

Tomáš Mrkvička, Vladimíra Petrášková,
Michal Houda, Jan Fiala

Pravděpodobnost a statistika s řešenými příklady v prostředí MS Excel



SERIA
EDUCATION

Tomáš Mrkvička, Vladimíra Petrášková,
Michal Houda, Jan Fiala

Pravděpodobnost a statistika s řešenými příklady v prostředí MS Excel



Pravděpodobnost a statistika s řešenými příklady v prostředí MS Excel

Autoři: Tomáš Mrkvička
Vladimíra Petrášková
Michal Houda
Jan Fiala

Recenzenti: prof. RNDr. Jindřich Klůfa, CSc.
doc. RNDr. Kateřina Helisová, Ph.D.
RNDr. Jakub Staněk, Ph.D.

Vydalo nakladatelství: Jihočeská univerzita v Českých Budějovicích,
Pedagogická fakulta

Ilustrace: Daniel Kyncl

Grafický návrh obálky: Přemysl Rosa

Sazba: Přemysl Rosa

Michal Houda

Tisk: Tiskárna Protisk, s.r.o.

1. vydání

© Tomáš Mrkvička, Vladimíra Petrášková, Michal Houda, Jan Fiala, 2025

© Jihočeská univerzita v Českých Budějovicích, Pedagogická fakulta, 2025

ISBN 978-80-7694-134-2 (tisk), 978-80-7694-135-9 (e-kniha)

Obsah

Předmluva	9
1 Úvod	11
1.1 Stručně z historie	11
1.2 Náhodný pokus	12
1.3 Seznam použitých symbolů	14
1.4 Seznam symbolů rozdělení náhodné veličiny	15
I Teorie pravděpodobnosti	17
2 Náhodný jev	19
2.1 Axiomatická definice pravděpodobnosti	19
2.2 Klasický pravděpodobnostní prostor	25
2.3 Geometrická pravděpodobnost	29
2.4 Obecné pravděpodobnostní prostory	31
2.5 Úlohy	33
3 Podmíněná pravděpodobnost	39
3.1 Podmíněná pravděpodobnost	39
3.2 Nezávislost	44
3.3 Bernoulliho proces	48
3.4 Úlohy	51
4 Celková pravděpodobnost, Bayesův vzorec	57
4.1 Celková pravděpodobnost	57
4.2 Bayesův vzorec	59

4.3	Úlohy	64
5	Náhodná veličina	67
5.1	Zavedení náhodné veličiny	67
5.2	Rozdělení náhodné veličiny	69
5.3	Distribuční funkce	71
5.4	Diskrétní náhodné veličiny	75
5.5	Absolutně spojité náhodné veličiny	78
5.6	Zobecnění	83
5.7	Úlohy	88
6	Charakteristiky náhodných veličin	93
6.1	Střední hodnota náhodné veličiny	93
6.2	Další charakteristiky polohy	98
6.3	Rozptyl náhodné veličiny	101
6.4	Momenty vyšších řádů	104
6.5	Úlohy	107
7	Rozdělení diskrétních veličin	111
7.1	Rovnoměrné diskrétní rozdělení	111
7.2	Degenerované rozdělení	113
7.3	Alternativní rozdělení	113
7.4	Binomické rozdělení	115
7.5	Geometrické rozdělení	121
7.6	Negativně-binomické rozdělení	126
7.7	Hypergeometrické rozdělení	129
7.8	Poissonovo rozdělení	132
7.9	Úlohy	138
8	Rozdělení spojitých veličin	143
8.1	Rovnoměrné spojité rozdělení	143
8.2	Exponenciální rozdělení	145
8.3	Erlangovo rozdělení a rozdělení gama	151
8.4	Obecné normální rozdělení	155
8.5	Normované normální rozdělení	159

8.6	Logaritmicko-normální rozdělení	164
8.7	Pearsonovo rozdělení χ^2	166
8.8	Studentovo t -rozdělení	169
8.9	Fisherovo–Snedecorovo rozdělení	171
8.10	Úlohy	174
9	Náhodný vektor	179
9.1	Náhodný vektor	179
9.2	Charakteristiky náhodného vektoru	185
9.3	Některá mnohorozměrná rozdělení	189
9.3.1	Multinomické rozdělení	189
9.3.2	Obecné dvourozměrné normální rozdělení	190
9.4	Nezávislé náhodné veličiny	191
9.5	Úlohy	195
10	Funkce náhodných veličin	203
10.1	Funkce náhodných veličin	203
10.2	Funkce náhodného vektoru	206
10.3	Konvoluce	208
10.4	Úlohy	214
11	Zákon velkých čísel, CLV	219
11.1	Konvergence náhodných veličin	219
11.2	Zákon velkých čísel	221
11.3	Centrální limitní věta	222
11.4	Úlohy	229
II	Matematická statistika	231
12	Zpracování statistického materiálu	233
12.1	Základní pojmy	233
12.2	Rozložení četností a třídění	235
12.3	Charakteristiky polohy	243
12.4	Charakteristiky variability (měnlivosti)	249
12.5	Obecné výběrové momenty	253

12.6	Charakteristiky symetrie a koncentrace	255
12.7	Úlohy	256
13	Stavební kameny statistiky	259
13.1	Základní a výběrový soubor	259
13.2	Rozdělení výběrových statistik	263
13.3	Kritické hodnoty rozdělení	264
13.4	Úlohy	266
14	Odhady parametrů	269
14.1	Kvalita odhadu	271
14.2	Bodové odhady střední hodnoty a rozptylu	273
14.3	Intervalové odhady	273
14.3.1	Intervalový odhad střední hodnoty normálního rozdělení	274
14.3.2	Odhad rozsahu výběru pro danou šířku intervalového odhadu střední hodnoty	275
14.3.3	Intervalový odhad rozptylu	277
14.4	Intervalový odhad střední hodnoty pomocí CLV	278
14.5	Úlohy	281
15	Parametrické testy	285
15.1	Principy testování hypotéz	285
15.2	Jednovýběrový t -test	288
15.2.1	Výpočet oboustranné a jednostranné p -hodnoty t -testu	289
15.3	Test rozptylu normálního rozdělení	291
15.4	Párový t -test	292
15.5	Dvouvýběrový t -test	295
15.6	F -test	297
15.7	Testování rovnosti středních hodnot souborů s různými rozptyly	300
15.8	Test střední hodnoty pomocí CLV	303
15.9	Úlohy	306
16	Neparametrické testy	313
16.1	Znaménkový test	314
16.2	Jednovýběrový Wilcoxonův test	316

16.3	Dvouvýběrový Wilcoxonův test	319
16.4	Úlohy	321
17	Porovnání více výběrů	325
17.1	Analýza rozptylu jednoduchého třídění	325
17.1.1	Levenův test	328
17.2	Kruskalův–Wallisův test	332
17.3	Analýza rozptylu dvojného třídění	333
17.4	Friedmanův test	339
17.5	Úlohy	343
18	Korelační analýza	353
18.1	Pearsonův výběrový korelační koeficient	353
18.2	Spearmanův korelační koeficient	355
18.3	Korelace versus kauzalita	357
18.4	Úlohy	359
19	Regresní analýza	363
19.1	Lineární regrese s jednou proměnnou	363
19.2	Lineární regrese s více proměnnými	370
19.3	Polynomiální regrese	376
19.4	Nelineární regrese	380
19.5	Úlohy	384
20	Testy dobré shody	389
20.1	Pearsonův χ^2 test	389
20.2	Test normality	391
20.3	Test Poissonova rozdělení	393
20.4	Kolmogorovův–Smirnovův jednovýběrový test	395
20.5	Úlohy	399
21	Kontingenční tabulky	403
21.1	Test nezávislosti	404
21.2	Test homogenity multinomických rozdělení	406
21.3	Test χ^2 ve čtyřpolních tabulkách	408

21.4 Fisherův faktoriálový (exaktní) test	409
21.5 McNemarův test	410
21.6 Test symetrie	412
21.7 Úlohy	415
Závěr	425
Summary	427
Rejstřík	429
Funkce a nástroje MS Excel	437
Statistické tabulky	445

Předmluva

Pravděpodobnostní úsudky, modely a předpovědi jsou dnes již běžnou součástí výzkumné práce, v řadě vědních oblastí nalézají i široké uplatnění v praxi. Pravděpodobnostní metody se používají při zkoumání procesů ovlivněných náhodou. V situacích, kdy umíme s vlivem náhody počítat, je často možné pomocí těchto metod dojít k optimálním řešením, k racionálnímu rozhodování apod.



CELA' TATO KNIHA
SE DA' VTĚSNAT DO HLAVY
JEDNOHO STATISTIKA

S počtem pravděpodobnosti je úzce spojena matematická statistika, které se užívá při vyhodnocování různých testů a experimentů. Veškerá moderní analýza v přírodních i humanitních vědách je založena na statistické inferenci, tedy vytváření závěrů o populaci na základě statistického (vědeckého) zkoumání dat, podléhajících variabilitě nebo nejistotě. Bez variability není co statisticky zkoumat.

Učebnice je psána formou přístupnou širšímu okruhu zájemců, což někdy vede k menší matematické obecnosti. Jednotlivé kapitoly obsahují nejdříve výklad teorie, na který navazují řešené příklady. Příklad řešený v MS Excel je uvozen ikonou . Na konci každé kapitoly jsou úlohy k samostatnému řešení (vždy s výsledky, někde i dílčími pokyny a doporučeními). Kapitola označená hvězdičkou * obsahuje rozšiřující učivo, které zobecňuje a rozšiřuje ostatní text.

Autoři

Kapitola 1

Úvod

1.1 Stručně z historie

Teorie pravděpodobnosti a matematická statistika jsou matematické disciplíny spadající souhrnně do vědního oboru *stochastika*¹, která se zabývá zkoumáním a modelováním náhodných jevů. První zmínku o stochastice lze sice nalézt již v díle Platona *Philebos*, zásadního rozmachu dosáhly však obě uvedené disciplíny až ve 20. století. Základy *teorie pravděpodobnosti* byly položeny zhruba v období renesance, kdy se nejslavnější matematikové té doby zabývali mimo jiné otázkami nad snad nejběžnější zábavou té doby – hazardní hrou v kostky². Za prvního západního matematika studujícího hazardní hry je dnes označován Girolamo Cardano³, který kolem roku 1564 napsal knihu *Liber de Ludo Aleae* (Kniha o hře náhodné). Kniha ovšem vyšla až téměř o století později v roce 1663, kdy již byla překonána pracemi Pascala⁴ a Fermata⁵. V té době Pascalův přítel, jistý francouzský amatérský matematik, svými přáteli přezdívaný rytíř z Méré (Chevalier de Méré)⁶, zkoušel vydělat jmění s následující dvojicí kostkových her.



¹Ze starořeckého stóchos – mířit na cíl, hádat. Anglicky *stochastics*.

²Arabsky se kostka řekne *az-zahr*. Odtud se slovo dostalo patrně do španělštiny v podobě *azar* a do francouzštiny v podobě *hasard*.

³Girolamo Cardano (1501–1576), italský fyzik, matematik a hazardní hráč.

⁴Blaise Pascal (1623–1662), francouzský matematik, fyzik, vynálezce, filozof, spisovatel a katolický teolog.

⁵Pierre de Fermat (1601–1665), francouzský matematik a právník.

⁶Antoine Gombaud (1607–1684), francouzský spisovatel a salónní teoretik. Svoji přezdívku Chevalier de Méré získal podle jedné ze svých literárních postav.

Příklad 1.1 (Kostkové hry rytíře z Méré). Rytíř z Méré s oblibou přijímal sázky na to, že při čtyřech hodech jednou kostkou hodí alespoň jednu šestku. Spočítal si totiž, že jeho šance na výhru v tomto případě činí $4 \cdot \frac{1}{6} = \frac{2}{3}$. Přestože je tento jeho výpočet chybný, dokázal na rovných sázkách v této hře vyhrávat. Takové štěstí však již neměl ve své druhé kostkové hře. Tentokrát házel dvěma kostkami celkem 24krát a bank bral v případě, že během těchto hodů padla alespoň jednou dvojice šestek. Jeho podobně chybný výpočet byl následující: šance na padnutí dvojice šestek je $\frac{1}{36}$, tudíž v 24 hodech jeho šance na výhru činí $24 \cdot \frac{1}{36} = \frac{2}{3}$, tedy stejně jako v první hře.

Protože ve druhé hře navzdory svému výpočtu prohrál značnou sumu peněz, požádal členy Pařížské akademie a především Pascala o pomoc s vysvětlením svého neúspěchu. Úvahy nad způsobem řešení této úlohy (a ještě o něco slavnější souběžně řešené úlohy o rozdělení sázky), vtělené do korespondence Pascala s Fermatem, otevřely stavidla k obecnějším úvahám o matematické „naději“, tedy o pravděpodobnosti chápané jako vědní obor. Konkrétnějšími výstupy v souvislosti s Méréovými otázkami byly to, co dnes známe pod názvy Pascalův trojúhelník, binomické rozdělení a očekávaná neboli střední hodnota.

1.2 Náhodný pokus

Ústředním pojmem teorie pravděpodobnosti a matematické statistiky je náhodný pokus. *Pokus* je obecně jakýkoliv proces generující nějaké výsledky (data). S pokusy jsme se všichni setkali během školních let v hodinách fyziky či chemie, jako například tření ebonitové tyče lištěm ohonem. Pokusy, které při opakování vedou k jednomu a témuž výsledku (v tomto případě ke vzniku záporného elektrického náboje na tyči), nazýváme pokusy *deterministické*. Zachováme-li identické vnější podmínky, výsledek opakovaného deterministického pokusu bude vždy stejný (na tyči vždy vznikne elektrický náboj). V chemii si můžeme představit jiný deterministický pokus, při kterém dojde k neutralizaci kyseliny přidáním přesně specifikovaného množství zásady. Toto množství je pro daný druh a koncentrace kyseliny a zásady identické a při opakování pokusu za jinak stejných podmínek se nemění.

Zajímavější z hlediska zkoumání jsou pokusy, jejichž výsledky se při opakování liší (podléhají určité nejistotě), i když jinak zachováme veškeré vnější experimentální podmínky. Takové pokusy označujeme jako *náhodné*. Jak jsme již zmínili



v úvodu, nejstaršími zkoumanými náhodnými pokusy byly hody kostkou a elementární prvky podobných her (hod mincí, karetní hry, sportka, ruleta aj.). Moderní teorie pravděpodobnosti se však nezaměřuje pouze na náhodné pokusy v oblasti hazardních her (ty ve skutečnosti nejsou příliš velkého významu), ale nabízí velmi obecné koncepty důležité k řešení mnoha praktických úloh a životních situací: namátkou zmiňme zkoumání účinnosti léčivých přípravků, výnosnost zemědělských plodin, ověřování kvality průmyslové výroby apod. V této publikaci se budeme zabývat výhradně náhodnými pokusy.

1.3 Seznam použitých symbolů

$A, B, \dots$	náhodný jev
$\{A_i\}$	posloupnost jevů A_i
$\emptyset$	nemožný jev, prázdná množina
$\bar{A}$	opačný náhodný jev k jevu A
Ω	jistý jev, množina všech možných výsledků náhodného jevu
ω	výsledek náhodného pokusu
$P(\cdot)$	pravděpodobnost náhodného jevu jako proměnná nebo hodnota proměnné
$\mathcal{P}(\Omega)$	potenční množina množiny Ω
$\mathcal{A}$	σ -algebra
f_n	relativní četnost
$\mathbb{P}(A)$	pravděpodobnost náhodného jevu A ve významu pravděpodobnostní funkce
$\mathbb{P}(A B)$	podmíněná pravděpodobnost jevu A za podmínky, že nastal jev B
$X, Y, \dots$	náhodné veličiny
$p_X(\cdot)$	pravděpodobnostní funkce diskrétní náhodné veličiny X
$\varphi(\cdot)$	hustota pravděpodobnosti normovaného normálního rozdělení
$\Phi(\cdot)$	distribuční funkce normovaného normálního rozdělení
$f_X(\cdot)$	hustota pravděpodobnosti spojité náhodné veličiny X
$F_X(\cdot)$	distribuční funkce náhodné veličiny X
m	moment náhodné veličiny
m_A	počet výsledků příznivých jevu A
θ	teoretický parametr populace
μ	střední (očekávaná) hodnota jako parametr populace
μ^1, μ^2	jednorozměrná míra (délka), dvourozměrná míra (obsah)
$\mathbb{E}X$	střední hodnota náhodné veličiny X
$\text{var } X$	rozptyl náhodné veličiny X
σ^2	populační rozptyl
S^2	výběrový rozptyl (ve smyslu náhodné veličiny)
σ	populační směrodatná odchylka
S	výběrová směrodatná odchylka (ve smyslu náhodné veličiny)
v	variační koeficient
$[x]$	zaokrouhlení čísla x na celé číslo směrem dolů
$\mathbb{R}$	množina reálných čísel
$\mathbb{N}$	množina přirozených čísel

$\mathcal{B}, \mathcal{B}(\mathbb{R})$	borelovská σ -algebra (na množině reálných čísel)
$\bar{X}$	výběrový průměr (ve smyslu náhodné veličiny)
$\bar{x}$	aritmetický průměr
$\tilde{x}$	medián
$\hat{x}$	modus
$\bar{x}_g$	geometrický průměr
$\bar{x}_h$	harmonický průměr
R	variační rozpětí
R	výběrový korelační koeficient (ve smyslu náhodné veličiny)
r	vypočtená hodnota výběrového korelačního koeficientu
d	absolutní odchylka od aritmetického průměru
$\bar{d}$	průměrná absolutní odchylka
$s_{x,y}, \text{cov}(x, y)$	kovariance
n	četnost, rozsah souboru
$\bigcap_{i=1}^n A_i$	průnik jevů A_i
$\bigcup_{i=1}^n A_i$	sjednocení jevů A_i
$\prod_{i=1}^n \mathbb{P}(A_i)$	součin pravděpodobností jevů A_i
$\sum_{i=1}^n \mathbb{P}(A_i)$	součet pravděpodobností jevů A_i
$\mathbf{X}$	náhodný vektor
$\mathbf{X}^T$	transponovaný vektor (matice vektoru)
B	funkce beta
Γ	funkce gama

1.4 Seznam symbolů rozdělení náhodné veličiny

A	Alternativní (Bernoulliho, nula-jedničkové) rozdělení
Bi	binomické rozdělení
Dg	degenerované rozdělení
Erlang	Erlangovo rozdělení
Exp	exponenciální rozdělení
F	Fisherovo-Snedecorovo F -rozdělení
Γ	rozdělení gama
Ge	geometrické rozdělení
HGeom	hypergeometrické rozdělení
χ^2	rozdělení χ -kvadrát

LN	logaritmicko-normální rozdělení
N	normální rozdělení
NB	negativně binomické rozdělení
Po	Poissonovo rozdělení
q	studentizované rozpětí
Rd	rovnoměrné diskrétní rozdělení
t	Studentovo t -rozdělení
U	rovnoměrné spojité rozdělení
Wb	Weibullovo rozdělení

Část I

Teorie pravděpodobnosti

Kapitola 2

Jev, náhodný jev, pravděpodobnosti náhodného jevu

2.1 Axiomatická definice pravděpodobnosti

V úvodní kapitole jsme si vysvětlili, co představuje náhodný pokus. Každému náhodnému pokusu můžeme přiřadit množinu Ω , obsahující jako prvky všechny možné výsledky pokusu. Při hodu kostkou je například

$$\Omega = \{\square, \blacksquare, \blacklozenge, \blacktriangle, \blacktriangledown, \blacksquare\},$$

při hodu mincí je

$$\Omega = \{\text{rub}, \text{líc}\},$$

při hodu dvěma mincemi je

$$\Omega = \{(\text{rub}, \text{rub}), (\text{líc}, \text{líc}), (\text{líc}, \text{rub}), (\text{rub}, \text{líc})\}.$$

Množina Ω může být konečná nebo nekonečná. Má však smysl uvažovat pouze neprázdné množiny. Náhodný pokus má být množinou Ω jednoznačně popsán, tzn. že množina Ω musí být vyčerpávající (musí nastat vždy právě jeden z výsledků) a výsledky se musí navzájem vylučovat (nemůže se stát, aby dva výsledky nastaly současně). Prvky $\omega \in \Omega$ nazýváme *elementárními jevy*.

Otázka 2.1. Jak vypadají elementární jevy v obou kostkových hrách rytíře z Méré z příkladu 1.1?

Tabulka 2.1: Zápis základních pravděpodobnostních relací a operací

Zápis	Pravděpodobnostní interpretace
$\omega \in A$	Jev A nastal, výsledek ω náhodného pokusu je příznivý jevu A .
$A \subset B$	A je podjev jevu B , jev B nastane, kdykoliv nastane jev A .
$B - A$, také $B \setminus A$	Rozdíl jevu B a A v tomto pořadí je jev, který nastane právě tehdy, když nastane jev B a zároveň nenastane jev A .
$\bar{A} = \Omega - A$	Doplňěk jevu A je jev, který nastane právě tehdy, když nenastane jev A .
$A \cup B$	Sjednocení jevů A a B je jev, který nastane právě tehdy, nastane-li alespoň jeden z jevů A , B .
$A \cap B$	Průnik jevů A a B je jev, který nastane právě tehdy, nastanou-li oba dva jevy současně.
$A \cap B = \emptyset$	Jevy A a B nazveme disjunktní, nemohou nastat současně.
$A_i \cap A_j = \emptyset, i \neq j,$ $\bigcup_{i=1}^n A_i = C$	Jevy $A_1, \dots, A_n$ tvoří rozklad jevu C .

Podmnožiny Ω se nazývají *náhodné jevy*. Jelikož možné výsledky náhodných pokusů fakticky tvoří takové množiny, při práci s jevy užíváme množinových relací (např. podmnožina) a operací (průnik, sjednocení, rozdíl, doplněk), které ovšem doplňujeme o pravděpodobnostní interpretaci a někdy i vlastní názvosloví. Přehled uvádíme v tabulce 2.1. Průnik a sjednocení jevů se dají rozšířit i na více jevů (na konečnou i nekonečnou posloupnost jevů). Tzv. *jev jistý* je takový náhodný jev, který nastane při každé realizaci pokusu. Značíme ho Ω , neboť je ekvivalentní množině všech možných výsledků. Tzv. *jev nemožný* je takový jev, který nenastane při žádné realizaci pokusu. Značíme ho $\emptyset$, protože je ekvivalentní prázdné množině.

Ne všechny podmnožiny Ω nutně musí být náhodné jevy. Některé nejsou zajímavé z hlediska interpretace, jiné nemohou být jako jevy vůbec označeny, protože nejdou „změřit“ a není možné nad nimi vybudovat smysluplnou teorii. To vedlo v roce 1933 Kolmogorova¹ k obecné formalizaci pravděpodobnosti založené na kolekcích podmnožin, které jsou uzavřené vzhledem k doplňkům a spočetným sjednocením. Tyto kolekce se nazývají σ -algebry a obsahují všechny podmnožiny Ω , kterými má smysl se zabývat. S trochou nadsázky bychom σ -algebru mohli nazvat kolekcí všech „hezkých“ podmnožin Ω .

Definice 2.1 (σ -algebra)

Nechť $\mathcal{A}$ je neprázdný systém podmnožin množiny $\Omega \neq \emptyset$ takový, že

- a) $\emptyset \in \mathcal{A}$,
- b) je-li $A \in \mathcal{A}$, pak $\bar{A} \in \mathcal{A}$,
- c) jsou-li $A_i \in \mathcal{A}$, $i = 1, 2, \dots$, pak $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$.

Pak $\mathcal{A}$ nazýváme σ -algebrou.

Jednotlivé prvky vhodně zvolené σ -algebry $\mathcal{A}$ nazýváme *náhodné jevy*. Dvojici $(\Omega, \mathcal{A})$ nazýváme *jevové pole*. Uvedme si nyní bez důkazu některé další vlastnosti náhodných jevů. Definice 2.1 spolu s větou 2.1 zajišťují, že výsledky běžných množinových operací nad náhodnými jevy budou opět náhodné jevy.

Věta 2.1 (o některých vlastnostech náhodných jevů). *Nechť Ω je jistý jev a A, B jsou náhodné jevy v $\mathcal{A}$.*

- a) *Jev jistý je náhodný jev: $\Omega \in \mathcal{A}$.*
- b) *Rozdíl dvou náhodných jevů je náhodný jev:*

$$\text{jsou-li } A, B \in \mathcal{A}, \text{ pak } B - A \in \mathcal{A}.$$

¹ Andrej Nikolajevič Kolmogorov (1903–1987), ruský matematik, tvůrce axiomatické definice pravděpodobnosti.

c) Průnik spočetně mnoha náhodných jevů je náhodný jev:

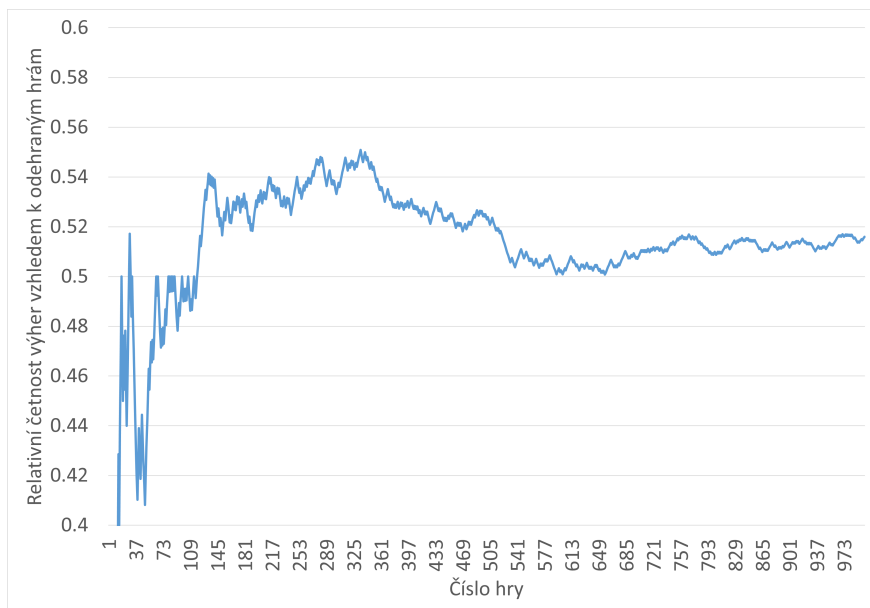
$$\text{jsou-li } A_1, A_2, \dots \in \mathcal{A}, \text{ pak } \bigcap_{i=1}^{\infty} A_i \in \mathcal{A}.$$

Příklad 2.1. Necht $\Omega = \{1, \dots, n\}$. Pak potenční množina $\mathcal{P}(\Omega)$ (tj. množina všech podmnožin Ω) je σ -algebra a neexistuje menší σ -algebra obsahující všechny elementární jevy $\{\omega\}$, $\omega \in \Omega$.

Příklad 2.2. Necht $\Omega = \mathbb{R}$. Pak potenční množina $\mathcal{P}(\mathbb{R})$ je také σ -algebra, ale existuje i menší σ -algebra, které dáváme (z důvodů zmíněných dále) přednost. Je to především tzv. *borelovská* σ -algebra, kterou definujeme jako nejmenší σ -algebru obsahující všechny otevřené podmnožiny $\mathbb{R}$. Borelovská σ -algebra obsahuje všechna spočetná sjednocení otevřených množin, ale také všechny uzavřené podmnožiny $\mathbb{R}$. Označujeme ji $\mathcal{B}(\mathbb{R})$ nebo jen krátce $\mathcal{B}$.

Každému náhodnému jevu lze přiřadit reálné číslo reprezentující „šanci“, s kterou jev nastane. Toto číslo nazveme *pravděpodobnost* náhodného jevu. S pojmem pravděpodobnosti jsme se jistě každý v běžném životě již setkali a mnohokrát použili, aniž bychom znali přesnou definici tohoto pojmu. Vezměme si například za náhodný pokus hod běžnou hrací kostkou. Při něm je pro každý výsledek $\square, \blacksquare, \dots, \boxtimes$ pravděpodobnost padnutí rovna $\frac{1}{6}$. Při hodu mincí je pravděpodobnost padnutí rubu stejná jako pravděpodobnost padnutí líce, tj. $\frac{1}{2}$. Tato tvrzení o pravděpodobnosti výsledků náhodného pokusu hodu kostkou a hodu mincí plynou z naší zkušenosti. Na pojem pravděpodobnosti se můžeme podívat i ze statistické stránky: házejme kostkou a vyšetřujeme relativní četnosti $f_n = \frac{m}{n}$ zvoleného jevu (n je počet hodů, m je počet hodů, ve kterých padl výsledek příznivý vyšetřovanému jevu), zjistíme při větším počtu hodů, že relativní četnosti elementárních výsledků budou kolísat kolem $\frac{1}{6} = 0,1\bar{6}$. U hodů mincí budou relativní četnosti kolísat kolem $\frac{1}{2} = 0,5$.

Příklad 2.3 (Relativní četnosti první kostkové hry rytíře z Méré z příkladu 1.1 na straně 12).  Tento statistický přístup lze snadno simulovat na počítači. Podívejme se, jak by si mohl rytíř z Méré poradit se svou otázkou ohledně šance na výhru v první kostkové hře, kdyby byla tato vymoženost k dispozici již v 17. století. V MS Excelu k tomu využijeme funkci na generování náhodných čísel. Hod jednou kostkou lze simulovat pomocí `=CEILING.MATH(NÁHČÍSLO()*6)` nebo jednodušeji `=RANDBETWEEN(1;6)`. Každá hra sestává ze čtyř takových hodů, vzorec tedy umístíme například do čtyř buněk B2 až E2 vedle sebe. Rytíř z Méré v jedné hře vítězí a vyhrává bank, pokud padla alespoň jedna šestka: `= (COUNTIF(B2:E2;6)>0)*1`. Nyní celý řádek zkopírujeme tolikrát, kolik her hodláme simulovat (například



Obrázek 2.1: Relativní četnost výhry v první kostkové hře rytíře de Méré z příkladu 2.3

tisíc). Sečteme počet jedniček a vydělíme počtem her, čímž získáme požadovanou relativní četnost výhry. Po určitém počtu her (při jedné simulaci ilustrované na obrázku 2.1 to bylo po asi 430 hrách) bude relativní četnost výhry jen málo kolísat kolem hodnoty 0,52.

Jak se naučíme počítat později, skutečná pravděpodobnost výhry v této hře je 0,518, tedy velmi blízko námi simulované hodnotě. Rytíř z Méré měl ve skutečnosti možnost takovou simulaci provést i ve své době, jenom by provedení simulace bez použití počítače (zato se skutečnou kostkou) trvalo o něco déle. Kdyby si to uvědomil, ušetřil by jistě velké jmění a nepokoušel by se hrát druhou ze svých her.



Otázka 2.2. Pomocí simulace odhadněte pravděpodobnost úspěchu rytíře z Méré v jeho druhé kostkové hře.

I jen z intuitivní představy můžeme dedukovat, že relativní četnosti mají následující vlastnosti:

- 1) $0 \leq f_n(A) \leq 1$ pro libovolný náhodný jev $A \in \mathcal{A}$,

- 2) je-li $A_1, A_2, \dots$ posloupnost náhodných jevů taková, že pro libovolná $i \neq j$, $i, j = 1, 2, \dots$ platí $A_i \cap A_j = \emptyset$, pak

$$f_n \left(\bigcup_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} f_n(A_i).$$

- 3) $f_n(\emptyset) = 0$ a $f_n(\Omega) = 1$,

Má-li pravděpodobnost odpovídat relativním četnostem jevů, jak to naznačuje příklad 2.3, pak musí vykazovat obdobné vlastnosti. Přesně toto si uvědomil Kolmogorov a odvodil následující definici pravděpodobnosti:

Definice 2.2 (Pravděpodobnost náhodného jevu)

Nechť $\Omega \neq \emptyset$ a $\mathcal{A}$ je σ -algebra definovaná na Ω , tj. $(\Omega, \mathcal{A})$ je jevové pole. Pak pravděpodobnostní mírou (krátce pravděpodobností) nazveme libovolnou reálnou funkci $\mathbb{P}: \mathcal{A} \rightarrow \mathbb{R}$, která splňuje následující trojici vlastností:

- a) (nezápornost) $\mathbb{P}(A) \geq 0$ pro libovolný náhodný jev $A \in \mathcal{A}$,
b) (σ -aditivita) pro každou posloupnost disjunktních jevů $\{A_n\}_{n=1}^{\infty}$ platí

$$\mathbb{P} \left(\bigcup_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

- c) (normovanost) $\mathbb{P}(\Omega) = 1$.

Trojice $(\Omega, \mathcal{A}, \mathbb{P})$ se nazývá pravděpodobnostní prostor. Hodnota $\mathbb{P}(A)$ se nazývá pravděpodobnost jevu A .

Pravděpodobnostní míra takto definovaná vyžaduje řadu dalších intuitivních vlastností. Shrňme si je v následující větě.



Věta 2.2 (o vlastnostech pravděpodobnosti náhodného jevu, bez důkazů). Pro pravděpodobnost $\mathbb{P}$ jevů $A, B \in \mathcal{A}$ platí:

- a) $\mathbb{P}(\emptyset) = 0$,
b) $\mathbb{P}$ je konečně aditivní, tj. jestliže $A_1, \dots, A_n \in \mathcal{A}$, $A_i \cap A_j = \emptyset$ pro libovolná $i \neq j$, $i, j = 1, \dots, n$, pak $\mathbb{P} \left(\bigcup_{i=1}^n A_i \right) = \sum_{i=1}^n \mathbb{P}(A_i)$,
c) $\mathbb{P}$ je monotónní, tj. jsou-li $A, B \in \mathcal{A}$ a $A \subset B$, pak $\mathbb{P}(A) \leq \mathbb{P}(B)$,
d) jsou-li $A, B \in \mathcal{A}$ a $A \subset B$, pak $\mathbb{P}(B - A) = \mathbb{P}(B) - \mathbb{P}(A)$,
e) $\mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A)$ pro libovolný náhodný jev $A \in \mathcal{A}$,
f) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ pro libovolné náhodné jevy $A, B \in \mathcal{A}$,
g) $\mathbb{P} \left(\bigcup_{i=1}^{\infty} A_i \right) \leq \sum \mathbb{P}(A_i)$ pro libovolnou posloupnost $\{A_i\}$ v $\mathcal{A}$,

- h) *jestliže pro všechna n je $A_n \in \mathcal{A}$, $A_1 \subset A_2 \subset A_3 \dots$, $A = \bigcup_{n=1}^{\infty} A_n$,
pak $\mathbb{P}(A) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n)$,*
- i) *jestliže pro všechna n je $A_n \in \mathcal{A}$, $A_1 \supset A_2 \supset A_3 \dots$, $A = \bigcap_{n=1}^{\infty} A_n$,
pak $\mathbb{P}(A) = \lim_{n \rightarrow \infty} \mathbb{P}(A_n)$.*

Vlastnost f) lze indukcí rozšířit na libovolný konečný počet jevů takto:

$$\begin{aligned} \mathbb{P}\left(\bigcup_{i=1}^n A_i\right) &= \sum_{i=1}^n \mathbb{P}(A_i) - \sum_{i=1}^{n-1} \sum_{j=i+1}^n \mathbb{P}(A_i \cap A_j) + \\ &+ \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^n \mathbb{P}(A_i \cap A_j \cap A_k) + \dots + (-1)^{n-1} \mathbb{P}\left(\bigcap_{i=1}^n A_i\right). \end{aligned} \quad (2.1)$$

2.2 Klasický pravděpodobnostní prostor

Kolmogorovova axiomatická definice pravděpodobnosti je nekonstruktivní: nespécifikuje vůbec, jak má jevové pole a pravděpodobnost vypadat, pouze specifikuje, že jevy musí tvořit σ -algebru dle definice 2.1 a pravděpodobnost splňovat podmínky definice 2.2. Starší pokusy o vymezení pravděpodobnosti byly vesměs konstruktivního charakteru – specifikovaly konkrétní vlastnosti pravděpodobnosti, ale zároveň tím přišly o potřebnou obecnost, aby byly schopny popsat náhodné chování složitějších procesů. Nejstarším zkoumaným pravděpodobnostním prostorem je tzv. *klasická pravděpodobnost*.

Definice 2.3 (Klasický pravděpodobnostní prostor²)

Pravděpodobnostní prostor $(\Omega, \mathcal{A}, \mathbb{P})$ nazveme klasickým pravděpodobnostním prostorem, jestliže

- množina Ω je konečná o m prvcích,*
- za σ -algebru $\mathcal{A}$ vezmeme systém $\mathcal{P}(\Omega)$ všech podmnožin množiny Ω ,*
- všechny elementární jevy jsou stejně pravděpodobné, tedy označíme-li $p_1, \dots, p_m$ pravděpodobnosti jednotlivých elementárních jevů, pak*

$$p_1 = p_2 = \dots = p_m = \frac{1}{m}.$$

Pravděpodobnost $\mathbb{P}$ náhodného jevu A je potom rovna

$$\mathbb{P}(A) = \frac{m_A}{m},$$

²Uvedená definice se nazývá též Laplaceova definice pravděpodobnosti. Jejím autorem je Pierre Simon de Laplace (1749–1827), francouzský matematik, fyzik, astronom a politik.

kde m_A je počet výsledků příznivých jevů A a m je počet všech možných výsledků náhodného pokusu. Pravděpodobnost takto definovaná se nazývá klasická pravděpodobnost.

Otázka 2.3. Ověřte, že klasická pravděpodobnost je pravděpodobností ve smyslu definice 2.2, tj. splňuje všechny tři axiomy.

V úlohách, při jejichž řešení se využije klasická pravděpodobnost, jde o nalezení množiny elementárních výsledků tak, aby tyto výsledky byly stejně pravděpodobné. Jakmile se nám to podaří, je výpočet pravděpodobností redukován na užití kombinatorických pravidel.

Příklad 2.4. Házíme dvěma kostkami. Jaká je pravděpodobnost jevu A „na kostkách padnou stěny s puntíky, jejichž součet je menší než 5“?

Řešení. Výsledky pokusu jsou uspořádané dvojice stěn kostek: první člen dvojice odpovídá výsledku hodu 1. kostkou a druhý člen odpovídá výsledku hodu 2. kostkou. Všechny možné výsledky jsou tedy reprezentovány množinou

$$\Omega = \{(\square, \square), (\square, \blacksquare), \dots, (\blacksquare, \blacksquare), (\blacksquare, \blacksquare)\}.$$

Počet všech možných výsledků je $|\Omega| = m = 36$. Výsledky příznivé sledovanému jevu A jsou $(\square, \square), (\square, \blacksquare), (\blacksquare, \square), (\blacksquare, \blacksquare), (\blacksquare, \blacksquare), (\blacksquare, \blacksquare)$, tedy $m_A = 6$. Hledaná pravděpodobnost jevu A je tedy rovna

$$\mathbb{P}(A) = \frac{6}{36} = \frac{1}{6}.$$

△

Příklad 2.5 (Pravděpodobnost výhry v druhé kostkové hře rytíře z Méré). Jaká je pravděpodobnost výhry v druhé kostkové hře rytíře z Méré z příkladu 1.1?

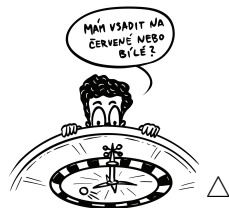
Řešení. Při hodu dvěma kostkami současně máme celkem 36 možných výsledků, jako v příkladu 2.4. Pouze jeden z nich je $(\blacksquare, \blacksquare)$, tedy pravděpodobnost, že v jednom hodu dvěma kostkami hodíme dvě šestky, je $\frac{1}{36}$. Podle pravidla pro pravděpodobnost opačného jevu (bod e) věty 2.2) je tedy pravděpodobnost, že dvě šestky nehodíme, $\frac{35}{36}$. Pravděpodobnost, že nehodíme dvě šestky v žádném z 24 pokusů, je tedy

$$\mathbb{P}(\text{bez dvojice šestek v 24 pokusech}) = \left(\frac{35}{36}\right)^{24} \doteq 0,5086.$$

Výsledná pravděpodobnost, že v celé hře padne alespoň jednou dvojice šestek, rytíř z Méré zvítězí a bude brát bank, je tedy

$$\mathbb{P}(\text{alespoň jedna dvojice šestek v 24 pokusech}) = 1 - \left(\frac{35}{36}\right)^{24} \doteq 0,4914.$$

Výhoda soupeřů rytíře z Méré tedy v této hře činí asi 1,8 %. Pro srovnání, výhoda kasina v americké ruletě (s jednoduchou a dvojitou nulou) činí 5,26 % a ve francouzské (s jednou zelenou nulou) 2,7 %. Rytíř z Méré tedy neprohrál své peníze tak rychle, jako v kasinu, ale s obdobnou mírou jistoty.



Otázka 2.4. Využitím postupu z příkladu 2.5 spočítejte pravděpodobnost výhry rytíře z Méré v první kostkové hře, kterou jsme odhadli simulací v příkladu 2.3.

Příklad 2.6. V balíčku je 32 hracích karet, z toho čtyři esa. Dvakrát za sebou vytáhneme náhodně jednu kartu, přitom po prvním tahu vytaženou kartu

- vrátíme zpět do balíčku,
- nevrátíme zpět do balíčku.

Stanovte pravděpodobnost jevu A „alespoň jedna z vytažených karet je eso“.



Řešení. Výsledky pokusu jsou v obou případech opět uspořádané dvojice: první člen dvojice odpovídá kartě vytažené v 1. tahu a druhý člen kartě vytažené v 2. tahu. Dále se postup liší.

- V 1. tahu můžeme kartu vytáhnout 32 způsoby. Protože vytaženou kartu vracíme zpět do balíčku, i v 2. tahu máme 32 možností. Počet všech možných případů je tedy $m = 32^2 = 1\,024$. Příznivým případům odpovídají tahy s uspořádanými dvojicemi typu (libovolná karta, eso), (eso, libovolná karta), (eso, eso). Počet příznivých případů je $m_A = 28 \cdot 4 + 4 \cdot 28 + 4 \cdot 4 = 240$. Hledaná pravděpodobnost je rovna $\mathbb{P}(A) = \frac{240}{1024} = \frac{15}{64} \doteq 0,234$.
- Počet možných případů je vzhledem k tomu, že po prvním tahu kartu nevrátíme, $m = 32 \cdot 31 = 992$. Příznivým případům odpovídají opět tahy (libovolná karta, eso), (eso, libovolná karta), (eso, eso). Počet příznivých případů je nyní $m_A = 28 \cdot 4 + 4 \cdot 28 + 4 \cdot 3 = 236$. Hledaná pravděpodobnost je rovna $\mathbb{P}(A) = \frac{236}{992} = \frac{59}{248} \doteq 0,238$.

Postupovat lze v tomto druhém případě i volbou reprezentace, ve které vytáhneme obě karty současně. Výsledky pokusu jsou všechny nespořádané dvojice možných tahů, kterých je $\binom{32}{2}$. Počet případů, kdy vytáhneme alespoň jedno eso, můžeme vypočítat použitím doplňkového jevu, tj. pomocí počtu případů, kdy nevytáhneme ani jedno eso. Těchto případů je $\binom{28}{2}$. Celkem tedy $\mathbb{P}(A) = \frac{\binom{32}{2} - \binom{28}{2}}{\binom{32}{2}} = \frac{59}{248} \doteq 0,238$.

△

Příklad 2.7. Jednou házíme šesti kostkami, přičemž každou kostku očíslováme právě jedním z čísel $1, \dots, 6$. Jaká je pravděpodobnost, že alespoň na jedné kostce bude počet ok souhlasit s číslem, jímž jsme kostku označili?



Řešení. Necht A_i je takový jev, že na i -té kostce souhlasí počet ok s číslem, jímž je kostka označena, a necht jev A je jev „alespoň na jedné kostce souhlasí počet ok s číslem, jímž je kostka označena“. Pak $A = \bigcup_{i=1}^6 A_i$. Protože jevy $A_1, \dots, A_6$ nejsou disjunktní, je podle rovnice (2.1) na straně 25

$$\begin{aligned} \mathbb{P}\left(\bigcup_{i=1}^6 A_i\right) &= \sum_{i=1}^6 \mathbb{P}(A_i) - \sum_{i=1}^5 \sum_{j=i+1}^6 \mathbb{P}(A_i \cap A_j) + \\ &+ \sum_{i=1}^4 \sum_{j=i+1}^5 \sum_{k=j+1}^6 \mathbb{P}(A_i \cap A_j \cap A_k) + \dots + (-1)^5 \mathbb{P}\left(\bigcap_{i=1}^6 A_i\right). \end{aligned} \quad (2.2)$$

Hledejme pravděpodobnosti vyskytující se v (2.2). Hodu šesti kostkami odpovídá uspořádaná šestice, jejíž prvky se mohou opakovat. Počet možných výsledků je tedy $6^6 = 46\,656$. Jevu A_i , $i = 1, 2, \dots, 6$, jsou příznivé ty výsledky, u kterých souhlasí u i -té kostky počet ok s jejím číslem, na ostatních kostkách může padnout cokoliv. Takových výsledků je $6^5 = 7\,776$. Z toho plyne, že

$$\mathbb{P}(A_i) = \frac{6^5}{6^6} = \frac{1}{6}, \quad i = 1, 2, \dots, 6.$$

Nyní budeme počítat pravděpodobnosti $\mathbb{P}(A_i \cap A_j)$, kde $i, j = 1, 2, \dots, 6$, $i \neq j$. Počet možných výsledků je opět $6^6 = 46\,656$. Příznivé výsledky jsou ty, u kterých na i -té a j -té kostce souhlasí počet ok s jejich číslem, na ostatních kostkách může padnout cokoliv. Počet příznivých výsledků je tedy $6^4 = 1\,296$ a

$$\mathbb{P}(A_i \cap A_j) = \frac{6^4}{6^6} = \frac{1}{6^2} \doteq 0,028, \quad i, j = 1, 2, \dots, 6, \quad i \neq j.$$

Analogicky

$$\begin{aligned}\mathbb{P}(A_i \cap A_j \cap A_k) &= \frac{6^3}{6^6} = \frac{1}{6^3} \doteq 0,004\,6, \quad i < j < k, \\ &\vdots \\ \mathbb{P}\left(\bigcap_{i=1}^6 A_i\right) &= \frac{1}{6^6}.\end{aligned}$$

Hledaná pravděpodobnost jevu A je

$$\mathbb{P}(A) = 6 \cdot \frac{1}{6} - \binom{6}{2} \cdot \frac{1}{6^2} + \cdots + (-1)^5 \cdot \frac{1}{6^6} \doteq 0,665.$$

Zmíněný postup je dosti komplikovaný. Jednodušší je v tomto případě využít opačného jevu $\bar{A}$ (na žádném kostce nebude souhlasit počet ok s jejím číslem). Počet příznivých jevů $\bar{A}$ je 5^6 , tedy

$$\mathbb{P}(A) = 1 - \mathbb{P}(\bar{A}) = 1 - \frac{5^6}{6^6} \doteq 0,665.$$

△

Příklad 2.8. Při zahájení hry Pocker dostane každý hráč pět karet. Jaká je pravděpodobnost, že hráč dostane dva krále a současně dvě esa?

Řešení. Uvažujeme karty se čtyřmi různými „barvami“, od každé barvy 13 karet s hodnotami 2–10, joker, královna, král a eso. Počet všech možností, jak mohou vypadat různé pětky karet, je $\binom{52}{5} = 2\,598\,960$. Počet příznivých výsledků sledovanému jevu je $\binom{4}{2} \cdot \binom{4}{2} \cdot \binom{44}{1} = 1\,584$. Výsledná pravděpodobnost je $\mathbb{P} = \frac{1\,584}{2\,598\,960} \doteq 0,000\,609$. △

2.3 Geometrická pravděpodobnost

Někdy lze elementární výsledky reprezentovat pomocí nějakého geometrického útvaru (úsečkou, obdélníkem, koulí apod.), obecně podmnožinou bodů z $\mathbb{R}^d$. Jestliže všechny body mají stejnou „pravděpodobnostní váhu“, tj. žádný elementární výsledek není z hlediska šance na výskyt preferován před jiným, mluvíme o tzv. modelu geometrické pravděpodobnosti.



Definice 2.4 (Geometrická pravděpodobnost)

Pravděpodobnostní prostor $(\Omega, \mathcal{A}, \mathbb{P})$ nazveme geometrickou pravděpodobností, *jestliže*

- a) $\Omega \subset \mathbb{R}^d$,
- b) $\mathcal{A} = \mathcal{B}(\Omega)$ je borelovská σ -algebra na Ω (tj. nejmenší σ -algebra obsahující všechny otevřené podmnožiny Ω),
- c) pravděpodobnost jevu $A \in \mathcal{A}$ je přímo úměrná d -rozměrné Lebesgueově³ míře μ^d množiny A^4 :

$$\mathbb{P}(A) = \frac{\mu^d(A)}{\mu^d(\Omega)}.$$

Příklad 2.9. Autobusy přijíždějí na zastávku pravidelně v 10minutových intervalech. Student přijde na zastávku v náhodném čase. Jaká je pravděpodobnost, že bude čekat déle než 5 minut?



Řešení. Výsledkem pokusu je doba čekání studenta na autobus. Student může čekat 0 až 10 minut, prostor elementárních jevů je tedy úsečka $\Omega = [0; 10)$. Jev A (doba čekání delší než 5 minut) je otevřený interval $(5; 10)$. Hledaná pravděpodobnost je tedy:

$$\mathbb{P}(A) = \frac{\mu^1((5; 10))}{\mu^1([0; 10))} = \frac{5}{10} = \frac{1}{2} = 0,5.$$

△

Příklad 2.10. Dvě osoby (I, II) přijdou na místo schůzky mezi 12. a 13. hodinou. Časové okamžiky příchodu obou osob jsou náhodné a nezávislé. Ten, kdo přijde na místo schůzky, čeká 20 minut a nedočká-li se druhého, odchází. Jaká je pravděpodobnost, že se osoby setkají?

Řešení. Elementární jev je zde $\omega = (x, y)$, kde x značí dobu příchodu osoby I a y dobu příchodu osoby II. Stavový prostor je tedy $\Omega = [0, 60] \times [0, 60]$. Označme A jev, že „se osoby setkají“, pak máme

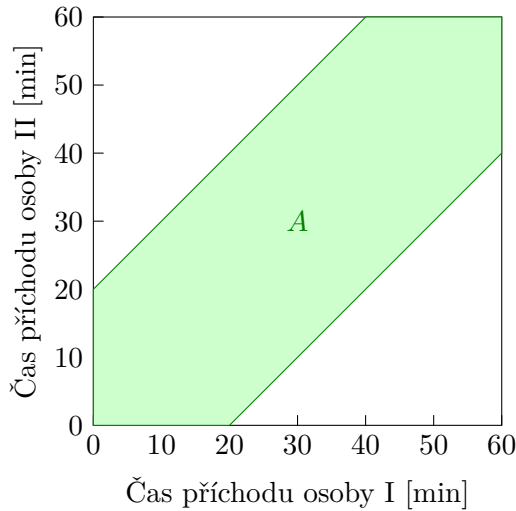
$$A = \{(x, y) \in \Omega : |x - y| \leq 20\}$$

(viz obrázek 2.2). Jev doplňkový je

$$\bar{A} = \{(x, y) \in \Omega : y < x - 20 \text{ nebo } y > x + 20\}.$$

³Henri Léon Lebesgue (1875–1941), francouzský matematik.

⁴Pro naše účely postačí, pokud si pod $\mu^1(A)$ představíme délku množiny A , pod $\mu^2(A)$ obsah A a pod $\mu^3(A)$ objem A .

Obrázek 2.2: Oblast setkání (jev A) k příkladu 2.10

Obsah množiny $\bar{A}$ skládající se ze dvou trojúhelníků je $\mu^2(\bar{A}) = 40 \cdot 40 = 1600$. Obsah Ω je $\mu^2(\Omega) = 3600$, což dohromady dává:

$$\mathbb{P}(A) = 1 - \mathbb{P}(\bar{A}) = 1 - \frac{\mu^2(\bar{A})}{\mu^2(\Omega)} = 1 - \frac{4}{9} = \frac{5}{9} = 0,5\bar{5}.$$

△

2.4 Obecné pravděpodobnostní prostory

Klasická pravděpodobnost nedokáže reprezentovat systémy, ve kterých je nekonečné množství elementárních jevů (např. náhodné jevy reprezentující počty událostí). Podobně ani klasická, ani geometrická pravděpodobnost si neporadí s případy, ve kterých nemají elementární jevy stejnou váhu (např. pravděpodobnost narození chlapce je o něco málo vyšší než pravděpodobnost narození dívky). Proto zavádíme obecnější pravděpodobnostní prostory, ve kterých tyto požadavky vypouštíme. Tyto prostory najdou využití později při zavedení pojmu náhodné veličiny.

Diskrétní pravděpodobnostní prostor je zobecněním klasické pravděpodobnosti, který může mít spočetně mnoho elementárních jevů a kde se pravděpodobnosti jednotlivých elementárních jevů mohou lišit.

Definice 2.5 (Diskrétní pravděpodobnostní prostor)

Pravděpodobnostní prostor $(\Omega, \mathcal{A}, \mathbb{P})$ nazveme diskrétní pravděpodobnostní prostor, jestliže

- a) $\Omega = \{\omega_1, \omega_2, \dots\}$ má nejvýše spočetně mnoho prvků,
 b) $\mathcal{A} = \mathcal{P}(A)$ je potenční množina (tj. množina všech podmnožin),
 c) každému elementárnímu jevu $\omega_i \in \Omega$, $i \in \mathbb{N}$ je přiřazena pravděpodobnost $p_i = \mathbb{P}(\omega_i)$ a to tak, aby

$$\sum_{i=1}^{\infty} p_i = 1.$$

Pravděpodobnost libovolného jevu $A \in \mathcal{A}$ je tedy potom jednoznačně dána vztahem

$$\mathbb{P}(A) = \sum_{\omega_i \in A} p_i. \quad (2.3)$$

Podobně spojitý případ je zobecněním jednorozměrné geometrické pravděpodobnosti. V kapitole 9 o náhodných vektorech zavedeme vícerozměrnou geometrickou pravděpodobnost.

Definice 2.6 (Jednorozměrný spojitý pravděpodobnostní prostor)

Pravděpodobnostní prostor $(\Omega, \mathcal{A}, \mathbb{P})$ nazveme jednorozměrný spojitý pravděpodobnostní prostor, jestliže

- a) $\Omega = \mathbb{R}$,
 b) $\mathcal{A} = \mathcal{B}(\mathbb{R})$ je borelovská σ -algebra⁵ nad $\mathbb{R}$,
 c) je dána funkce $f: \mathbb{R} \rightarrow [0; +\infty]$ taková, že

$$\int_{\mathbb{R}} f(x) \, dx = 1.$$

Pravděpodobnost libovolného jevu $A \in \mathcal{A}$ je dána jednoznačně vztahem

$$\mathbb{P}(A) = \int_A f(x) \, dx. \quad (2.4)$$

⁵Viz příklad 2.2 na straně 22.

2.5 Úlohy

2.1 V urně jsou kuličky tří barev (jinak shodných fyzikálních vlastností). Nechtě jevy A , B , C postupně znamenají, že „náhodně vybraná kulička je černá“, „kulička je červená“, „kulička je bílá“. Určete význam následujících jevů:

- a) $\bar{A} \cap \bar{B}$,
- b) $(A \cup C) \cap B$,
- c) $(A \cap C) \cup B$,
- d) $A \cup B \cup C$.



2.2 Házíme jednou kostkou. Jev A znamená, že „při hodu padne číslo menší než 4“, a jev B , že „při hodu kostkou padne číslo menší než 5“. Pomocí jevů A , B , $\bar{A}$, $\bar{B}$ vyjádřete následující jevy:

- a) „při hodu kostkou padne číslo 4“,
- b) „při hodu kostkou padne číslo větší než 3“,
- c) „při hodu kostkou padne číslo menší nebo rovno 6“.

2.3 Při zkoušce z biologie dostane náhodně vybraný student tři otázky. Nechtě jev A znamená, že „student zodpoví správně první otázku“, jev B , že „student zodpoví správně druhou otázku“ a jev C , že „student zodpoví správně třetí otázku“. Vyjádřete pomocí jevů A , B , C , $\bar{A}$, $\bar{B}$, $\bar{C}$, že tento náhodně vybraný student:

- a) „zodpoví správně jen první otázku“,
- b) „zodpoví správně alespoň dvě otázky“,
- c) „zodpoví správně právě jednu otázku“,
- d) „zodpoví správně maximálně dvě otázky“.

2.4 Házíme třemi kostkami. Jaká je pravděpodobnost, že součet ok, které padnou na vrchních stěnách těchto tří kostek, je roven 5?

2.5 V urně je 10 lístků. Na šesti lístcích jsou dvojčíselná čísla a na čtyřech jednocíselná čísla. Vytáhneme dvakrát po jednom lístku a lístky už nevracíme zpět do urny. Vytažené lístky uložíme vedle sebe v pořadí, v jakém jsme je vytáhli. Jaká je pravděpodobnost, že takto vzniklé číslo je čtyřčíselné?

2.6 Mezi čtrnácti lidmi je osm s vysokoškolským vzděláním, zbylí mají středoškolské vzdělání. Náhodně vybereme čtyři lidi. Jaká je pravděpodobnost, že

- a) „všichni čtyři mají středoškolské vzdělání“ (jev A),
- b) „právě jeden má vysokoškolské vzdělání“ (jev B),
- c) „alespoň jeden má vysokoškolské vzdělání“ (jev C).

2.7 Manželé Novákovi chtějí mít čtyři děti a vědí, že pravděpodobnost narození chlapce je podle dlouhodobých statistik asi 0,514, narození dívky 0,486. Tchyně pana Nováka ovšem tvrdí, že s největší pravděpodobností budou mít Novákovi dva syny a dvě dcery. Rozhodněte, zda má tchyně pravdu. Zdůvodněte.

2.8 Hráč I vyhraje, jestliže při šesti hodech kostkou hodí alespoň jednu šestku. Hráč II vyhraje, jestliže hodí alespoň dvě šestky z dvanácti hodů kostkou. Který hráč má větší pravděpodobnost výhry?⁶

2.9 Čtyři jeleni byli chyceni z populace N jelenů, byli označeni a vypuštěni zpět. Abychom ověřili, že označení jeleni jsou náhodně rozmístěni v populaci, odchytli jsme pět jelenů po dostatečně dlouhé době ze stejné populace. Jaká je pravděpodobnost, že „právě jeden označený jelen bude znovu odchycen“ (jev A), jestliže $N = 8$ a $N = 30$? Sestavte bodový graf závislosti pravděpodobnosti $\mathbb{P}(A)$ na velikosti populace jelenů ($N = 8, 9, \dots, 60$).



2.10 Házíme n -krát po sobě dvěma kostkami. Jaká je pravděpodobnost, že „alespoň při jednom hoďu padne součet 12“?

2.11 Student si při zkoušce z matematiky tahá tři otázky ze 30. Ve 30 otázkách je 10 otázek z algebry, 15 z matematické analýzy a 5 z geometrie. Jaká je pravděpodobnost, že „si student vytáhne alespoň dvě otázky ze stejné disciplíny“?

2.12 Z urny, která obsahuje n kuliček, vytáhneme najednou několik kuliček. Určete, jaká je pravděpodobnost, že jsme vytáhli sudý počet kuliček? Před-

⁶Tento problém formuloval jako první Samuel Pepys (1633–1703), vysoký úředník anglického námořnictva a politik. Úlohu vyřešil v roce 1693 sir Isaac Newton (1643–1727).

pokládejme, že stejně pravděpodobná jsou možná tažení kuliček, tj. volím náhodně mezi všemi možnými tahy (rozlišitelných) kuliček.

2.13 V osudí je a lístků se sudými čísly a b lístků s lichými čísly. Jedním tahem vytáhneme k lístků. Vytažené lístky vložíme do druhého prázdného osudí a z něho pak vytáhneme jeden lístek. Jaká je pravděpodobnost, že „tento lístek je se sudým číslem“?

2.14 Ve třídě je n žáků.

- Jaká je pravděpodobnost, že se ve třídě najde alespoň jedna dvojice, která má narozeniny stejný den v roce?
- Jaká je pravděpodobnost, že ve třídě existuje alespoň jeden spolužák, který má narozeniny stejný den jako třídní profesor?
- Jaké je nejmenší n takové, aby pravděpodobnost, že dva žáci mají narozeniny ve stejný den, byla větší než $\frac{1}{2}$?



V úlohách neuvažujte přestupné roky, přítomnost dvojčat a předpokládejte, že se během celého roku děti rodí rovnoměrně.

2.15 Dva parníky, které používají jediné stejné přístaviště, mohou připlout kdykoliv během 24 hodin. Jejich příjezdy jsou nezávislé. První parník obsadí přístaviště na jednu hodinu, druhý na dvě hodiny. Jaká je pravděpodobnost, že ani jeden parník nebude muset čekat na uvolnění přístaviště?

2.16 Necht $x, y \in (0; 1)$ jsou náhodně zvolená čísla. Jaká je pravděpodobnost, že jejich součet je menší než 1 a součin větší než 0,09?

2.17 Na úsečce délky l jsou náhodně umístěny dva body X a Y , které náhodně rozdělují úsečku na tři části. S jakou pravděpodobností lze z takto vzniklých tří úseček sestavit trojúhelník?

Řešení úloh

2.1

- „vybraná kulička není černá a současně není červená“ neboli „vybraná kulička je bílá“,
- „vybraná kulička je černá nebo bílá a zároveň červená“, jev nemožný,

- c) „vybraná kulička je černá a zároveň bílá, nebo červená“ neboli „vybraná kulička je červená“,
 d) „vybraná kulička je černá, nebo červená, nebo bílá“, jev jistý.

2.2 a) $B \cap \bar{A} = B \setminus A$, b) $\bar{A}$, c) $B \cup \bar{B}$ nebo $A \cup \bar{A}$.

2.3

- a) $A \cap \bar{B} \cap \bar{C}$,
 b) $(A \cap B) \cup (B \cap C) \cup (A \cap C)$,
 c) $(A \cap \bar{B} \cap \bar{C}) \cup (\bar{A} \cap B \cap \bar{C}) \cup (\bar{A} \cap \bar{B} \cap C)$
 d) $(\bar{A} \cap \bar{B} \cap \bar{C}) \cup (A \cap \bar{B} \cap \bar{C}) \cup (\bar{A} \cap B \cap \bar{C}) \cup (\bar{A} \cap \bar{B} \cap C) \cup (A \cap B \cap \bar{C}) \cup (A \cap \bar{B} \cap C) \cup (\bar{A} \cap B \cap C) = (A \cap B \cap C)$.

2.4 $\frac{1}{36} = 0,02\bar{7}$

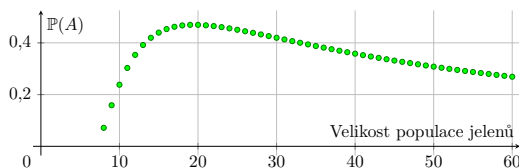
2.5 $\frac{1}{3} = 0,\bar{3}$

2.6 a) $\mathbb{P}(A) \doteq 0,0150$; b) $\mathbb{P}(B) \doteq 0,1598$; c) $\mathbb{P}(C) \doteq 0,985$.

2.7 Pravděpodobnost narození dvou synů a dvou dcer je po zaokrouhlení 0,374, což je nejvyšší pravděpodobnost, které může být při různých kombinacích počtu synů a dcer dosaženo. Tchyně má tedy pravdu.

2.8 Pravděpodobnost výhry hráče I je asi 0,665, pravděpodobnost výhry hráče II je asi 0,619. Hráč I má vyšší šanci na výhru.

2.9 Pro $N = 8$ je $\mathbb{P}(A) \doteq 0,071$, pro $N = 30$ je $\mathbb{P}(A) \doteq 0,42$. Bodový graf je na obrázku 2.3.



Obrázek 2.3: Bodový graf závislosti pravděpodobnosti $\mathbb{P}(A)$ na velikosti populace N jelenů

2.10 $1 - \left(\frac{35}{36}\right)^n$

2.11 0,815

$$2.12 \frac{2^{n-1} - 1}{2^n - 1}$$

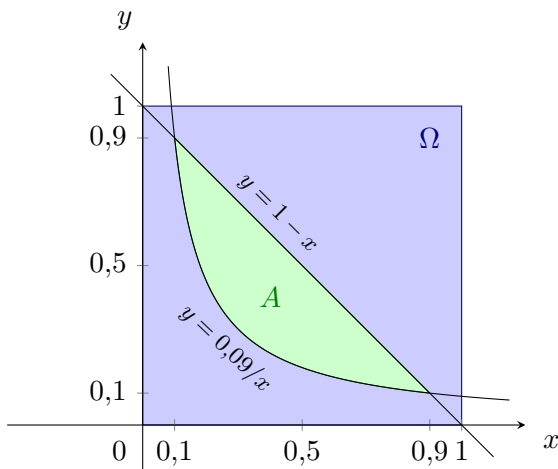
$$2.13 \frac{a}{a+b}$$

2.14

- a) $\mathbb{P}(A) = 1 - 1 \cdot \left(1 - \frac{1}{365}\right) \cdot \left(1 - \frac{2}{365}\right) \cdots \left(1 - \frac{n-1}{365}\right) =$
 $= 1 - \frac{365 \cdot 364 \cdots (365-n+1)}{365^n}$.
- b) $\mathbb{P}(B) = 1 - \left(\frac{364}{365}\right)^n$.
- c) $\mathbb{P}(A) > \frac{1}{2} \Leftrightarrow n > 23$, tedy alespoň 24 žáků.

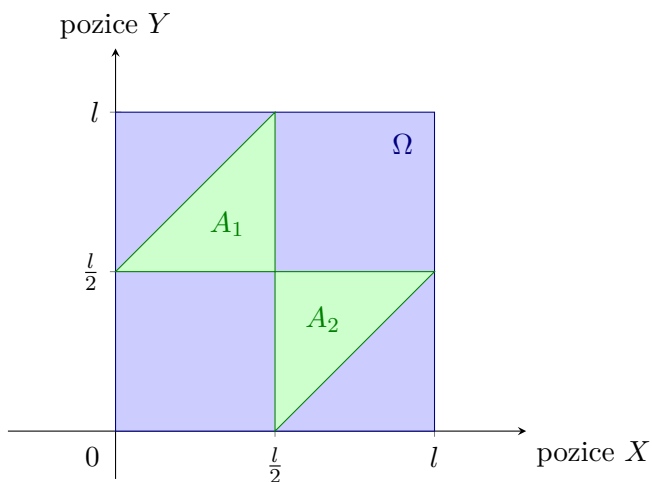
2.15 $\mathbb{P} \doteq 0,878$.

$$2.16 \mathbb{P}(A) = \frac{\mu^2(A)}{\mu^2(\Omega)} = \frac{\int_{0,1}^{0,9} \left(1 - x - \frac{0,09}{x}\right) dx}{1} \doteq 0,202.$$



Obrázek 2.4: Množina všech možných výsledků a výsledků příznivých jevu A v úloze 2.16

$$2.17 \mathbb{P}(A) = \frac{\mu^2(A)}{\mu^2(\Omega)} = \frac{l^2/4}{l^2} = \frac{1}{4}.$$



Obrázek 2.5: Množina všech možných výsledků a výsledků příznivých jevu $A = A_1 \cup A_2$ v úloze 2.17

Kapitola 3

Podmíněná pravděpodobnost, nezávislost

3.1 Podmíněná pravděpodobnost

Pravděpodobnostní prostor $(\Omega, \mathcal{A}, \mathbb{P})$ je matematickou reprezentací všech možných výsledků náhodného pokusu spolu s informací o pravděpodobnosti jejich výskytu. Je běžné, že v průběhu času může dojít k nějaké aktualizaci informace o tomto pokusu, čímž může dojít i ke změně pravděpodobnosti, že nastane určitý jev A . Teorie pravděpodobnosti disponuje nástrojem k popsání této situace – nazývá se *podmíněná pravděpodobnost*.

Předpokládejme, že zkoumáme pravděpodobnost výskytu náhodného jevu A nad pravděpodobnostním prostorem $(\Omega, \mathcal{A}, \mathbb{P})$. Nechť je nově k dispozici dodatečná informace o tom, že nastal nějaký jiný jev B s nenulovou pravděpodobností $\mathbb{P}(B) > 0$. Situaci můžeme ilustrovat na následujícím příkladu.

Příklad 3.1. Házejme postupně dvěma kostkami a zkoumejme náhodný jev $A =$ „součet ok na obou kostkách je nejvýše 5“. Prostor elementárních jevů je stejný jako v příkladu 2.4

$$\Omega = \{(\square, \square), (\square, \blacksquare), \dots, (\blacksquare, \blacksquare), (\blacksquare, \blacksquare)\}$$

a pravděpodobnost jevu A před zahájením pokusu činí

$$\mathbb{P}(A) = \frac{10}{36} = 0,2\bar{7},$$

neboť 10 z 36 možných výsledků je jevu A příznivých (nechť si je čtenář sám vyčíslí).

Nyní zkoumejme situaci po prvním ze dvou hodů. Dejme tomu, že v prvním hodu na kostce padla $\blacksquare$. Tuto informaci reprezentujeme nad prostorem $(\Omega, \mathcal{A}, \mathbb{P})$ jako jev

$$B = \{(\blacksquare, \blacksquare), (\blacksquare, \blacksquare), \dots, (\blacksquare, \blacksquare)\}.$$

Pravděpodobnost výskytu tohoto jevu je zřejmě $\mathbb{P}(B) = 6/36$. Jaká je nyní v této nové situaci pravděpodobnost výskytu jevu A ? Z deseti uspořádaných dvojic množiny A jich mají $\blacksquare$ jako první prvek celkem čtyři dvojice tvořící podmnožinu Ω :

$$\{(\blacksquare, \blacksquare), (\blacksquare, \blacksquare), (\blacksquare, \blacksquare), (\blacksquare, \blacksquare)\}.$$

Všimněte si, že tato množina je formálně průnikem množin A a B , tj. $A \cap B$, a její původní (před prvním hodem spočtená) pravděpodobnost je $4/36$. Protože však již víme, že jev B nastal, je žádoucí pravděpodobnost jevu A aktualizovat tak, aby vzala v úvahu novou informaci o realizaci jevu B . Její nová hodnota má činit $4/6$ – pro druhý hod připadá v úvahu šest možných výsledků a čtyři z nich jsou příznivé jevu A . S využitím původních informací o pravděpodobnosti ji můžeme dostat jako podíl

$$\frac{4/36}{6/36} = \frac{2}{3} = 0,\bar{3},$$

tedy jako podíl pravděpodobnosti průniku jevů $A \cap B$ a pravděpodobnosti podmínky B . Vidíme tedy, že pravděpodobnost výskytu jevu A se po prvním ze dvou hodů zvýšila z původních přibližných 28 % na asi 33 %.

V tomto a podobných případech ve skutečnosti nehovoříme o „aktualizované“ pravděpodobnosti jevu A , nýbrž o pravděpodobnosti jevu A podmíněné jevem B nebo též o podmíněné pravděpodobnosti jevu A za podmínky B , kterou značíme $\mathbb{P}(A|B)$.

Definice 3.1 (Podmíněná pravděpodobnost)

Nechť je dán pravděpodobnostní prostor $(\Omega, \mathcal{A}, \mathbb{P})$ a náhodné jevy A, B , kde $\mathbb{P}(B) > 0$. Podmíněnou pravděpodobnost jevu A za podmínky, že nastal jev B , definujeme vztahem

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \quad (3.1)$$

Následující věta říká, že podmíněná pravděpodobnost má všechny základní vlastnosti pravděpodobnosti nepodmíněné dle definice 2.2, a tudíž je to také pravděpodobnost.

Věta 3.1. *Nechť je dán pravděpodobnostní prostor $(\Omega, \mathcal{A}, \mathbb{P})$ a náhodný jev B , kde $\mathbb{P}(B) > 0$. Pak pro libovolný jev $A \in \mathcal{A}$ platí:*

- a) (nezápornost) $\mathbb{P}(A|B) \geq 0$,
 b) (σ -aditivita) pro každou posloupnost $\{A_n\}_{n=1}^{\infty}$ disjunktních jevů platí

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n | B\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n | B),$$

- c) (normovanost) $\mathbb{P}(\Omega|B) = 1$.

Důkaz.

- a) $\mathbb{P}(A|B) \geq 0$, neboť podle (3.1) jde o podíl nezáporného a kladného čísla.
 b) Protože $A_1, A_2, \dots$ jsou disjunktní, tak i $A_1 \cap B, A_2 \cap B, \dots$ jsou disjunktní. Z axiomu b) definice 2.2 a z definice 3.1 plyne

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n | B\right) = \frac{\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n \cap B\right)}{\mathbb{P}(B)} = \frac{\sum_{n=1}^{\infty} \mathbb{P}(A_n \cap B)}{\mathbb{P}(B)} = \sum_{n=1}^{\infty} \mathbb{P}(A_n | B).$$

- c) Z definice 3.1 plyne, že

$$\mathbb{P}(\Omega|B) = \frac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1.$$

□

Velmi často jsou v definici 3.1 známy podmíněné i nepodmíněné pravděpodobnosti, což nám umožní ze vztahu (3.1) odvodit pravděpodobnost, že dva jevy nastaly současně:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B) \cdot \mathbb{P}(B). \quad (3.2)$$

Příklad 3.2 (o vadných pojistkách). Mějme krabici s 20 tepelnými pojistkami, o nichž víme, že nedopatřením obsahuje 5 vadných. Potřebujeme opravit dva přístroje výměnou jejich pojistek. Postupně tedy z krabice vyndáme dvě pojistky (bez vracení). Jaká je pravděpodobnost, že budou obě dvě vadné?



Řešení. Označme

A = „druhá pojistka vadná“

B = „první pojistka vadná“

$A \cap B$ = „obě pojistky jsou vadné“

Známe podmíněnou pravděpodobnost jevu $A|B$ a nepodmíněnou pravděpodobnost jevu B ,

$$\mathbb{P}(B) = \frac{5}{20}, \quad \mathbb{P}(A|B) = \frac{4}{19},$$

obě plynou z jednoduché kombinatorické úvahy. Dosazením do vztahu (3.2) dostáváme

$$\mathbb{P}(A \cap B) = \frac{4}{19} \cdot \frac{5}{20} = \frac{1}{19}.$$

△

Toto jednoduché pravidlo lze zobecnit na více jevů, jak je formulováno ve větě 3.2.

Věta 3.2 (o násobení pravděpodobností, řetězové pravidlo). *Jestliže pro libovolnou posloupnost náhodných jevů $A_1, A_2, \dots, A_n$ platí podmínka $\mathbb{P}(A_1 \cap A_2 \cap \dots \cap A_{n-1}) > 0$, potom*

$$\begin{aligned} \mathbb{P}\left(\bigcap_{i=1}^n A_i\right) &= \mathbb{P}(A_1) \cdot \mathbb{P}(A_2|A_1) \cdot \mathbb{P}(A_3|A_1 \cap A_2) \cdots \\ &\quad \cdots \mathbb{P}(A_n|A_1 \cap A_2 \cap \dots \cap A_{n-1}). \end{aligned} \quad (3.3)$$

Důkaz. Vzhledem k monotonii pravděpodobnosti a předpokladu věty máme

$$\mathbb{P}(A_1) \geq \mathbb{P}(A_1 \cap A_2) \geq \dots \geq \mathbb{P}(A_1 \cap \dots \cap A_{n-1}) > 0,$$

a tedy všechny podmíněné pravděpodobnosti v tvrzení věty jsou dobře definovány. Opakovaným použitím vzorce (3.2) dostáváme

$$\begin{aligned} \mathbb{P}\left(\bigcap_{i=1}^{n-1} A_i \cap A_n\right) &= \mathbb{P}\left(\bigcap_{i=1}^{n-1} A_i\right) \mathbb{P}\left(A_n \mid \bigcap_{i=1}^{n-1} A_i\right) = \\ &= \mathbb{P}\left(\bigcap_{i=1}^{n-2} A_i\right) \mathbb{P}\left(A_{n-1} \mid \bigcap_{i=1}^{n-2} A_i\right) \mathbb{P}\left(A_n \mid \bigcap_{i=1}^{n-1} A_i\right) = \\ &= \dots = \\ &= \mathbb{P}(A_1) \mathbb{P}(A_2|A_1) \mathbb{P}(A_3|A_1 \cap A_2) \cdots \mathbb{P}\left(A_n \mid \bigcap_{i=1}^{n-1} A_i\right). \end{aligned}$$

□

Příklad 3.3. Z urny, ve které je n bílých a n černých kuliček, náhodně vybereme n -krát po dvou kuličkách bez vracení vytažených kuliček. Určete, jaká je pravděpodobnost, že vždy vytáhneme jednu bílou a jednu černou kuličku.

Řešení. Označme A_i jev „v i -tém tahu vytáhneme jednu kuličku bílou a jednu kuličku černou“, $i = 1, \dots, n$. Hledáme pravděpodobnost náhodného jevu

$$A_1 \cap A_2 \cap \dots \cap A_n = \bigcap_{i=1}^n A_i;$$

ta je dána vzorcem (3.3) ve větě 3.2. Hledejme pravděpodobnosti vyskytující se v rovnici (3.3). Nejprve spočteme $\mathbb{P}(A_1)$: počet možných případů je $\binom{2n}{2}$ a počet příznivých případů jevu A_1 je $n \cdot n$, z čehož plyne, že

$$\mathbb{P}(A_1) = \frac{n \cdot n}{\binom{2n}{2}}.$$

Při hledání pravděpodobnosti $\mathbb{P}(A_2|A_1)$ vycházíme z toho, že v urně je již pouze $n - 1$ bílých a $n - 1$ černých kuliček, takže

$$\mathbb{P}(A_2|A_1) = \frac{(n-1)(n-1)}{\binom{2n-2}{2}}.$$

Analogicky

$$\mathbb{P}(A_3|A_1 \cap A_2) = \frac{(n-2)(n-2)}{\binom{2n-4}{2}}$$

$$\vdots$$

$$\mathbb{P}\left(A_n \mid \bigcap_{i=1}^{n-1} A_i\right) = 1.$$

Hledaná pravděpodobnost je tedy rovna

$$\mathbb{P}\left(\bigcap_{i=1}^n A_i\right) = \frac{n \cdot n}{\binom{2n}{2}} \cdot \frac{(n-1)(n-1)}{\binom{2n-2}{2}} \cdot \frac{(n-2)(n-2)}{\binom{2n-4}{2}} \dots 1 = \frac{n!n!2^n}{(2n)!}.$$

△

3.2 Nezávislost

Uvažujme nyní dva náhodné jevy A a B . Jestliže výskyt jevu B neovlivní nijak pravděpodobnost jevu A , tedy jestliže

$$\mathbb{P}(A|B) = \mathbb{P}(A), \quad (3.4)$$

lze mluvit o tom, že jev A je na jevu B nezávislý. Dosazením (3.4) do vzorce (3.2) dostáváme, že v tom případě musí platit

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B). \quad (3.5)$$

Obráceným postupem lze ze vzorce (3.5) odvodit, že vztah (3.4) platí symetricky i pro jev B :

$$\mathbb{P}(B|A) = \mathbb{P}(B). \quad (3.6)$$

Oba vzorce (3.4) a (3.6) však kvůli definici podmíněné pravděpodobnosti vyžadují, aby

$$\mathbb{P}(A) > 0 \text{ a } \mathbb{P}(B) > 0.$$

Vzorec (3.5) však funguje dobře i bez tohoto požadavku, navíc dobře reflektuje i symetrii pojmu nezávislost, tudíž mu dnes dáváme přednost při definování nezávislosti dvou náhodných jevů.

Definice 3.2 (Nezávislé náhodné jevy)

Náhodné jevy A a B jsou (vzájemně¹) nezávislé, jestliže platí

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B). \quad (3.7)$$

Příklad 3.4. Jaká je pravděpodobnost, že ve dvou hodech hrací kostkou padne v prvním hodu číslo sudé (jev A) a ve druhém hodu padne prvočíslo (jev B)?

Řešení. Množinu Ω tvoří všechny možné uspořádané dvojice sestavené postupně z výsledků obou házení hrací kostkou. Pravděpodobnost jevů A a B je stejná, $\mathbb{P}(A) = \mathbb{P}(B) = \frac{1}{2}$ (prvočísla jsou 2, 3 a 5, číslo 1 podle definice není prvočíslo). Protože výsledek druhého hodu není ovlivněn výsledkem hodu předchozího, oba sledované jevy jsou navzájem nezávislé a tedy pravděpodobnost jejich současného výskytu je

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} = 0,25.$$

△

¹Slovo *vzájemně* se často vynechává a mluvíme pouze o nezávislosti jevů.

Pojem nezávislosti můžeme rozšířit i na skupinu náhodných jevů, ovšem několika způsoby.

Definice 3.3 (Po dvou nezávislé náhodné jevy)

Nechť $A_1, \dots, A_n$ jsou náhodné jevy. Řekneme, že jsou po dvou nezávislé, jestliže jevy A_i, A_j jsou nezávislé pro všechna $i, j = 1, \dots, n$, $i \neq j$.

Příklad 3.5. Při hodu dvěma mincemi uvažujeme tyto náhodné jevy: A_1 je jev „na 1. minci padne panna“, A_2 je jev „na 2. minci padne orel“, A_3 je jev „na obou mincích padne stejný obrazec“. Zjistěte, zda jsou dané jevy po dvou nezávislé.



Řešení.

$$\begin{aligned} \mathbb{P}(A_1) &= \frac{1}{2}, & \mathbb{P}(A_2) &= \frac{1}{2}, & \mathbb{P}(A_3) &= \frac{1}{2}, \\ \mathbb{P}(A_1 \cap A_2) &= \frac{1}{4}, & \mathbb{P}(A_1 \cap A_3) &= \frac{1}{4}, & \mathbb{P}(A_2 \cap A_3) &= \frac{1}{4}, \\ \mathbb{P}(A_1) \cdot \mathbb{P}(A_2) &= \frac{1}{4}, & \mathbb{P}(A_1) \cdot \mathbb{P}(A_3) &= \frac{1}{4}, & \mathbb{P}(A_2) \cdot \mathbb{P}(A_3) &= \frac{1}{4}, \end{aligned}$$

Protože pro všechny tři dvojice jevů A_1, A_2, A_3 platí rovnost

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i) \cdot \mathbb{P}(A_j), \quad i \neq j,$$

jsou tyto jevy po dvou nezávislé. △

Definice 3.4 (Skupinově nezávislé náhodné jevy)

Nechť $A_1, A_2, \dots, A_n$ jsou náhodné jevy. Řekneme, že jsou skupinově (totálně) nezávislé, jestliže pro libovolnou posloupnost indexů $\{k_1, k_2, \dots, k_r\} \subset \{1, \dots, n\}$, $r = 2, \dots, n$ platí

$$\mathbb{P}(A_{k_1} \cap A_{k_2} \cap \dots \cap A_{k_r}) = \mathbb{P}(A_{k_1}) \cdot \mathbb{P}(A_{k_2}) \cdots \mathbb{P}(A_{k_r}). \quad (3.8)$$

Příklad 3.6 (pokračování příkladu 3.5). Zjistěte, zda jsou jevy z příkladu 3.5 skupinově nezávislé.

Řešení. $\mathbb{P}(A_1 \cap A_2 \cap A_3) = 0$, neboť $A_1 \cap A_2 \cap A_3$ je jev nemožný. Protože $\mathbb{P}(A_1) \cdot \mathbb{P}(A_2) \cdot \mathbb{P}(A_3) \neq 0$, nejsou jevy A_1, A_2, A_3 skupinově nezávislé, přestože jsou nezávislé po dvou. △

Dále uvedeme některá tvrzení o nezávislosti náhodných jevů.

Věta 3.3 (o nezávislosti jevů).

- a) Jev nemožný a libovolný náhodný jev A jsou nezávislé.
- b) Jev jistý a libovolný náhodný jev A jsou nezávislé.
- c) Necht jsou jevy A, B disjunktní. Pak jsou jevy A, B nezávislé právě tehdy, když $\mathbb{P}(A) \cdot \mathbb{P}(B) = 0$.

Důkaz.

$$a) \left. \begin{aligned} \mathbb{P}(\emptyset \cap A) &= \mathbb{P}(\emptyset) = 0 \\ \mathbb{P}(\emptyset) \cdot \mathbb{P}(A) &= 0 \cdot \mathbb{P}(A) = 0 \end{aligned} \right\} \Rightarrow \mathbb{P}(\emptyset \cap A) = \mathbb{P}(\emptyset) \cdot \mathbb{P}(A).$$

$$b) \mathbb{P}(\Omega \cap A) = \mathbb{P}(A) = \mathbb{P}(A) \cdot 1 = \mathbb{P}(A) \cdot \mathbb{P}(\Omega).$$

- c) Důkaz ekvivalence provedeme ověřením jednotlivých implikací. Implikace zleva doprava: pro disjunktní jevy A, B platí $A \cap B = \emptyset$. Jelikož předpokládáme, že jsou nezávislé, platí

$$\mathbb{P}(A) \cdot \mathbb{P}(B) = \mathbb{P}(A \cap B) = \mathbb{P}(\emptyset) = 0.$$

Důkaz implikace zprava doleva: A a B jsou disjunktní, tedy $\mathbb{P}(A \cap B) = \mathbb{P}(\emptyset) = 0$. Jelikož dle předpokladu je rovněž $\mathbb{P}(A) \cdot \mathbb{P}(B) = 0$, jsou A, B nutně také nezávislé. □

Věta 3.4. Necht A, B jsou nezávislé náhodné jevy. Pak dvojice jevů $(A, \bar{B})$, $(\bar{A}, B)$, $(\bar{A}, \bar{B})$ jsou nezávislé.

Důkaz.

$$\begin{aligned} \mathbb{P}(\bar{A} \cap B) &= \mathbb{P}(B - A) = \mathbb{P}(B - (A \cap B)) = \mathbb{P}(B) - \mathbb{P}(A \cap B) = \\ &= \mathbb{P}(B) - \mathbb{P}(B) \cdot \mathbb{P}(A) = \mathbb{P}(B) \cdot (1 - \mathbb{P}(A)) = \mathbb{P}(B) \cdot \mathbb{P}(\bar{A}). \end{aligned}$$

Nezávislost jevů $A, \bar{B}$ se dokáže analogicky.

Jsou-li A, B nezávislé jevy, pak podle předchozího jsou nezávislé jevy $A, \bar{B}$, ale odtud opět podle předchozího jsou nezávislé i jevy $\bar{A}, \bar{B}$. □

Věta 3.5. Necht $A_1, \dots, A_n$ jsou skupinově (totálně) nezávislé jevy. Pak platí

$$\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = 1 - \prod_{i=1}^n [1 - \mathbb{P}(A_i)].$$

Důkaz. Z de Morganových vzorců pro operace s množinami plyne

$$\bigcup_{i=1}^n A_i = \Omega - \bigcap_{i=1}^n \bar{A}_i,$$

tedy

$$\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = \mathbb{P}(\Omega) - \mathbb{P}\left(\bigcap_{i=1}^n \bar{A}_i\right) = 1 - \mathbb{P}\left(\bigcap_{i=1}^n \bar{A}_i\right).$$

Protože $A_1, \dots, A_n$ jsou nezávislé, jsou i $\bar{A}_1, \dots, \bar{A}_n$ nezávislé, takže

$$\mathbb{P}\left(\bigcap_{i=1}^n \bar{A}_i\right) = \prod_{i=1}^n \mathbb{P}(\bar{A}_i) = \prod_{i=1}^n (1 - \mathbb{P}(A_i)).$$

□

Příklad 3.7. Během dne se v porodnici narodilo deset dětí. Pravděpodobnost narození chlapce je $p = 0,514$. Jaká je pravděpodobnost, že se během tohoto dne narodil v porodnici alespoň jeden chlapec?



Řešení. Nechť A_i , $i = 1, \dots, 10$, značí jev „ i -té narozené dítě je chlapec“. Jev $\bigcup_{i=1}^{10} A_i$ znamená „alespoň jedno narozené dítě je chlapec“. Jelikož $A_1, \dots, A_{10}$ jsou skupinově nezávislé, je podle věty 3.5

$$\mathbb{P}\left(\bigcup_{i=1}^{10} A_i\right) = 1 - \prod_{i=1}^{10} (1 - \mathbb{P}(A_i)) = 1 - 0,486^{10} \doteq 0,999\,265.$$

△

Věta 3.6. Náhodné jevy A , B , kde $0 < \mathbb{P}(B) < 1$, jsou nezávislé právě tehdy, když

$$\mathbb{P}(A|B) = \mathbb{P}(A|\bar{B}).$$

Důkaz. Rovnost

$$\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A \cap \bar{B})}{\mathbb{P}(\bar{B})}$$

platí právě tehdy, když $\mathbb{P}(\bar{B}) \cdot \mathbb{P}(A \cap B) = \mathbb{P}(\bar{A} \cap B) \cdot \mathbb{P}(B)$. Po dosazení za $\mathbb{P}(\bar{B}) = 1 - \mathbb{P}(B)$ dostáváme $\mathbb{P}(A \cap B) - \mathbb{P}(B) \cdot \mathbb{P}(A \cap B) = \mathbb{P}(B) \cdot \mathbb{P}(A \cap \bar{B})$ a po úpravě $\mathbb{P}(A \cap B) = \mathbb{P}(B) \cdot (\mathbb{P}(A \cap B) + \mathbb{P}(A \cap \bar{B})) = \mathbb{P}(B) \cdot \mathbb{P}(A)$. □

3.3 Bernoulliho² proces

V úvodních kapitolách jsme mluvili o náhodných pokusech jako o jednorázových experimentech, při nichž s určitou pravděpodobností očekáváme výskyt právě jednoho z možných výsledků. Nyní budeme zkoumat opakované pokusy, tj. takové pokusy, kdy jeden a týž pokus opakujeme vícekrát za sebou. Speciálním případem je opakování pokusů s dvěma alternativními navzájem se vylučujícími se výsledky, například 0–1, chlapec–dívka, tažení z urny s černými a bílými koulemi apod. Jednotlivé pokusy tohoto typu se nazývají *Bernoulliho pokusy*, jejichž základní rys je ten, že výsledky pokusu jsou charakterizovány pouze dvěma navzájem se vylučujícími znaky. Jestliže jsou navíc jednotlivé pokusy na sobě nezávislé (výsledek jednoho pokusu není ovlivněn výsledky pokusů předcházejících), jejich posloupnost se nazývá *Bernoulliho proces*. V kapitole 7 se budeme zabývat náhodnými veličinami, které s Bernoulliho procesem souvisí (např. alternativní rozdělení v odst. 7.3, binomické v odst. 7.4 nebo geometrické v odst. 7.5). Bernoulliho proces se dá znázornit pomocí tzv. *Bernoulliho schématu*, které si nyní představíme.



Příklad 3.8 (Bernoulliho schéma). V osudí je b koulí bílých a c koulí černých. n -krát po sobě vytáhneme vždy po jedné kouli, přičemž tuto kouli vrátíme po tahu vždy zpět. Jaká je pravděpodobnost, že mezi vytaženými koulemi je právě k koulí bílých a $(n - k)$ koulí černých?

Řešení. V osudí máme při každém opakovaném tahu neustále b koulí bílých a c koulí černých. V každém z n tahů máme tedy $(b + c)$ možností, jak vytáhnout jednu kouli. Protože každá možnost prvního tahu může být zkombinována s každou možností druhého tahu a tyto možnosti se dají zkombinovat s každou možností tahu třetího, ..., je počet možných případů konaného pokusu $(b + c)^n$.

Nyní budeme řešit počet příznivých případů. V n tazích máme vytáhnout k koulí bílých, tzn. rozmístujeme k koulí na n míst. Takovýchto rozmístění je $\binom{n}{k}$. Budeme-li uvažovat jedno pevné rozmístění, například nejprve bude vytaženo k bílých koulí a pak $(n - k)$ černých koulí, vidíme, že takové rozmístění můžeme vytáhnout $b^k \cdot c^{n-k}$ způsoby. Tedy počet příznivých případů

²Jacob Bernoulli (1655–1705), švýcarský matematik a fyzik.

je $\binom{n}{k} b^k c^{n-k}$. Hledaná pravděpodobnost $P_{k,n}$ je tedy rovna

$$P_{k,n} = \binom{n}{k} \frac{b^k \cdot c^{n-k}}{(b+c)^n},$$

což není nic jiného než

$$P_{k,n} = \binom{n}{k} \left(\frac{b}{b+c} \right)^k \left(\frac{c}{b+c} \right)^{n-k}. \quad (3.9)$$

Označme si $p = \frac{b}{b+c}$, $q = \frac{c}{b+c}$; zde tedy p značí pravděpodobnost, že v jednotlivém tahu vytáhneme bílou kouli, a q značí pravděpodobnost, že v jednotlivém tahu vytáhneme černou kouli. Pak (3.9) můžeme psát ve tvaru

$$P_{k,n} = \binom{n}{k} p^k q^{n-k}, \quad \text{kde } q = 1 - p, \quad p \in (0; 1). \quad (3.10)$$

△

U mnohonásobného nezávisle opakovaného alternativního pokusu počítáme pravděpodobnost, že v n nezávislých pokusech náhodný jev A nastane k -krát, podle (3.10), přičemž v každém pokusu náhodný jev A nastane s pravděpodobností p . Pravděpodobnost, že v n nezávislých pokusech nastane náhodný jev A alespoň k -krát, počítáme podle vzorce

$$S_{k,n} = \sum_{i=k}^n P_{i,n}. \quad (3.11)$$

Z (3.11) dále speciálně pro $k = 1$ dostáváme

$$S_{1,n} = \sum_{i=1}^n P_{i,n} = 1 - P_{0,n} = 1 - q^n, \quad (3.12)$$

což je pravděpodobnost, že v n nezávislých pokusech nastane náhodný jev A alespoň jednou.

Vraťme se nyní k příkladu 3.7, který jsme řešili pomocí věty 3.5. Tento příklad můžeme vyřešit i pomocí výše uvedeného Bernoulliho schématu, jestliže uvažujeme porod za Bernoulliho pokus. V příkladu 3.7 tedy máme deset nezávislých Bernoulliho pokusů; výsledek každého takového pokusu je buď narození chlapce nebo dívky. Pravděpodobnost narození chlapce je při každém porodu 0,514; podle (3.12) tedy dostáváme

$$S_{1,n} = 1 - (0,486)^{10} \doteq 0,999\,265.$$

Vidíme, že Bernoulliho schéma je poměrně efektivní metodou při řešení podobných úloh.

Příklad 3.9. Házíme n -krát po sobě jednou hrací kostkou. Vypočtete pravděpodobnost, že v těchto n hodech padne alespoň jednou šestka.

Řešení. Máme opět n nezávislých pokusů. Výsledek každého pokusu je buď šestka s pravděpodobností $p = \frac{1}{6}$, nebo s pravděpodobností $q = \frac{5}{6}$ některý ze zbývajících výsledků. Podle (3.12) dostáváme

$$S_{1,n} = 1 - \left(\frac{5}{6}\right)^n,$$

což je pravděpodobnost, se kterou v n hodech hrací kostkou padne alespoň jednou šestka. $\triangle$



Často nás spíše než pravděpodobnost zajímá počet pokusů n , které musíme vykonat, abychom s pravděpodobností rovnou nejméně P mohli tvrdit, že náhodný jev A nastal alespoň jednou, přičemž pravděpodobnost, že náhodný jev A nastane v každém jednotlivém pokusu, je rovna p . Neboli požadujeme, aby

$$P \leq S_{1,n},$$

odkud po úpravách dostaneme

$$\begin{aligned} P &\leq 1 - (1 - p)^n, \\ \ln(1 - P) &\geq n \ln(1 - p), \end{aligned}$$

a tedy

$$n \geq \frac{\ln(1 - P)}{\ln(1 - p)}. \quad (3.13)$$

Příklad 3.10. Hráč podává v sázkové kanceláři 30 tiketů Sportky. Pravděpodobnost, že vyhraje 5. cenu (uhodne tři čísla ze 49, vyplňuje pouze jeden sloupec), je $p = 0,017\,65$. Určete počet tiketů, které musí hráč podat, aby mezi nimi byl alespoň jeden vyhrávající 5. cenu s pravděpodobností rovnající se alespoň 0,75.

Řešení. Výpočet provedeme podle (3.13), kde $p = 0,017\,65$, $P = 0,75$:

$$n \geq \frac{\ln(1 - 0,75)}{\ln(1 - 0,017\,65)} = \frac{\ln 0,25}{\ln 0,982\,35} \doteq 77,8.$$

Hráč musí podat 78 tiketů Sportky, aby s pravděpodobností alespoň 0,75 vyhrál alespoň jednu 5. cenu. $\triangle$



3.4 Úlohy

3.1 V urně je šest kuliček. Každá kulička má jinou barvu (červená, bílá, černá, žlutá, modrá, zelená). Budeme nezávisle pětkrát po sobě losovat jednu kuličku, přičemž po každém tahu kuličku vracíme zpět do urny. Jaká je pravděpodobnost, že

- „při 1. a 5. tahu vytáhneme kuličku žluté barvy“ (A),
- „kuličku žluté barvy vytáhneme právě dvakrát“ (B)?

3.2 Stroj sestává ze dvou součástek, které jsou zapojeny

- sériově,
- paralelně,

a pracují na sobě nezávisle. Pravděpodobnost poruchy první součástky je 0,04, druhé 0,03. Zařízení se sériově zapojenými součástkami je funkční, pracují-li obě součástky současně bez problémů. Zařízení s paralelně zapojenými součástkami

je funkční, pracuje-li bez problémů alespoň jedna součástka. Jaká je při jednotlivých zapojeních pravděpodobnost, že

- „stroj pracuje“ (A),
- „stroj nepracuje“ ($\bar{A}$)?



3.3 Stroj sestává ze tří součástek, které jsou všechny navzájem zapojeny paralelně a pracují na sobě nezávisle. Pravděpodobnost funkčnosti první součástky je $\mathbb{P}(A_1)$, druhé $\mathbb{P}(A_2)$ a třetí $\mathbb{P}(A_3)$. Zařízení je funkční, pracuje-li alespoň jedna součástka bez problémů. Obecně vyjádřete pravděpodobnost $\mathbb{P}(A)$ jevu A , že „stroj bude pracovat bez problémů“.

3.4 Dvacetkrát hodíme hrací kostkou. Jaká je pravděpodobnost, že

- „ve všech dvaceti hodech padne stěna s nejvýše 4 puntíky“ (jev A),
- „až při dvacátém hodu nepadne stěna s nejvýše 4 puntíky“ (jev B),
- „právě desetkrát padne stěna s nejvýše 4 puntíky“ (jev C).

3.5 Dvakrát po sobě vytáhneme jednu kartu z balíčku 32 karet, přičemž první vytaženou kartu do balíčku zpět nevrátíme. Jaká je pravděpodobnost, že:

- a) „v druhém tahu vytáhneme libovolného krále za předpokladu, že jsme v prvním tahu vytáhli také libovolného krále“,
- b) „v druhém tahu nevytáhneme krále za předpokladu, že jsme v prvním tahu vytáhli libovolného krále“,
- c) „v druhém tahu vytáhneme krále za předpokladu, že jsme v prvním tahu nevytáhli libovolného krále“,
- d) „v druhém tahu nevytáhneme krále za předpokladu, že jsme ho ani v prvním tahu nevytáhli“.



3.6 Dva sportovci hází oštěpem nezávisle jeden na druhém. Každý má pouze jeden pokus. Pravděpodobnost, že první hodí 80 metrů, je 0,8. Pravděpodobnost, že druhý hodí 80 metrů, je 0,75. Jaká je pravděpodobnost, že 80 metrů nehodí ani jeden z nich?



3.7 Profesor zapomene deštník při každé návštěvě obchodu s pravděpodobností $\frac{1}{4}$. Jestliže navštívil čtyři obchody a přišel domů bez deštníku, jaká je pravděpodobnost, že jej zapomněl v posledním obchodě?



3.8 Je dáno: $\mathbb{P}(A) = 0,4$, $\mathbb{P}(B) = 0,5$
a $\mathbb{P}(\bar{A} \cap \bar{B}) = 0,3$.

- a) Jsou jevy A a B nezávislé?
- b) Vypočítejte $\mathbb{P}(A \cup B)$, $\mathbb{P}(A \cap B)$, $\mathbb{P}(\bar{A} \cup \bar{B})$ a $\mathbb{P}(A \cup \bar{B})$.
- c) Vypočítejte $\mathbb{P}(A|\bar{B})$ a $\mathbb{P}(\bar{B}|\bar{A})$.

3.9 Uvažujeme dvě osudí. V každém jsou obsaženy čtyři lístky. Na dvou lístcích jsou čísla dělitelná dvěma a na druhých dvou jsou čísla dělitelná třemi, přičemž na žádném z lístků není číslo dělitelné šesti. Nechť A je takový jev, že z prvního osudí vytáhneme lístek, na němž je číslo dělitelné dvěma; nechť B je takový jev, že z druhého osudí vytáhneme lístek s číslem dělitelným třemi, a C je jev, že v obou dvou tazích vytáhneme lístky s čísly

dělitelnými tímž číslem. Zjistěte, zda jevy A , B , C jsou nezávislé po dvou a po třech.

3.10 Dětská obrázková skládačka obsahuje dvanáct kostek, které mají na každé stěně část jednoho ze šesti obrázků. Aby se obrázek správně složil z dílčích obrázků na jednotlivých stěnách kostek, musí být každá kostka na správném místě, musí být kostka otočena správnou stěnou vzhůru a obrázek natočený správným směrem. Necht' jev A znamená, že „každou kostku uložíme do krabičky na její správné místo“, jev B , že „bude správnou stěnou vzhůru“ (všechny stěny příslušející stejnému obrázku) a jev C , že „správnou stěnu otočíme do správné polohy“ (přičemž pozici hlavou dolů nepokládáme pro jednoduchost za správnou). Vypočítejte $\mathbb{P}(A \cap B \cap C)$, tj. že při náhodném uložení kostek do krabice složíme správně některý z šesti obrázků.

3.11 Zahradnictví Truhlík a spol. dodává obchodu s květinami zásilku sta kusů květin v květináčích. Prodavačka při přejímání zásilky dělá namátkovou kontrolu kvality zboží. Vybere náhodně pět květin. Jestliže alespoň jedna květina je napadená škůdcem, zásilku nepřijme. Určete pravděpodobnost zamítnutí zásilky, jestliže v celkovém dodaném množství květin jsou 2 % květin napadených škůdcem.

3.12 Z 32 karet, mezi nimiž jsou čtyři esa, taháme jednu kartu po druhé, přičemž vytažené karty nevracíme zpět do balíčku. Vytáhneme celkem 7 karet. Vypočtete pravděpodobnost, že se eso vyskytne pouze jako druhá a sedmá karta.



3.13 U zkoušky z pravděpodobnosti a statistiky je z šedesáti otázek 30 % otázek z teorie pravděpodobnosti. Jednou vytažené otázky se už zpět nevrací. Doposud bylo vyzkoušeno deset studentů. Určete pravděpodobnost, že

- „právě čtyři z vytažených otázek byly z teorie pravděpodobnosti“ (jev A),
- „nejvýše tři byly z teorie pravděpodobnosti“ (jev B),
- „alespoň dvě otázky byly z teorie pravděpodobnosti“ (jev C).



3.14 (Banachova³ úloha) Prodavačka má na pultě rozdělené pulovry na dvě stejně početné hromádky. V každé hromádce je jich n . Při prodeji pulovru vybírá náhodně rovnoměrně jednu z hromádek, ze které pulovr odebere. Časem dojde k tomu, že vybere hromádku, ze které jsou již pulovry vyprodané. Vypočítejte pravděpodobnost, že v druhé hromádce v tomto okamžiku zůstalo k pulovrů.

3.15 (De Montmortův⁴ přiřazovací problém) Máme n různých dopisů a n různých obálek. Dopisy byly do obálek umístěny náhodně, do každé právě jeden. Jaká je pravděpodobnost, že se alespoň jeden dopis dostane do správné obálky? Pravděpodobnost vypočítejte pro $n = 4$.

3.16 Hráči A, B, C střídavě házejí mincí; A hází první, B hází jako druhý a C hází jako třetí. Hra končí, jakmile jednomu hráči padne líc a ten se stává výhercem. Spočtete pravděpodobnosti:

- a) výhry hráče A ,
- b) výhry hráče B .

3.17 Urna obsahuje šest koulí, z nichž čtyři jsou bílé. Náhodným způsobem vybereme bez vracení čtyři koule. Označme jevy: A „právě jedna z prvních dvou tažených koulí je bílá“, B „čtvrtá tažená koule je bílá“, C „ve výběru jsou právě dvě koule bílé“. Určete pravděpodobnost jevů A, B, C . Jsou jevy A, B, C skupinově nezávislé?

3.18 Házíme dvěma hracími kostkami. Jev A znamená, že na modré kostce padlo liché číslo, jev B znamená, že na zelené kostce padlo sudé číslo, jev C znamená, že součet obou čísel je lichý. Jsou náhodné jevy A, B, C nezávislé? Jsou náhodné jevy A, B, C po dvou nezávislé?

3.19 Ve vězení jsou tři lotři: Al Capone, Babin-
ský a Cimrman. Losem jsou určeni dva z nich,
kteří budou popraveni. Al Capone se chce do-
vědět, zda je mezi vylosovanými. Ví, že dozor-
ce by mu na přímou otázku neodpověděl, pro-
to ho žádá, aby mu jmenoval jednoho z jeho
spoluvězňů, který bude popraven. Dozorce je
pravdomluvný a má-li dvě možnosti, volí mezi



³Stefan Banach (1892–1945), polský matematik.

⁴Pierre Remond de Montmort (1678–1719), francouzský matematik.

nimi náhodně rovnoměrně. Jmenuje Babinského. Představuje tato odpověď pro Al Capona nějakou informaci o jeho osudu? (Před rozhovorem s dozorcem byla Al Caponova pravděpodobnost, že bude popraven, rovna $\frac{2}{3}$; je podmíněná pravděpodobnost po dozorcově odpovědi jiná?)

Řešení úloh

3.1 a) $\mathbb{P}(A) = \frac{1}{36} \doteq 0,028$; b) $\mathbb{P}(B) \doteq 0,161$.

3.2

(S) a) $\mathbb{P}(A) = 0,9312$; b) $\mathbb{P}(\bar{A}) = 0,0688$.

(P) a) $\mathbb{P}(A) = 0,9988$; b) $\mathbb{P}(\bar{A}) = 0,0012$.

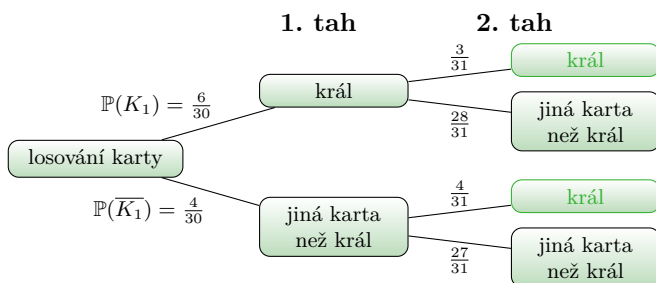
3.3 $\mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) - \mathbb{P}(A_1)\mathbb{P}(A_2) - \mathbb{P}(A_1)\mathbb{P}(A_3) - \mathbb{P}(A_2)\mathbb{P}(A_3) + \mathbb{P}(A_1)\mathbb{P}(A_2)\mathbb{P}(A_3)$.

3.4 a) $\mathbb{P}(A) \doteq 3,01 \cdot 10^{-4}$; b) $\mathbb{P}(B) \doteq 1,50 \cdot 10^{-4}$;

c) $\mathbb{P}(C) = \binom{20}{10} \cdot \left(\frac{4}{6}\right)^{10} \cdot \left(\frac{2}{6}\right)^{10} \doteq 0,054$.

3.5 a) $\mathbb{P}(K_2|K_1) = \frac{3}{31} \doteq 0,097$; b) $\mathbb{P}(\bar{K}_2|K_1) = \frac{28}{31} \doteq 0,903$;

c) $\mathbb{P}(K_2|\bar{K}_1) = \frac{4}{31} \doteq 0,129$; d) $\mathbb{P}(\bar{K}_2|\bar{K}_1) = \frac{27}{31} \doteq 0,871$.



3.6 0,05

3.7 0,154

3.8

a) Jevy A a B jsou nezávislé, neboť jsou nezávislé i jevy $\bar{A}$ a $\bar{B}$:
 $\mathbb{P}(\bar{A}) \cdot \mathbb{P}(\bar{B}) = 0,6 \cdot 0,5 = 0,3 = \mathbb{P}(\bar{A} \cap \bar{B})$.

- b) Vzhledem k nezávislosti jevů A a B je výhodné nejdříve spočítat jejich průnik: $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B) = 0,4 \cdot 0,5 = 0,2$. Poté určíme $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) = 0,4 + 0,5 - 0,2 = 0,7$.
Dále podle téhož vzorce: $\mathbb{P}(\bar{A} \cup B) = 0,6 + 0,5 - 0,6 \cdot 0,5 = 0,8$,
 $\mathbb{P}(A \cup \bar{B}) = 0,4 + 0,5 - 0,4 \cdot 0,5 = 0,7$.
- c) Z nezávislosti jevů A a B plyne $\mathbb{P}(A|\bar{B}) = \mathbb{P}(A) = 0,4$ a $\mathbb{P}(\bar{B}|\bar{A}) = \mathbb{P}(\bar{B}) = 0,5$.

3.9 Jevy A , B , C jsou nezávislé po dvou, nejsou nezávislé po třech.

3.10 $\frac{1}{12!} \cdot \frac{6}{6!} \cdot \frac{1}{4!} \doteq 3,4 \cdot 10^{-25}$

3.11 0,098

3.12 0,00834

3.13 a) $\mathbb{P}(A) \doteq 0,213$; b) $\mathbb{P}(B) \doteq 0,6575$; c) $\mathbb{P}(C) \doteq 0,874$.

3.14 $\binom{2n-k}{n} \cdot \left(\frac{1}{2}\right)^n \cdot \left(\frac{1}{2}\right)^{n-k} = \binom{2n-k}{n} \cdot \left(\frac{1}{2}\right)^{2n-k}$

3.15 $\mathbb{P}_n = 1 - \frac{D_n}{n!}$, kde $D_n = n! \left(1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \dots + \frac{(-1)^n}{n!}\right)$. $\mathbb{P}_{n=4} = 1 - \frac{9}{24} = \frac{5}{8} = 0,625$.

3.16 a) $\frac{4}{7} \doteq 0,571$; b) $\frac{2}{7} \doteq 0,286$.

3.17 $\mathbb{P}(A) = \frac{8}{15}$, $\mathbb{P}(B) = \frac{2}{3}$, $\mathbb{P}(C) = \frac{2}{5}$, A, B, C nejsou skupinově nezávislé.

3.18 A, B, C jsou po dvou nezávislé, nejsou skupinově nezávislé.

3.19 Pravděpodobnost, že bude Al Capone ve vylosované dvojici vězňů k popravě, je $\frac{2}{3}$. Uvede-li dozorce pravdivě Babinského za vězně k popravě, bude podmíněná pravděpodobnost, že druhým odsouzeným k popravě je Al Capone, $\frac{1}{6}$. Ovšem Al Capone nadále zůstává v dalších třech zbylých možných dvojicích dvou vězňů k popravě. Informace od dozorce tedy pouze říká Al Caponovi, že bude popraven ve dvojici s Babinským s pravděpodobností $\frac{1}{6}$. Pravděpodobnost, že bude popraven Al Capone zůstává na $\frac{2}{3}$.

Kapitola 4

Celková pravděpodobnost, Bayesův vzorec

4.1 Celková pravděpodobnost

Podmíněná pravděpodobnost v primárním pohledu slouží, jak bylo zmíněno v odstavci 3.1, k aktualizaci pravděpodobnosti výskytu nějakého jevu, jestliže došlo k nějakému upřesnění informace související s tímto zkoumaným jevem, konkrétně tedy k realizaci podmiňujícího jevu. Koncept podmíněné pravděpodobnosti ovšem můžeme použít i pro výpočet tzv. *celkové (úplné) pravděpodobnosti* zkoumaného jevu v situaci, kdy k realizaci podmiňujícího jevu ještě nedošlo, ale jsou k dispozici jeho možné výsledky a příslušející pravděpodobnosti a související podmíněné pravděpodobnosti. Uvedme si příklad, kde bude tato situace ilustrována.

Příklad 4.1. V tréninkové skupině sportovců je dvacet lyžařů, šest cyklistů a čtyři běžci. Pravděpodobnost splnění tréninkové normy pro lyžaře je 0,9, pro cyklistu 0,8 a pro běžce 0,75. Jaká je pravděpodobnost toho, že náhodně vybraný sportovec splní normu?

Řešení. Označíme A_1 jev „náhodně vybraný sportovec je lyžař“, A_2 jev „náhodně vybraný sportovec je cyklista“ a A_3 jev „náhodně vybraný sportovec je běžec“, dále B jev „náhodně vybraný sportovec splnil normu“. Pak

$$\mathbb{P}(A_1) = \frac{20}{30} = \frac{2}{3}, \quad \mathbb{P}(A_2) = \frac{6}{30} = \frac{1}{5}, \quad \mathbb{P}(A_3) = \frac{4}{30} = \frac{2}{15}.$$

Tyto pravděpodobnosti se nazývají *apriorní* – dopředu známe rozložení výskytu jednotlivých sportů v populaci sportovců. Dále známe *podmíněné*

pravděpodobnosti splnění normy u jednotlivých kategorií sportovců, tj.

$$\mathbb{P}(B|A_1) = 0,9, \quad \mathbb{P}(B|A_2) = 0,8 \quad \text{a} \quad \mathbb{P}(B|A_3) = 0,75.$$

Jev B rozložíme podle kategorie sportu na tři disjunktní podmnožiny (tudiž můžeme jejich pravděpodobnosti sečíst) a opakovaně použijeme vzorec (3.2) pro výpočet pravděpodobnosti jevu B :

$$\begin{aligned} \mathbb{P}(B) &= \mathbb{P}(B \cap A_1) + \mathbb{P}(B \cap A_2) + \mathbb{P}(B \cap A_3) = \\ &= \mathbb{P}(B|A_1) \cdot \mathbb{P}(A_1) + \mathbb{P}(B|A_2) \cdot \mathbb{P}(A_2) + \mathbb{P}(B|A_3) \cdot \mathbb{P}(A_3). \end{aligned} \quad (4.1)$$

Protože máme k dispozici veškerou potřebnou informaci, můžeme dosadit do vzorce (4.1) aktuální hodnoty, čímž dostáváme

$$\mathbb{P}(B) = \frac{2}{3} \cdot 0,9 + \frac{1}{5} \cdot 0,8 + \frac{2}{15} \cdot 0,75 = 0,86.$$

Náhodně vybraný sportovec tedy splní normu s pravděpodobností 0,86. $\triangle$

Postup z řešení příkladu 4.1 zobecníme do věty 4.1 i jejího důkazu.

Věta 4.1 (o celkové pravděpodobnosti). *Nechť $A_1, A_2, \dots$ jsou náhodné jevy tvořící rozklad jevu jistého, tzn.*

$$A_i \cap A_j = \emptyset \text{ pro libovolná } i \neq j \quad \text{a} \quad \bigcup_{i=1}^{\infty} A_i = \Omega.$$

Nechť tyto náhodné jevy mají postupně pravděpodobnosti $\mathbb{P}(A_1), \mathbb{P}(A_2), \dots$, přičemž $\mathbb{P}(A_i) > 0$ pro libovolné $i = 1, 2, \dots$. Uvažujme libovolný náhodný jev B , u něhož známe podmíněné pravděpodobnosti $\mathbb{P}(B|A_i)$ pro libovolné $i = 1, 2, \dots$. Pak platí tzv. formule o celkové pravděpodobnosti

$$\mathbb{P}(B) = \sum_{i=1}^{\infty} \mathbb{P}(B|A_i) \cdot \mathbb{P}(A_i). \quad (4.2)$$

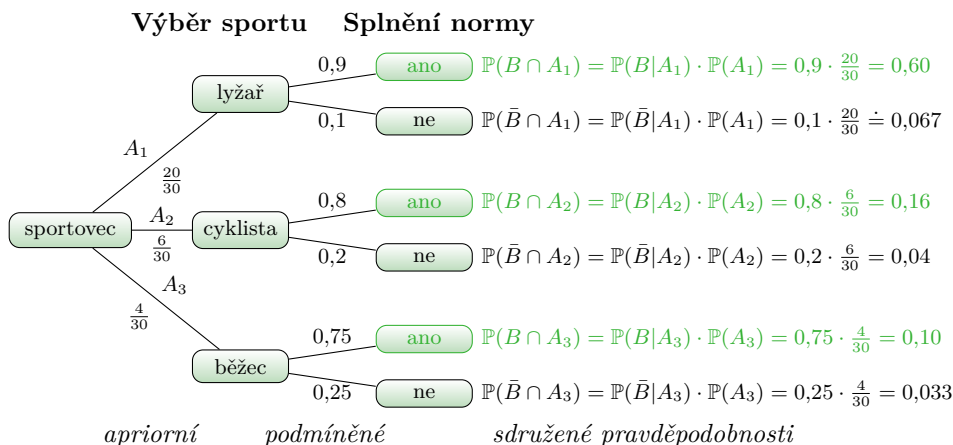
Důkaz. Jevy $A_1, \dots, A_n$ tvoří disjunktní rozklad. Z toho vyplývá, že jestliže pro libovolná $i \neq j$ platí $(A_i \cap B) \cap (A_j \cap B) = \emptyset$, pak

$$\bigcup_{i=1}^{\infty} (A_i \cap B) = B.$$

$$\mathbb{P}(B) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} (A_i \cap B)\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i \cap B) = \sum_{i=1}^{\infty} \mathbb{P}(B|A_i) \cdot \mathbb{P}(A_i).$$

V důkazu jsme využili vlastnosti b) z věty 2.2 na straně 24 a vzorce (3.2). $\square$

Situaci náhodného výběru jednoho sportovce je také možné znázornit stromovým diagramem, jak ukazuje obrázek 4.1. Celková pravděpodobnost je součtem sdružených pravděpodobností, ležících na všech větvích s označením „ano“ (splnil normu).



Obrázek 4.1: Stromový diagram náhodného výběru sportovce z úlohy 4.1

4.2 Bayesův¹ vzorec

Vraťme se k příkladu 4.1 ze strany 57 a položme si novou otázku.

Příklad 4.2. Ve skupině třiceti sportovců je dvacet lyžařů, šest cyklistů a čtyři běžci. Lyžař splní normu s pravděpodobností 0,9, cyklista s pravděpodobností 0,8 a běžec s pravděpodobností 0,75. Nyní je známo, že náhodně vybraný sportovec splnil normu. Jaká je pravděpodobnost, že je to cyklista?

Řešení. Necht' jevy A_1, A_2, A_3, B znamenají totéž, co v řešení příkladu 4.1. V něm jsme na základě znalostí apriorních a podmíněných pravděpodobností $\mathbb{P}(A_i)$ a $\mathbb{P}(B|A_i)$ dopočítali celkovou pravděpodobnost jevu B . Nyní navíc víme, že jev B nastal, a chceme příslušným způsobem „aktualizovat“ pravděpodobnosti jevů A_i (konkrétně tedy A_2).

Z pravděpodobnostního pohledu hledáme pravděpodobnost, s jakou bude vylosován cyklista (jev A_2), jestliže víme, že splnil normu (jev B), tj. chceme dle vzorce (3.1) z definice 3.1 vypočítat hodnotu podmíněné pravděpodobnosti

$$\mathbb{P}(A_2|B) = \frac{\mathbb{P}(A_2 \cap B)}{\mathbb{P}(B)}.$$

¹Thomas Bayes (1701–1761), anglický duchovní a statistik.

Čítatel vzorce je sdružená pravděpodobnost

$$\mathbb{P}(A_2 \cap B) = \mathbb{P}(B|A_2) \cdot \mathbb{P}(A_2) = 0,8 \cdot \frac{6}{30} = 0,16,$$

kterou jsme kupříkladu zanesli do stromového diagramu 4.1, a jmenovatel $\mathbb{P}(B) = 0,86$ je hodnota celkové pravděpodobnosti, kterou jsme také vypočítali v příkladu 4.1. Hledaná výsledná pravděpodobnost je tedy

$$\mathbb{P}(A_2|B) = \frac{\mathbb{P}(A_2 \cap B)}{\mathbb{P}(B)} = \frac{0,16}{0,86} \doteq 0,186. \quad (4.3)$$

Výpočet příkladu 4.2 ukrývá náznak důkazu tzv. Bayesova vzorce, který nyní uvedeme v obecném tvaru.

Věta 4.2 (Bayesova věta). *Nechť jsou splněny předpoklady věty 4.1. Pak pro libovolné $j = 1, 2, \dots$ platí*

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A_j) \cdot \mathbb{P}(A_j)}{\sum_{i=1}^{\infty} \mathbb{P}(B|A_i) \cdot \mathbb{P}(A_i)}. \quad (4.4)$$

Důkaz. Podle vzorce (3.1) pro podmíněnou pravděpodobnost je

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(A_j \cap B)}{\mathbb{P}(B)}. \quad (4.5)$$

Po dosazení (4.2) do (4.5) dostáváme

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(A_j \cap B)}{\sum_{i=1}^{\infty} \mathbb{P}(B|A_i) \cdot \mathbb{P}(A_i)}.$$

V čitateli ještě použijeme (3.2) a dostáváme vzorec (4.4), což jsme měli dokázat. $\square$

Vzorec (4.4) nazýváme první Bayesův vzorec, který využijeme tak, že podle textu úlohy sestavíme jevy $A_i, i \in I$, tzv. hypotézy, které se navzájem vylučují a současně jejich sjednocení reflektuje všechny možnosti – tedy tvoří rozklad Ω . Hledáme pravděpodobnost jednoho z těchto jevů, například A_j pro nějaké $j \in I$, za podmínky, že nastal jev B . Dopředu známe tzv. apriorní pravděpodobnosti $\mathbb{P}(A_i)$ a podmíněné pravděpodobnosti $\mathbb{P}(B|A_i)$



pro všechna $i \in I$. Pravděpodobnost $\mathbb{P}(A_j|B)$ nazýváme aposteriorní pravděpodobnost a vypočítáme ji dosazením známých hodnot do vzorce (4.4).

Vraťme se nyní k příkladu 4.2 a ověřme jeho řešení pomocí prvního Bayesova vzorce, což je ovšem přímo výpočet (4.3).

Řešení. (pokračování ze strany 59) S využitím již známých pravděpodobností z příkladu 4.1 vypočteme $\mathbb{P}(A_2|B)$. Po dosazení do (4.4) dostaneme

$$\mathbb{P}(A_2|B) = \frac{\mathbb{P}(B|A_2) \cdot \mathbb{P}(A_2)}{\sum_{j=1}^3 \mathbb{P}(B|A_j) \cdot \mathbb{P}(A_j)} = \frac{0,8 \cdot \frac{1}{5}}{0,9 \cdot \frac{2}{3} + 0,8 \cdot \frac{1}{5} + 0,75 \cdot \frac{2}{15}} \doteq 0,186.$$

Pravděpodobnost, že náhodně vybraný sportovec splňující normu je cyklista, je rovna po zaokrouhlení 0,186. $\triangle$

Otázka 4.1. Spočítejte pravděpodobnost, že náhodně vybraný sportovec splňující normu je lyžař, resp. běžec.

Příklad 4.3. Při vyšetřování pacienta je podezření na tři navzájem se vylučující nemoci. Pravděpodobnost výskytu první nemoci je 0,3, druhé 0,5 a třetí nemoci 0,2. Laboratorní zkouška dává pozitivní výsledek u 15 % nemocných na první nemoc, u 30 % nemocných na druhou nemoc a u 30 % na třetí nemoc. Jaká je pravděpodobnost výskytu druhé nemoci, jestliže po vykonání laboratorní zkoušky je výsledek pozitivní (pacient některou z nemocí má)?



Řešení. Označme si A_1 jev „pacient má 1. nemoc“, A_2 jev „pacient má 2. nemoc“, A_3 jako jev „pacient má 3. nemoc“ a B jev „laboratorní zkouška dává pozitivní výsledek“. Jevy A_1, A_2, A_3 tvoří rozklad jevu jistého, neboť

$$A_i \cap A_j = \emptyset \text{ pro libovolná } i \neq j, \ i, j = 1, 2, 3, \text{ a } \bigcup_{i=1}^3 A_i = \Omega.$$

Víme, že $\mathbb{P}(A_1) = 0,3$, $\mathbb{P}(A_2) = 0,5$ a $\mathbb{P}(A_3) = 0,2$. Známe i podmíněné pravděpodobnosti: $\mathbb{P}(B|A_1) = 0,15$, $\mathbb{P}(B|A_2) = 0,30$, $\mathbb{P}(B|A_3) = 0,30$. Chceme zjistit hodnotu $\mathbb{P}(A_2|B)$. Opět použijeme Bayesův vzorec. Po dosazení známých pravděpodobností do (4.4) dostáváme

$$\mathbb{P}(A_2|B) = \frac{0,5 \cdot 0,3}{0,3 \cdot 0,15 + 0,5 \cdot 0,3 + 0,2 \cdot 0,3} \doteq 0,588.$$

Jestliže je po vykonání laboratorní zkoušky výsledek pozitivní, pravděpodobnost výskytu druhé nemoci je po zaokrouhlení 0,588. $\triangle$

Příklad 4.4. V celé společnosti je naočkováno 60 % obyvatelstva, 40 % všech lidí nemá očkování. Ze statistik nemocnic je známo, že pravděpodobnost, že pacient je očkovaný, resp. neočkovaný, za předpokladu že skončí na JIP, je 0,1, resp. 0,9. Jaké je riziko, že neočkovaný člověk skončí na JIP?

Řešení. Označme náhodné jevy: O „Člověk je očkovaný“, J „Člověk skončil na JIP.“ Ze zadání je známo: $\mathbb{P}(O) = 0,6$, $\mathbb{P}(\bar{O}) = 0,4$, $\mathbb{P}(O|J) = 0,1$, $\mathbb{P}(\bar{O}|J) = 0,9$.

Platí

$$\mathbb{P}(J|O) = \frac{\mathbb{P}(O|J) \cdot \mathbb{P}(J)}{\mathbb{P}(O)} = \frac{0,1 \cdot \mathbb{P}(J)}{0,6}$$

a podobně

$$\mathbb{P}(J|\bar{O}) = \frac{\mathbb{P}(\bar{O}|J) \cdot \mathbb{P}(J)}{\mathbb{P}(\bar{O})} = \frac{0,9 \cdot \mathbb{P}(J)}{0,4}.$$

$$\frac{\mathbb{P}(J|O)}{\mathbb{P}(J|\bar{O})} = \frac{\frac{0,1}{0,6}}{\frac{0,9}{0,4}} \doteq 0,074.$$

$\triangle$

Z uvedených příkladů je vidět, že v úlohách řešitelných podle Bayesova vzorce postupujeme následovně:

- 1) stanovíme náhodné jevy $A_1, \dots, A_n$, které tvoří rozklad jevu jistého, tzn., že jsou disjunktní a vyčerpávají všechny možnosti,
- 2) stanovíme jev B ,
- 3) identifikujeme apriorní pravděpodobnosti $\mathbb{P}(A_1), \dots, \mathbb{P}(A_n)$ a podmíněné pravděpodobnosti $\mathbb{P}(B|A_1), \dots, \mathbb{P}(B|A_n)$ a
- 4) dosadíme do vzorce (4.4).

Příklad 4.5 (AIDS). Krevní test na přítomnost viru HIV nemusí vždy správně identifikovat chorobu. Mohou nastat dva druhy chyby:

- a) test špatně indikuje pozitivitu (přítomnost viru), nebo
- b) test špatně indikuje negativitu (nepřítomnost viru).

Statistickým pozorováním bylo zjištěno, že tento test je velmi spolehlivý, přesněji, je-li objekt infikován, bude test pozitivní s pravděpodobností 0,995 (tato vlastnost se nazývá senzitivita testu). Neboli $\mathbb{P}(Poz|Inf) = 0,995$, odkud dostáváme, že pravděpodobnost chyby (falešné negativy testu) je

$\mathbb{P}(Neg|Inf) = 0,005$. Podobně je identifikován neinfikovaný objekt s pravděpodobností $\mathbb{P}(Neg|NeInf) = 0,996$ (tato vlastnost se nazývá specificita testu) a pravděpodobnost chyby (falešné positivity testu) je $\mathbb{P}(Poz|NeInf) = 0,004$.

Předpokládejme, že bude vydán zákon, který nařídí všem lidem podstoupit tento „přesný“ test, aby mohli být identifikováni všichni infikovaní lidé. Jestliže pak náhodně vybereme jednoho člověka s pozitivním výsledkem testu, ptáme se, jaká je pravděpodobnost, že skutečně má HIV. Neboli zajímá nás pravděpodobnost $\mathbb{P}(Inf|Poz)$, kterou určíme podle Bayesova vzorce:

$$\mathbb{P}(Inf|Poz) = \frac{\mathbb{P}(Poz|Inf) \cdot \mathbb{P}(Inf)}{\mathbb{P}(Poz|Inf) \cdot \mathbb{P}(Inf) + \mathbb{P}(Poz|NeInf) \cdot \mathbb{P}(NeInf)}.$$

Zbývá nám určit apriorní pravděpodobnosti $\mathbb{P}(Inf)$, $\mathbb{P}(NeInf)$. V České republice bylo k 31. červenci 2021 registrováno 2 990 infikovaných osob. Ke stejnému datu měla Česká republika 10 524 167 obyvatel. Zjištění vede k odhadům

$$\mathbb{P}(Inf) = \frac{2\,990}{10\,524\,167} \doteq 0,000\,284 \quad \text{a} \quad \mathbb{P}(NeInf) \doteq 0,999\,716.$$

V lékařské diagnostice se $\mathbb{P}(Inf)$ nazývá prevalence nemoci². Po dosazení do (4.4) dostáváme

$$\mathbb{P}(Inf|Poz) = \frac{0,995 \cdot 0,000\,284}{0,995 \cdot 0,000\,284 + 0,004 \cdot 0,999\,716} \doteq 0,066.$$

Vydávat takový zákon tedy není příliš chytré, protože pouze 6,6 % pozitivně testovaných lidí by bylo skutečně infikovaných HIV.

²Při analýze šíření chorob se místo prevalence obvykle používá tzv. incidence nemoci, tj. počet nově zjištěných pacientů s chorobou vztažený k celkové velikosti populace.

4.3 Úlohy

4.1 Dvě školky produkují sazenice smrků. První školka vyprodukuje 80 % sazenic, přičemž ze sta je jich devadesát první jakosti. Druhá školka kryje výsadbu z 20 %, přičemž ze sta sazenic je šedesát první jakosti.

- Jaká je pravděpodobnost, že náhodně vybraná sazenice je první jakosti?
- Jaká je pravděpodobnost, že náhodně vybraná sazenice první jakosti je z první školky?

4.2 Ve městě jsou tři obchodní společnosti. Pod první obchodní společností spadá dvacet obchodů, pod druhou patnáct obchodů a pod třetí deset obchodů. Při návštěvě obchodu první společnosti budete ošizen s pravděpodobností 0,15, v obchodě druhé společnosti 0,08 a u třetí společnosti s pravděpodobností 0,05. Jaká je pravděpodobnost, že při nákupu v tomto městě budete ošizen?

4.3 Příjímací zkoušky na určitou střední školu se konají ve čtyřech třídách. V první třídě s padesáti studenty zkoušku udělalo 80 %, ve druhé třídě s 35 studenty prospělo 40 %, ve třetí třídě se 40 studenty zkoušky složilo 65 % studentů a ve čtvrté třídě se 30 studenty byla úspěšnost 60 %. Ze všech studentů, kteří absolvovali přijímací zkoušky, vybereme náhodně jednoho. Jaká je pravděpodobnost, že tento student zkoušky úspěšně složil?



4.4 V autobusu je n cestujících. Na následující zastávce každý z nich vystupuje s pravděpodobností p . Kromě toho do autobusu nenastoupí ani jeden cestující s pravděpodobností p_0 a s pravděpodobností $1 - p_0$ nastoupí jeden nový cestující. Jaká je pravděpodobnost toho, že když se autobus znovu rozjede, bude v něm po následující zastávce n cestujících jako na začátku?

4.5 Ve třídě je 36 žáků, z toho je 15 chlapců a 21 děvčat. Z chlapců je pět obézních a z dívek je obézních sedm. Jaká je pravděpodobnost, že vybrané obézní dítě bude chlapec?

4.6 Na hřišti se sešly dvě skupinky hochů. V první skupince bylo dvanáct hochů a ve druhé deset. Hoši si chtějí zahrát fotbal. Jelikož na každé straně by mělo být jedenáct hráčů, je z první skupiny náhodně vybrán jeden, který je převeden do druhé skupiny. Jaká je pravděpodobnost, že náhodně vybraný hoch z doplněné druhé skupiny je dobrým fotbalistou, jestliže první skupinka před odchodem jednoho z hochů měla osm dobrých a čtyři horší fotbalisty a druhá šest dobrých a čtyři horší fotbalisty?



4.7 Jsou dána čtyři osudí. Nechť pravděpodobnost volby kteréhokoliv z nich je stejná. V každém z prvních tří osudí jsou obsaženy tři lístky se sudými čísly a dva lístky s lichými čísly. Ve čtvrtém osudí jsou čtyři lístky se sudými čísly a jeden lístek s lichým číslem. Náhodně zvolíme osudí, vytáhneme z něj jeden lístek a vložíme ho do jiného osudí (opět náhodně zvoleného). Z tohoto osudí pak vytáhneme jeden lístek. Jaká je pravděpodobnost, že poslední vytážený lístek je se sudým číslem?

4.8 12 studentů oborů M-F, 14 studentů oboru M-Bi a 12 studentů M-VT umí řešit příklad na Bayesovy vzorce pořadě s pravděpodobnostmi 0,80, 0,65 a 0,95. S jakou pravděpodobností je Pavel posluchač oboru

- a) M-F,
- b) M-Bi,
- c) M-VT,

jestliže neumí řešit příklady na Bayesovy vzorce?

4.9 Pro zprávy vysílané Morseovou abecedou platí: místo vyslaného signálu „tečka“ přijde omylem ve $\frac{2}{5}$ případů „čárka“ a místo vyslaného signálu „čárka“ přijde v $\frac{1}{3}$ případů omylem „tečka“. Vyslané signály tečka a čárka jsou v poměru 5 : 3. Zjistěte, jaká je pravděpodobnost toho, že byl vyslaný signál

- a) tečka za podmínky, že jsme přijali signál čárka (jev A);
- b) tečka za podmínky, že jsme přijali signál tečka (jev B)?



4.10 Běhu na 100 m se účastní tři skupiny běžců. První skupina s deseti běžci uběhne trasu v limitu

s pravděpodobností 0,90, druhá skupina šesti běžců s pravděpodobností 0,85 a třetí skupina šesti běžců s pravděpodobností 0,75. Náhodně jsme vybrali ze všech běžců jednoho, o kterém jsme zjistili, že uběhne 100 m v limitu. Jaká je pravděpodobnost, že tento vybraný běžec pochází z druhé skupiny?



Řešení úloh

4.1 a) 0,84; b) 0,857.

4.2 0,1044

4.3 0,6323

4.4 $p_0 \cdot (1 - p)^n + (1 - p_0) \cdot n \cdot p \cdot (1 - p)^{n-1}$

4.5 0,417

4.6 0,606

4.7 0,65

4.8 a) 0,304; b) 0,620; c) 0,076.

4.9 a) $\mathbb{P}(A) = 0,5$; b) $\mathbb{P}(B) = 0,75$

4.10 0,274

Kapitola 5

Náhodná veličina

Práce s náhodnými jevy, tedy s množinou elementárních jevů Ω a σ -algebrou $\mathcal{A}$ může být někdy obtížná, neboť konstrukce obou množin může být značně komplikovaná: elementárními jevy mohou být kupříkladu libovolné objekty či jejich k -tice, kterých může být vysoký počet. Naštěstí mnoho výsledků náhodných pokusů je číselných: buď je výsledek číslo již z podstaty pokusu, nebo jej lze číslem vyjádřit. Pro reprezentaci takových výsledků tedy můžeme používat pravděpodobnostní prostor vybudovaný nad množinou reálných čísel. Matematické objekty reprezentující náhodné pokusy s číselnými výsledky nazýváme *náhodné veličiny*.

5.1 Zavedení náhodné veličiny

Uvažujme pravděpodobnostní prostor $(\Omega, \mathcal{A}, \mathbb{P})$, kde Ω je neprázdná množina všech možných výsledků náhodného pokusu, $\mathcal{A}$ je σ -algebra sestavená na množině Ω a $\mathbb{P}: \mathcal{A} \rightarrow [0; 1]$ je pravděpodobnost. Pokud je výsledek pokusu číselný, mohli bychom množinu Ω definovat rovnou jako množinu všech možných číselných výsledků tohoto pokusu – tedy například u měření tělesné výšky studentů bychom vzali $\Omega = (0; \infty)$, za systém jevů borelovskou σ -algebru $\mathcal{B}([0; \infty))$ a pracovali přímo s tímto prostorem. Náhodné veličiny ovšem můžeme zavést i pro případy, kdy prostor výsledků sice není číselný, ale můžeme jej vhodným způsobem číselně zakódovat, jak učiníme v příkladech 5.1 a 5.2. Z tohoto důvodu definujeme náhodnou veličinu jako funkci, která libovolnému (číselnému či abstraktnímu) náhodnému jevu přiřadí vhodnou číselnou reprezentaci.

Definice 5.1 (Náhodná veličina)

Nechť $(\Omega, \mathcal{A}, \mathbb{P})$ je pravděpodobnostní prostor. Reálnou funkci X definova-

nou na Ω nazýváme náhodnou veličinou, jestliže X je měřitelné zobrazení¹ definované jako $X: (\Omega, \mathcal{A}) \rightarrow (\mathbb{R}, \mathcal{B})$, tj. pro libovolnou borelovskou množinu $B \in \mathcal{B}$ platí

$$\{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{A}, \quad (5.1)$$

kde $\mathcal{B} = \mathcal{B}(\mathbb{R})$ je σ -algebra borelovských² podmnožin $\mathbb{R}$.

Příklad 5.1. Jedenkrát hodíme jednou kostkou. Množina elementárních jevů je $\Omega = \{\square, \blacksquare, \boxdot, \boxtimes, \boxplus, \boxminus\}$, σ -algebra (kolekce náhodných jevů) je potenční množina $\mathcal{A} = \mathcal{P}(\Omega)$. Náhodnou veličinou X rozumíme číselný výsledek hodu kostkou a definujeme ji jako funkci $X: \Omega \rightarrow \mathbb{R}$ s hodnotami

$$X(\square) = 1, X(\blacksquare) = 2, X(\boxdot) = 3, X(\boxtimes) = 4, X(\boxplus) = 5, X(\boxminus) = 6. \quad (5.2)$$

Všimněte si, že takto definovaná funkce je skutečně náhodnou veličinou, tj. je měřitelnou funkcí: pro libovolnou borelovskou podmnožinu reálných čísel platí, že její vzor patří do $\mathcal{A}$. Například pro interval $(\pi, 5]$ je vzorem zobrazení množina $X^{(-1)}((\pi; 5]) = \{\boxtimes, \boxplus\}$ a ta je prvkem $\mathcal{A}$.

O tom, zda je funkce X náhodnou veličinou, rozhoduje mimo jiné i podobu zvolené σ -algebry. Pokud bychom v příkladu 5.1 pracovali se σ -algebrou například ve tvaru

$$\mathcal{A} = \{\emptyset; \{\square, \blacksquare\}; \{\boxdot, \boxtimes, \boxplus, \boxminus\}; \Omega\}, \quad (5.3)$$

pak funkce definovaná prostřednictvím (5.2) nemůže být náhodnou veličinou: v příkladu zmíněná množina $X^{(-1)}((\pi; 5]) = \{\boxtimes, \boxplus\}$ do této σ -algebry totiž nepatří, což je v rozporu s podmínkou měřitelnosti. Připomeňme v této souvislosti, že z praktického hlediska představuje σ -algebra příslušející náhodnému pokusu informaci, kterou můžeme identifikovat jako výsledek náhodného pokusu. Množina výsledků ve tvaru (5.3) odpovídá situaci, ve které například nemůže nastat jev „padla $\square$ “, resp. takový jev není možné identifikovat (hod byl například proveden tajně a statistikovi je pouze sdělena informace, zda výsledek převýšil dvojku). Konstrukce a využití takového specifických σ -algeber ovšem přesahuje rámec této knihy; pro běžné užití si vystačíme se σ -algebry zmíněnými v příkladech 2.1 a 2.2.

V následujícím příkladu si zkonstruujeme náhodnou veličinu, jejímž oborem hodnot bude něco jiného než přirozená čísla, jak jsme byli dosud zvyklí.

¹Dvojice $(\kappa, \mathcal{A})$ a $(\theta, \mathcal{B})$, kde κ a θ jsou neprázdné množiny a $\mathcal{A}, \mathcal{B}$ jsou σ -algebry jejich podmnožin, nazýváme měřitelné prostory. Zobrazení $f: \kappa \rightarrow \theta$ se nazývá měřitelné právě tehdy, když $\forall B \in \mathcal{B}$ platí $\{x \in \kappa : f(x) \in B\} \in \mathcal{A}$, tj. vzory měřitelných množin jsou měřitelné. Měřitelné zobrazení značíme $f: (\kappa, \mathcal{A}) \rightarrow (\theta, \mathcal{B})$.

²Tj. nejmenší σ -algebra obsahující všechny otevřené podmnožiny $\mathbb{R}$, viz příklad 2.2 na straně 22.

Příklad 5.2. Pohlaví narozeného dítěte můžeme prezentovat jako náhodný pokus se dvěma možnými výsledky, a to chlapec a děvče. Tento pokus tedy reprezentujeme množinou elementárních jevů $\Omega = \{\text{chlapec}, \text{děvče}\}$ a jevový prostor je potenční množina

$$\mathcal{A} = \{\emptyset, \{\text{chlapec}\}, \{\text{děvče}\}, \Omega\}.$$

Náhodnou veličinu X můžeme v tomto případě definovat například takto:

$$X(\omega) = \begin{cases} 0, & \text{jestliže } \omega = \text{chlapec}, \\ 1, & \text{jestliže } \omega = \text{děvče}. \end{cases}$$

V tomto případě mají výsledky náhodného pokusu kvalitativní charakter (chlapec, děvče) a my jsme je ohodnotili určitým, v tomto případě uměle zvoleným reálným číslem.

Náhodné veličiny se obvykle značí velkými písmeny z konce abecedy, například X, Y, Z . Hodnoty $X(\omega)$, kterých mohou náhodné veličiny nabývat, budeme značit malými písmeny x, y, z . Říkáme také, že x je tzv. *realizace* náhodné veličiny X (zejména v případech, kdy je již hodnota náhodné veličiny známa, protože náhodný pokus již byl proveden). Také běžně používáme zjednodušeného zápisu: místo $\{\omega \in \Omega : X(\omega) \in B\}$ budeme stručně psát $\{X \in B\}$ a místo $\{\omega \in \Omega : X(\omega) < x\}$ budeme psát $\{X < x\}$.

Součty, součiny a podíly náhodných veličin jsou opět náhodné veličiny. Stejně tak umocnění náhodné veličiny přirozeným číslem a násobení náhodné veličiny skalárem jsou opět náhodné veličiny. Důkazy jsou uvedeny například v [7].

5.2 Rozdělení náhodné veličiny

Stejně jako jsme v kapitole 2 na abstraktním jevovém prostoru $(\Omega, \mathcal{A})$ zavedli pravděpodobnost $\mathbb{P}$ (splňující požadavky definice 2.2), mohli bychom podobným způsobem zavést pravděpodobnostní míru na borelovských podmnožinách prostoru $(\mathbb{R}, \mathcal{B})$, reprezentující možné číselné výsledky náhodných veličin. Protože jsme však definovali náhodnou veličinu X nad již stanoveným pravděpodobnostním prostorem $(\Omega, \mathcal{A}, \mathbb{P})$, je pravděpodobnost výsledků určena již definovanou mírou $\mathbb{P}$ a předpisem funkce X .

Příklad 5.3 (pokračování příkladu 5.2). Pravděpodobnost narození chlapce je 0,514. Definujme náhodnou veličinu X tak, jak byla zavedena v příkladu 5.2. Chceme-li nyní změřit pravděpodobnost, že tato náhodná veličina

nabude hodnoty nula, budeme postupovat takto:

$$\begin{aligned}\mathbb{P}(X = 0) &= \mathbb{P}\{\omega \in \Omega : X(\omega) = 0\} = \mathbb{P}\{\omega \in X^{(-1)}(0)\} = \\ &= \mathbb{P}\{\omega \in \{\text{chlapec}\}\} = \mathbb{P}\{\omega = \text{chlapec}\} = 0,514.\end{aligned}$$

Podobně bychom určili, že $\mathbb{P}(X = 1) = \mathbb{P}\{\omega = \text{dívka}\} = 0,486$.

Vidíme, že tímto postupem můžeme určit pravděpodobnost libovolné borelovské množiny, jednoduše tak, že zjistíme pravděpodobnost jejího vzoru³. Takto zavedená pravděpodobnostní míra na borelovských množinách (indukovaná pravděpodobností $\mathbb{P}$) se nazývá *rozdělení pravděpodobnosti náhodné veličiny*, či krátce jen *rozdělení*⁴. Rozdělení pravděpodobnosti tedy plně charakterizuje pravděpodobnostní chování náhodné veličiny.

V praktických úlohách určujeme pravděpodobnost číselných výsledků tak, jak bychom to dělali pro jakýkoliv jiný pravděpodobnostní prostor. Například *diskrétní náhodné veličiny*, tedy veličiny s konečným nebo nejvýše spočetným množstvím hodnot, jsou úplně charakterizovány tzv. *pravděpodobnostní funkcí*, tj. výčtem hodnot náhodné veličiny a příslušejících pravděpodobností. Takovými veličinami jsou například výsledek hodu kostkou (viz příklad 5.1), číselně reprezentované pohlaví narozeného dítěte (příklady 5.2, 5.3), počet příchozích telefonních výzev (příklad 7.16) apod. Situaci budeme ilustrovat také na následujícím příkladu.

Příklad 5.4 (zkouška ze statistiky). V univerzitním studijním systému je možné vypořádat dlouhodobé rozložení známek z předmětu Statistika. Student může u zkoušky dostat jednu ze známek 1, 2, 3, nebo 4. Pravděpodobnost, s jakou náhodně vybraný student obdržel konkrétní známku, lze přehledně reprezentovat tabulkou hodnot a příslušejících pravděpodobností (viz tabulku 5.1). Definujeme náhodnou veličinu X = známka ze statistiky pro náhodně vybraného studenta. Potom tabulka 5.1 je rovněž reprezentací pravděpodobnostní funkce náhodné veličiny X , a tedy jejího rozdělení.

Otázka 5.1. K tabulce 5.1 z příkladu 5.4 sestavte v Excelu sloupcový graf reprezentující pravděpodobnostní funkci náhodné veličiny X .

Pro výpočet pravděpodobností užíváme pravidel zavedených v kapitole 2; v příkladu 5.4 například máme

$$\mathbb{P}(X = 2) = 0,25 \quad \text{nebo} \quad \mathbb{P}(X \geq 3) = 0,50 + 0,15 = 0,65.$$

³Náhodnou veličinu jsme právě proto definovali tak, abychom to mohli udělat: jako *měřitelnou* funkci.

⁴Rozdělení pravděpodobnosti můžeme značit například $\mu_X = \mathbb{P} \circ X^{(-1)}$, protože je z hlediska formální matematiky složenou funkcí míry $\mathbb{P}$ a zobrazení $X^{(-1)}$. Toto značení však nebudeme dále potřebovat.

Tabulka 5.1: Znamky a pravděpodobnosti pro náhodně vybraného studenta

Znamka	1	2	3	4
Pravděpodobnost	0,10	0,25	0,50	0,15

Všimněte si, že součet pravděpodobností v tabulce 5.1 činí 1. Rigorózní definici diskrétní náhodné veličiny uvedeme v odstavci 5.4. Je-li počet možných výsledků náhodné veličiny nespočetný, pravděpodobnostní funkce je zavedena pomocí integrálního počtu a představíme ji v odstavci 5.5. K tomu účelu se nám bude hodit ještě jedna alternativní úplná charakterizace rozdělení náhodné veličiny, kterou si představíme nyní.

5.3 Distribuční funkce

Definice 5.2 (Distribuční funkce)

Nechť X je náhodná veličina. Její distribuční funkci nazýváme reálnou funkcí F_X reálné proměnné x definovanou předpisem

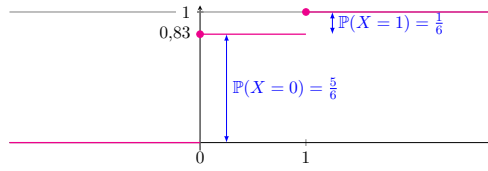
$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(\{\omega : X(\omega) \leq x\}). \quad (5.4)$$

Distribuční funkce je pravděpodobnost pro množiny specifického tvaru $\{\omega : X(\omega) \leq x\}$, které patří do σ -algebry $\mathcal{A}$, jsou to tedy náhodné jevy a je možné pro ně pravděpodobnost (hodnotu distribuční funkce) určit. Distribuční funkce je z tohoto titulu definována pro všechna $x \in \mathbb{R}$ a lze ukázat, že vedle pravděpodobnostní funkce také jednoznačně určuje rozdělení náhodné veličiny.

Příklad 5.5 (Člověče, nezlob se!). Ve hře Člověče, nezlob se! hráč nasadí figurku do hry v okamžiku, kdy mu při hodu kostkou padne šestka. Hod kostkou jako náhodný pokus reprezentujeme pomocí množiny elementárních jevů

$$\Omega = \{\square, \begin{smallmatrix} \square & \square \\ \square & \square \end{smallmatrix}, \begin{smallmatrix} \square & \square & \square \\ \square & \square & \square \end{smallmatrix}, \begin{smallmatrix} \square & \square & \square & \square \\ \square & \square & \square & \square \end{smallmatrix}, \begin{smallmatrix} \square & \square & \square & \square & \square \\ \square & \square & \square & \square & \square \end{smallmatrix}, \begin{smallmatrix} \square & \square & \square & \square & \square & \square \\ \square & \square & \square & \square & \square & \square \end{smallmatrix}\}$$

a za σ -algebru $\mathcal{A}$ vezmeme systém všech podmnožin Ω . Protože pravděpodobnosti všech elementárních jevů jsou totožné ($\frac{1}{6}$), jedná se o klasický pravděpodobnostní prostor a pravděpodobnost $\mathbb{P}(A)$ náhodného jevu $A \in \mathcal{A}$ je rovna poměru počtu příznivých elementárních jevů k počtu všech elementárních jevů. Nyní na tomto prostoru $(\Omega, \mathcal{A}, \mathbb{P})$ zkonstruujeme náhodnou veličinu X , reprezentující úspěch při hodu kostkou, který umožní nasadit figurku do hry. Uděláme to analogicky k příkladu 5.2 pomocí uměle zvolené



Obrázek 5.1: Distribuční funkce náhodné veličiny X z příkladu 5.5.

dvojice čísel takto:

$$X(\omega) = \begin{cases} 0, & \text{jestliže } \omega \in \{\square, \square\cdot, \square\cdot\cdot, \square\cdot\cdot\cdot, \square\cdot\cdot\cdot\cdot\}, \\ 1, & \text{jestliže } \omega = \square\cdot\cdot\cdot\cdot\cdot. \end{cases}$$

Pravděpodobnostní funkce náhodné veličiny X je dána hodnotami pravděpodobnosti $\mathbb{P}(X = 0) = \frac{5}{6}$ a $\mathbb{P}(X = 1) = \frac{1}{6}$. Distribuční funkce F je podle definice 5.2 pak definována takto:

$$F(x) = \begin{cases} 0, & \text{pokud } x < 0, \\ \frac{5}{6}, & \text{pokud } 0 \leq x < 1, \\ 1, & \text{pokud } x \geq 1. \end{cases}$$

Graf této distribuční funkce uvádíme na obrázku 5.1.

Všimněte si, že distribuční funkce z příkladu 5.5 je neklesající, neboť postupně kumuluje pravděpodobnosti (nejprve $\mathbb{P}(X = 0)$, potom $\mathbb{P}(X = 1)$). Proto se distribuční funkce také označují jako *kumulativní distribuční funkce*. Hodnoty distribuční funkce leží mezi nulou a jedničkou, neboť se jedná o pravděpodobnosti. A konečně, distribuční funkce diskrétní náhodné veličiny vykazuje schodovitý (skokovitý) charakter, přičemž skoky nastávají přesně v hodnotách, které náhodná veličina realizuje (v našem případě tedy v hodnotách 0 a 1), a mají velikost příslušejících pravděpodobností. Tyto vlastnosti jsou charakteristické pro libovolnou distribuční funkci a shrneme je jako větu 5.1, kam připojíme ještě informaci o spojitosti a nevlastních limitách distribuční funkce.

Věta 5.1 (o vlastnostech distribuční funkce). *Distribuční funkce $F_X(x)$ náhodné veličiny X je*

- neklesající, tj. pro libovolné $a, b \in \mathbb{R}$, $a \leq b$, platí $F_X(a) \leq F_X(b)$,*
- zprava spojitá v libovolném bodě $x \in \mathbb{R}$, tj. pro libovolné $x_0 \in \mathbb{R}$ platí*

$$\lim_{x \rightarrow x_0^+} F_X(x) = F_X(x_0),$$
- $\lim_{x \rightarrow -\infty} F_X(x) = 0$ a $\lim_{x \rightarrow \infty} F_X(x) = 1$,*

d) má nejvýše spočetně bodů nespojitosti; tyto body nespojitosti jsou prvního druhu (tj. skoky) a velikost skoku v bodě x_0 je

$$F_X(x_0) - F_X(x_0^-) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x_0\}),$$

$$\text{kde } F_X(x_0^-) = \lim_{x \rightarrow x_0^-} F_X(x).$$

Důkaz.

a) Podle definice 5.2 je

$$F_X(a) = \mathbb{P}(\{\omega : X(\omega) \leq a\}),$$

$$F_X(b) = \mathbb{P}(\{\omega : X(\omega) \leq b\}).$$

Jelikož $\{\omega : X(\omega) \leq a\} \subset \{\omega : X(\omega) \leq b\}$, je podle vlastnosti c) pravděpodobnosti ve větě 5.1 na straně 24

$$\mathbb{P}(\{\omega : X(\omega) \leq a\}) \leq \mathbb{P}(\{\omega : X(\omega) \leq b\}),$$

tj. $F_X(a) \leq F_X(b)$.

b) Zvolme si libovolný bod $x_0 \in \mathbb{R}$. Máme dokázat, že funkce F_X je v bodě x_0 zprava spojitá. Uvažujme libovolnou posloupnost $\{x_n\}$ reálných čísel takovou, že $x_n \searrow x_0$. Jestliže si označíme

$$A_n = \{\omega \in \Omega : X(\omega) \leq x_n\}, \quad n = 0, 1, 2, \dots,$$

pak

$$F_X(x_n) = \mathbb{P}(A_n), \quad n = 0, 1, 2, \dots,$$

$$A_n \subset A_{n+1}, \quad n = 1, 2, \dots$$

a

$$A_0 = \bigcup_{n=1}^{\infty} A_n.$$

Chceme dokázat, že $\lim_{n \rightarrow \infty} F_X(x_n) = F_X(x_0)$. To dokážeme s využitím vlastnosti h) pravděpodobnosti ve větě 2.2 na straně 25:

$$\begin{aligned} \lim_{n \rightarrow \infty} F_X(x_n) &= \lim_{n \rightarrow \infty} \mathbb{P}(\{\omega : X(\omega) \leq x_n\}) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \\ &= \mathbb{P}(A_0) = F_X(x_0). \end{aligned}$$

$$c) \lim_{x \rightarrow -\infty} F_X(x) = \lim_{n \rightarrow \infty} F_X(-n) = \lim_{n \rightarrow \infty} \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq -n\}).$$

Podle vlastnosti i) pravděpodobnosti na straně 25 platí

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq -n\}) &= \\ &= \mathbb{P}\left(\bigcap_{n=1}^{\infty} \{\omega \in \Omega : X(\omega) \leq -n\}\right) = \mathbb{P}(\emptyset) = 0, \\ \lim_{x \rightarrow \infty} F_X(x) &= \lim_{n \rightarrow \infty} F_X(n) = \lim_{n \rightarrow \infty} \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq n\}). \end{aligned}$$

Z vlastnosti h) pravděpodobnosti ve větě 2.2 plyne

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq n\}) &= \\ &= \mathbb{P}\left(\bigcup_{n=1}^{\infty} \{\omega \in \Omega : X(\omega) \leq n\}\right) = \mathbb{P}(\Omega) = 1. \end{aligned}$$

d) Označme si:

C jako množinu bodů nespojitosti dané distribuční funkce,

C_n jako množinu bodů nespojitosti se skokem větším než $\frac{1}{n}$, $n = 2, 3, \dots$

Pak $C = \bigcup_{n=2}^{\infty} C_n$. Protože distribuční funkce $F_X(x)$ je pro libovolné $x \in \mathbb{R}$ mezi nulou a jedničkou a je to funkce neklesající, obsahuje C_n nejvýše $(n-1)$ bodů nespojitosti. Množina C je tedy sjednocením spočetně mnoha konečných množin, přičemž takové sjednocení je nejvýše spočetná množina.

Nyní si ukážeme, že velikost skoku v bodě x_0 je

$$F_X(x_0) - F_X(x_0^-) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x_0\}).$$

$$\begin{aligned} \mathbb{P}(\{\omega \in \Omega : X(\omega) = x_0\}) &= \\ &= \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x_0\}) - \mathbb{P}(\{\omega \in \Omega : X(\omega) < x_0\}) = \\ &= F_X(x_0) - \mathbb{P}(\{\omega \in \Omega : X(\omega) < x_0\}) = \\ &= F_X(x_0) - \mathbb{P}\left(\bigcup_{n=1}^{\infty} \left\{\omega \in \Omega : X(\omega) \leq x_0 - \frac{1}{n}\right\}\right). \end{aligned}$$

Z vlastnosti i) pravděpodobnosti ve větě 2.2 na straně 25 plyne

$$\begin{aligned} \mathbb{P}\left(\bigcup_{n=1}^{\infty} \left\{\omega \in \Omega : X(\omega) \leq x_0 - \frac{1}{n}\right\}\right) &= \\ &= \lim_{n \rightarrow \infty} \mathbb{P}\left(\left\{\omega \in \Omega : X(\omega) \leq x_0 - \frac{1}{n}\right\}\right) = F_X(x_0^-). \end{aligned}$$



Nyní se vrátíme ke slíbenému předpisu diskrétních a spojitých náhodných veličin.

5.4 Diskrétní náhodné veličiny

Definice 5.3 (Diskrétní náhodná veličina)

Náhodná veličina X se nazývá diskrétní, nebo též náhodná veličina s diskrétním rozdělením pravděpodobnosti, jestliže existuje posloupnost reálných čísel $\{x_n\}$ a odpovídající posloupnost nezáporných čísel $\{p_n\}$ taková, že

$$p_n = \mathbb{P}(X = x_n)$$

a platí

$$\sum_{n=1}^{\infty} p_n = 1.$$

Pravděpodobnost, s jakou diskrétní náhodná veličina X nabude hodnoty x , budeme označovat $p_X(x)$ nebo jen $p(x)$; $p_X(x) = \mathbb{P}(X = x)$ je tedy pravděpodobnostní funkcí náhodné veličiny X , jak jsme ji zavedli v odstavci 5.2. Distribuční funkce diskrétní náhodné veličiny je po částech konstantní až na hodnoty x_n , ve kterých má skok velikosti $p_n = \mathbb{P}(X = x_n)$. Formálně má tvar

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{\{n: x_n \leq x\}} \mathbb{P}(X = x_n) = \sum_{\{n: x_n \leq x\}} p_n, \quad (5.5)$$

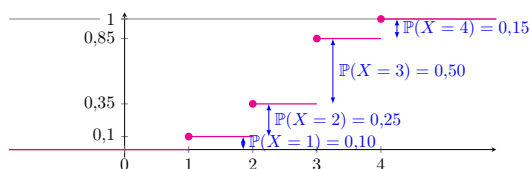
ve které čtenář může identifikovat charakter kumulujících se pravděpodobností. Distribuční funkce slouží mimo jiné k výpočtu pravděpodobnosti jevu $X \in (a; b]$, tj.

$$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a) = \sum_{\{n: a < x_n \leq b\}} \mathbb{P}(X = x_n) = \sum_{\{n: a < x_n \leq b\}} p_n$$

pro libovolná reálná čísla a, b , kde $a \leq b$.

Příklad 5.6 (zkouška ze statistiky). K náhodné veličině X z příkladu 5.4, jejíž pravděpodobnostní funkce je dána tabulkou 5.1, sestrojte distribuční funkci a vypočítejte pravděpodobnosti, že náhodně vybraný student

a) dostane známku 4, tj. zkouškou neprojde;



Obrázek 5.2: Graf distribuční funkce k příkladu 5.6

- b) projde zkouškou se známkou 1 až 3;
- c) obdrží známku 3 nebo 4;
- d) patří k „průměru ročníku“ se známkami 2 nebo 3.

Řešení. Distribuční funkce F_X náhodné veličiny X je po částech konstantní se skoky v hodnotách 1, 2, 3 a 4 velikosti příslušejících pravděpodobností. Graf distribuční funkce ukazuje obrázek 5.2.

Pro výpočty pravděpodobností obecných diskrétních náhodných veličin lze v MS Excel využít funkci `=PROB`. Argumenty funkce jsou oblast hodnot náhodné veličiny (v našem případě udělované známky, které mějme například v buňkách A2:A5), příslušejících pravděpodobností (buňky B2:B5) a meze intervalu, pro které hodnoty počítáme. Snadno tedy spočteme, že

- a) $\mathbb{P}(X = 4) = \text{PROB}(A2:A5; B2:B5; 4) = 0,15;$
- b) $\mathbb{P}(X \leq 3) = \text{PROB}(A2:A5; B2:B5; ; 3) = 0,85;$
- c) $\mathbb{P}(X \geq 3) = \text{PROB}(A2:A5; B2:B5; 3; 4) = 0,65;$
- d) $\mathbb{P}(2 \leq X \leq 4) = \text{PROB}(A2:A5; B2:B5; 2; 4) = 0,90.$

V tomto jednoduchém příkladu samozřejmě snadno výsledky ověříme ručním výpočtem. Všimněte si také, že

- a) $\mathbb{P}(X = 4) = p_X(4),$
- b) $\mathbb{P}(X \leq 3) = F_X(3),$
- c) $\mathbb{P}(X \geq 3) = 1 - \mathbb{P}(X < 3) = 1 - \mathbb{P}(X \leq 2) = 1 - F_X(2),$
- d) $\mathbb{P}(2 \leq X \leq 4) = \mathbb{P}(1 < X \leq 4) = F_X(4) - F_X(1).$

△

Příklad 5.7.  Uvažujme diskrétní náhodnou veličinu X , která udává počet padnutých líců v pěti hodech mincí. Stanovte její pravděpodobnostní a distribuční funkci a vytvořte jejich grafy.



Řešení. Opakovaný hod mincí je příkladem Bernoulliho procesu (viz odstavec 3.3); pravděpodobnost, že

Tabulka 5.2: Hodnoty $p(x)$ a $F(x)$ pro příklad 5.7

x	0	1	2	3	4	5
$p(x)$	0,031 25	0,156 25	0,312 5	0,312 5	0,156 25	0,031 25
$F(x)$	0,031 25	0,187 5	0,5	0,812 5	0,968 75	1

v n hodech padne k líců, je tedy určena vzorcem (3.10), ve kterém je $p = q = \frac{1}{2}$. Pravděpodobnostní funkce má tedy tvar

$$p(x) = \mathbb{P}(X = x) = \begin{cases} \binom{5}{x} \cdot \left(\frac{1}{2}\right)^x \cdot \left(\frac{1}{2}\right)^{5-x} & \text{pro } x \in \{0, 1, \dots, 5\}, \\ 0 & \text{jinak.} \end{cases}$$

Uvedené hodnoty si snadno vypočteme v Excelu užitím kombinatorických vzorců, například $\mathbb{P}(X = 2) = \text{KOMBINACE}(5; 2) * 0,5^2 * 0,5^{(5-2)}$, přičemž dvojku v tomto vzorci fakticky nahradíme odkazem na příslušnou buňku připravené tabulky hodnot (prvního řádku tabulky 5.2), abychom mohli vzorec zkopírovat i pro ostatní⁵ hodnoty x . Vypočtené hodnoty pravděpodobnostní funkce jsou také v tabulce 5.2. Ještě ověříme, že se jedná o rozdělení pravděpodobnosti:

$$\sum_{k=0}^5 p_k = \sum_{k=0}^5 \binom{5}{k} \cdot \left(\frac{1}{2}\right)^k \cdot \left(\frac{1}{2}\right)^{5-k} = 0,031\,25 + 0,156\,25 + \dots + 0,031\,25 = 1.$$

Distribuční funkce je pak kumulativním součtem výše uvedených pravděpodobností:

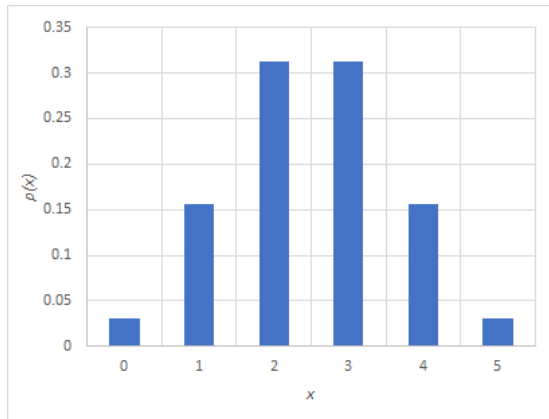
$$F(x) = \mathbb{P}(X \leq x) = \begin{cases} 0 & \text{pro } x < 0, \\ \sum_{k=0}^{\lfloor x \rfloor} \binom{5}{k} \cdot \left(\frac{1}{2}\right)^k \cdot \left(\frac{1}{2}\right)^{5-k} & \text{pro } 0 \leq x \leq 5, \\ 1 & \text{pro } x > 5. \end{cases}$$

Do tabulky 5.2 jsme doplnili i hodnoty distribuční funkce pro $x = 0, 1, \dots, 5$.

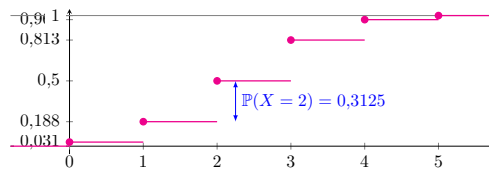
Pravděpodobnostní funkci obvykle znázorňujeme pomocí sloupcového grafu, který je na obrázku 5.3. Graf lze vytvořit například v MS Excel volbou sloupcového grafu na kartě Vložení a úpravou popisků vodorovné osy (v Návrhu grafu zvolíme Vybrat data a jako Popisky vodorovné osy (kategorie) zvolíme namísto výchozích hodnot oblast buněk s hodnotami x). Graf distribuční funkce ukazuje obrázek 5.4⁶.

⁵V odstavci 7.4 si představíme ještě o něco jednodušší vzorec na výpočet hodnot této pravděpodobnostní funkce.

⁶Grafy distribučních funkcí náhodných veličin lze v MS Excel vytvářet pouze přibližně a poměrně komplikovaně, proto je čtenáři předkládáme připravené v jiném grafickém softwaru (konkrétně PGF/TikZ).



Obrázek 5.3: Graf pravděpodobnostní funkce v příkladu 5.7



Obrázek 5.4: Graf distribuční funkce k příkladu 5.7

△

5.5 Absolutně spojitě náhodné veličiny

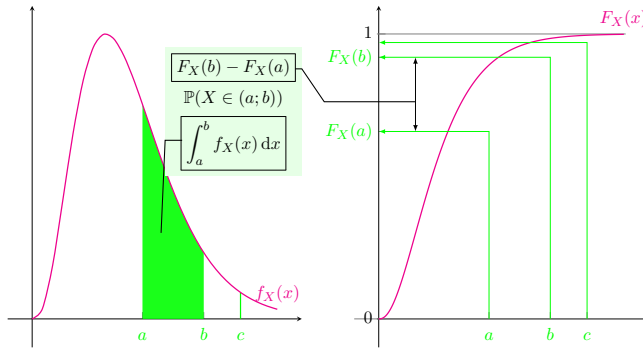
Je-li počet možných výsledků náhodné veličiny X nespočetný, není možné pro výpočet pravděpodobností použít „klasických“ součtových vzorců, které pro nespočetně mnoho hodnot divergují do nekonečna. Místo součtů použijeme pro výpočet určitý (Lebesgueův) integrál a roli pravděpodobnostní funkce přebírá tzv. *hustota rozdělení pravděpodobnosti*.

Definice 5.4 (Absolutně spojitá náhodná veličina)

Náhodná veličina X , resp. její distribuční funkce F_X se nazývá absolutně spojitá, jestliže existuje nezáporná integrovatelná funkce f_X taková, že pro libovolné $x \in (-\infty, \infty)$ platí

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f_X(t) dt. \quad (5.6)$$

Funkce f_X se nazývá hustota (rozdělení) pravděpodobnosti.



Obrázek 5.5: Příklad hustoty a distribuční funkce absolutně spojitě náhodné veličiny a výpočtu pravděpodobnosti

Připomeňme si krátce vlastnosti Lebesgueova integrálu. Definice 5.4 zajišťuje, že k hustotě f_X existuje primitivní funkce. Vzhledem k (5.6) je jednou z nich distribuční funkce (konkrétně ta, která má v $-\infty$ limitně nulovou hodnotu, jak předepisuje vlastnost c) ve větě 5.1). Dále pravděpodobnost, že $X \in (a; b]$ pro nějaká reálná čísla a, b , je vzhledem k vlastnosti d) ve větě 2.2

$$\mathbb{P}(X \in (a; b]) = \mathbb{P}(X \in (-\infty; b]) - \mathbb{P}(X \in (-\infty; a]) = F_X(b) - F_X(a),$$

což není nic jiného než pravidlo pro výpočet určitého (Newtonova) integrálu z hustoty f_X o mezích a, b ,

$$\mathbb{P}(X \in (a; b]) = \int_a^b f_X(x) dx,$$

který dále můžeme vizualizovat (v Riemannově⁷ smyslu) jako plochu pod grafem hustoty f_X mezi jejími body a, b (viz také obrázek 5.5). Z této geometrické interpretace mimo jiné vyplývá, že

$$\mathbb{P}(X = c) = \mathbb{P}(X \in [c; c)) = \int_c^c f_X(x) dx = 0$$

pro libovolnou elementární hodnotu $c \in \mathbb{R}$ („plocha“ pod jedním bodem c je samozřejmě nulová).

Nulovost pravděpodobnosti pro každou jednotlivou elementární hodnotu také znamená, že distribuční funkce v daném bodě nemá skok, je tam tedy spojitá, a je tedy spojitá na celé reálné ose $\mathbb{R}$. Vlastnost (5.6) je ale silnější, zajišťuje dokonce diferencovatelnost ve všech bodech až na nejvýše spočetně

⁷Georg Friedrich Bernhard Riemann (1826–1866), německý matematik.

početnou množinu nulové pravděpodobnosti, ve kterých nemusí být první derivace definována. Takové funkce označujeme jako *absolutně spojité*, odkud plyne i název pro odpovídající pravděpodobnostní rozdělení.

Důsledkem toho je rovněž, že při zkoumání pravděpodobností absolutně spojité náhodné veličiny X nemusíme brát v úvahu všechny její hodnoty v daném intervalu – příslušný integrál zůstane nezměněn, když jich konečně nebo nejvýše spočetně mnoho z intervalu vypustíme. Například pravděpodobnosti

$$\mathbb{P}(X \in (a; b]) = \mathbb{P}(X \in (a; b)) = \mathbb{P}(X \in [a; b)) = \mathbb{P}(X \in [a; b])$$

jsou všechny z tohoto důvodu totožné.

V teorii pravděpodobnosti mají speciální význam množiny hodnot V (říkejme krátce „vlastnost V “), kterých náhodná veličina nabývá s pravděpodobností 1, tedy

$$\mathbb{P}(X \text{ má vlastnost } V) = 1.$$

Z výše uvedeného vyplývá, že z hodnot příslušejících vlastnosti V je možné jich až spočetně mnoho vypustit (nebo změnit), aniž by se cokoliv změnilo na jednotkové pravděpodobnosti⁸. V těchto případech budeme říkat, že X má vlastnost V „skoro jistě“ (příp. „skoro všude“, jinými slovy až na množinu pravděpodobnosti 0) a užívat zkratku s. j.

Výše odvozené vlastnosti hustoty pravděpodobnosti nyní shrneme jako větu 5.2.

Věta 5.2 (o vlastnostech hustoty). *Nechť $f_X(x)$ je hustota rozdělení pravděpodobnosti náhodné veličiny X . Pak platí:*

$$a) \quad f_X(x) = \frac{d}{dx} F_X(x) \text{ s. j.,}$$

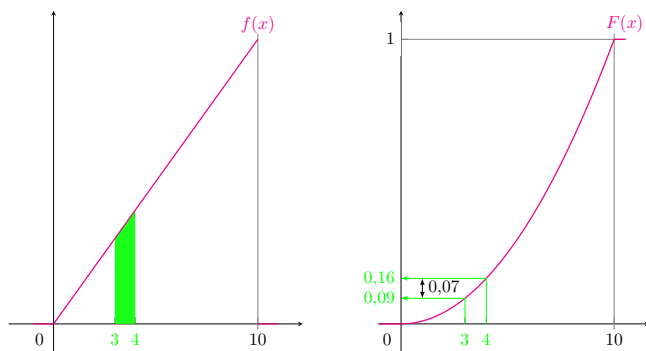
$$b) \quad \int_{-\infty}^{\infty} f_X(x) dx = 1,$$

$$c) \quad \mathbb{P}(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x) dx \text{ pro libovolná reálná čísla } a, b, \text{ kde } a \leq b.$$

Důkaz. Všechny uvedené vlastnosti plynou z definice 5.4 a z vlastností distribuční funkce. $\square$

Příklad 5.8. Uvažujeme hypotetickou populaci ryb. Je známo, že funkce dožití ryb závisí na kvadrátu délky života a že žádná ryba se nedožije více

⁸Z toho také plyne, že hodnoty V vypuštěné z uvažování mají nulovou pravděpodobnost.



Obrázek 5.6: Hustota, distribuční funkce a výpočet pravděpodobnosti k příkladu 5.8

než 10 let. Neboli $F(x)$ je dána vztahem

$$F(x) = \begin{cases} 0 & \text{pro } x \leq 0, \\ (c \cdot x)^2 & \text{pro } 0 < x \leq 10, \quad c \in \mathbb{R}^+, \\ 1 & \text{pro } x > 10. \end{cases}$$



Nechť X je náhodná veličina reprezentující věk ryby v čase úmrtí.

- Určete konstantu c tak, aby $F(x)$ byla distribuční funkce náhodné veličiny X .
- Vypočítejte hustotu $f(x)$ dožití v rybí populaci.
- Spočítejte pravděpodobnost, že ryba zemře mezi 3. a 4. rokem života.

Řešení.

- Musí platit, že $F(10) = 1$, tedy $(c \cdot 10)^2 = 1$, tedy $c = \frac{1}{10}$, tedy

$$F(x) = \begin{cases} 0 & \text{pro } x \leq 0, \\ \frac{x^2}{100} & \text{pro } 0 < x \leq 10, \\ 1 & \text{pro } x > 10. \end{cases}$$

- $f(x) = \left(\frac{x^2}{100}\right)' = \frac{x}{50}$, tedy

$$f(x) = \begin{cases} \frac{x}{50} & \text{pro } 0 < x \leq 10, \\ 0 & \text{jinak.} \end{cases}$$

c) Pravděpodobnost můžeme spočítat buď z hustoty

$$\mathbb{P}(3 \leq X \leq 4) = \int_3^4 f(x) dx = \int_3^4 \frac{1}{50} x dx = \frac{1}{50} \left[\frac{x^2}{2} \right]_3^4 = 0,07,$$

nebo přímo z distribuční funkce

$$\mathbb{P}(3 \leq X \leq 4) = F(4) - F(3) = \frac{16}{100} - \frac{9}{100} = 0,07.$$

Výpočet je ilustrován na obrázku 5.6.

△

Všimněte si ještě jednou, že oba přístupy řešení poslední otázky příkladu 5.8 jsou totožné: jak bylo řečeno výše, distribuční funkce je jednou z primitivních funkcí hustoty pravděpodobnosti. Tudíž výpočet pravděpodobnosti z hustoty (matematicky výpočet Riemannova určitého integrálu) je nutně ekvivalentní výpočtu pomocí distribuční funkce (matematicky výpočet Newtonova integrálu), neboť díky vlastnostem hustoty pravděpodobnosti oba integrály existují a musí si tedy být rovny.

Ukažme nyní řešení podobného příkladu, vyjdeme však ze známé hustoty pravděpodobnosti.

Příklad 5.9. Určete koeficient c tak, aby funkce

$$f(x) = \begin{cases} c \cdot x^2 \cdot e^x & \text{pro } 0 \leq x \leq 1, \\ 0 & \text{jinde,} \end{cases}$$

byla hustotou nějaké náhodné veličiny X . Vypočítejte distribuční funkci $F(x)$.

Řešení. Musíme najít c tak, aby

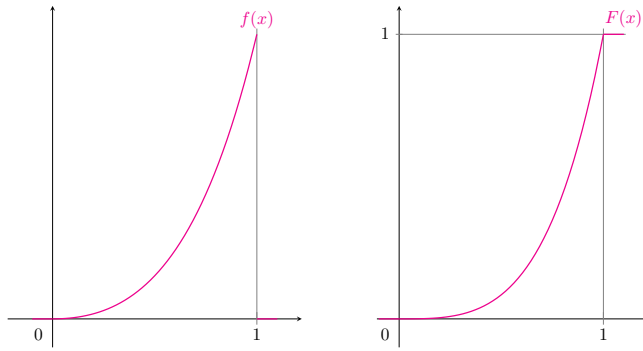
$$\int_{-\infty}^{\infty} f(x) dx = 1.$$

Ze zadání funkce $f(x)$ a s opakovaným využitím metody per partes dostáváme

$$\int_{-\infty}^{\infty} f(x) dx = \int_0^1 c \cdot x^2 \cdot e^x dx = c \cdot (e - 2) = 1, \text{ tedy } c = \frac{1}{e - 2}.$$

Tedy předpis hustoty náhodné veličiny X je

$$f(x) = \begin{cases} \frac{1}{e-2} \cdot x^2 \cdot e^x & \text{pro } 0 \leq x \leq 1, \\ 0 & \text{jinde.} \end{cases}$$



Obrázek 5.7: Hustota a distribuční funkce k příkladu 5.9

Distribuční funkce na intervalu $[0; 1]$ je

$$\begin{aligned} \int_0^x \frac{1}{e-2} \cdot t^2 \cdot e^t dt &= \frac{1}{e-2} \cdot [t^2 \cdot e^t - 2te^t + 2e^t]_0^x = \\ &= \frac{1}{e-2} \cdot (x^2 e^x - 2xe^x + 2e^x - 2). \end{aligned}$$

Kompletní předpis distribuční funkce tedy je

$$F(x) = \begin{cases} 0 & \text{pro } x < 0, \\ \frac{1}{e-2} \cdot (x^2 e^x - 2xe^x + 2e^x - 2) & \text{pro } 0 \leq x < 1, \\ 1 & \text{pro } x \geq 1. \end{cases}$$

Distribuční funkce i hustota pravděpodobnosti jsou zobrazeny na obrázku 5.7. △

5.6 Zobecnění[★]

Teorie míry zavádí pro účely měření „velikosti“ množin pojem *množinové míry*. Míra je nějaká σ -aditivní množinová funkce na prostoru $(\Omega, \mathcal{A})$, tj.

- a) $\mu: \mathcal{A} \rightarrow [0, \infty]$,
- b) $\mu(\emptyset) = 0$,

c) jsou-li $A_n \in \mathcal{A}$, $n \geq 1$ po dvou disjunktní, pak $\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n)$.

Je-li navíc $\mu(\Omega) = 1$, říkáme, že μ je *pravděpodobnostní míra*.

Pro účely charakterizace rozdělení náhodné veličiny ve smyslu zobrazení do prostoru $(\mathbb{R}, \mathcal{B})$ tedy na tomto prostoru zavádíme takovou pravděpodobnostní míru: každé borelovské množině $B \in \mathcal{B}$ lze přiřadit pravděpodobnostní míru na $(\mathbb{R}, \mathcal{B})$

$$\mu_X(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\}),$$

kteřou nazýváme rozdělení pravděpodobnosti náhodné veličiny X . Položíme-li speciálně $B = (-\infty, x]$, dostáváme distribuční funkci

$$\mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\}) = F_X(x).$$

Vidíme, že mezi distribuční funkcí a rozdělením pravděpodobností existuje vzájemně jednoznačný vztah.

Položíme-li $B = (a, b]$ pro $-\infty < a \leq b < \infty$, dostáváme

$$\mathbb{P}(X \in (a, b]) = F_X(b) - F_X(a) = \mu_F([a, b)).$$

Tímto vztahem je jednoznačně definována Lebesgueova–Stieltjesova míra indukovaná distribuční funkcí F , resp. náhodnou veličinou X . Výše uvedený vztah je definován pouze pro intervaly, ovšem to nám stačí pro jednoznačné definování míry pro všechny Borelovské množiny. Neboli platí

$$\mathbb{P}(X \in B) = \mu_{F_X}(B) = \int_B d\mu_{F_X} \text{ pro libovolný náhodný jev } B \in \mathcal{B}$$

a

$$\int_{\Omega} \phi(X(\omega)) d\mathbb{P}(\omega) = \int_{\mathbb{R}} \phi(x) d\mu_{F_X}(x)$$

pro libovolnou měřitelnou funkci ϕ , pro kterou existuje alespoň jeden z integrálů.

Integrál podle Lebesguovy–Stieltjesovy míry se někdy zkráceně zapisuje

$$\int_{\mathbb{R}} \phi(x) dF_X(x).$$

Jelikož není naším záměrem vykládat teorii míry v tomto textu, spokojíme se s intuitivním objasněním výše uvedených definic.

Je-li míra λ definována vztahem

$$\lambda([a, b]) = b - a, \quad -\infty < a \leq b < \infty,$$

pak se míře λ říká Lebesgueova míra. Tato míra není konečná, a tudíž není pravděpodobnostní. Integrál podle Lebesgueovy míry $\int \phi(x) d\lambda(x)$ je zobecněním Riemannova integrálu $\int \phi(x) dx$. Můžeme tedy s ním i tak pracovat.

Definujeme-li míru μ na omezeném intervalu $[A; B]$ vztahem

$$\mu([a; b]) = \frac{b - a}{B - A}, \quad A \leq a \leq b \leq B,$$

pak μ je již pravděpodobnostní míra.

Integrál podle míry μ pak vypočteme jako

$$\int \phi(x) d\mu(x) = \int \phi(x) \frac{1}{B - A} I_{[A, B]} x dx.$$

V obou těchto případech mají všechna x stejnou váhu. Lebesgueův–Stieltjesův integrál nám navíc umožňuje dát různým x různou váhu, která je určena přírůstkem distribuční funkce F_X .

Je-li X náhodná veličina definována v příkladu 5.5, pak distribuční funkce F_X má pouze dva skoky a žádné jiné přírůstky. Integrál podle μ_{F_X} je vlastně

$$\int \phi(x) d\mu_{F_X}(x) = \frac{5}{6}\phi(0) + \frac{1}{6}\phi(1).$$

Nemá-li naopak F_X žádné skoky a je spojitá v každém bodě, má pak F_X derivaci f v každém bodě, a ta ukazuje přírůstek F_X . Integrál podle μ_{F_X} je pak pouze

$$\int \phi(x) d\mu_{F_X}(x) = \int \phi(x) f(x) dx.$$

Lebesgueův–Stieltjesův integrál je tedy zobecněním integrálu, kde x mají různou váhu, a sumy, kde jednotlivé sčítance mají také různou váhu.

Tyto úvahy nás vedou k rozdělení náhodných veličin na dva základní typy, na diskrétní a absolutně spojitě náhodné veličiny. Mohou se však vyskytovat i jejich kombinace.

Příklad 5.10. Raketa se zaměřovacím systémem zasahuje s přesností 0 až 30 metrů od zaměřeného cíle, přičemž předpokládáme, že zasažené vzdálenosti od cíle jsou všechny stejně možné (jde tedy o geometrickou pravděpodobnost). Zaměřujeme střed budovy o průměru 10 metrů. Raketa způsobí stoprocentní poškození, pokud budovu jakkoliv zasáhne, poškození se kontinuálně snižuje o 10 % za každý metr od budovy. Jaká je pravděpodobnost, že budova bude po vystřelení rakety poškozena alespoň z 50 %?

Řešení. Zavedme nejprve pomocnou náhodnou veličinu Y s hodnotami v intervalu $[0; 30]$ reprezentující odchylku dopadu rakety od zaměřeného cíle

v metrech. Dále zavedme náhodnou veličinu X reprezentující způsobené poškození v procentních bodech. Platí zřejmě:

$$X = \begin{cases} 0 & \text{jestliže } Y \in [15; 30], \\ 10 \cdot (15 - Y) & \text{jestliže } Y \in (5; 15), \\ 100 & \text{jestliže } Y \in [0; 5]. \end{cases}$$

Dle principů geometrické pravděpodobnosti (viz odstavec 2.3 na straně 29) je

$$\mathbb{P}(X = 0) = \mathbb{P}(Y \in [15; 30]) = \frac{15}{30} = \frac{1}{2},$$

$$\mathbb{P}(X = 100) = \mathbb{P}(Y \in [0; 5]) = \frac{5}{30} = \frac{1}{6}.$$



Distribuční funkce F_X náhodné veličiny X má tedy dva skoky v bodech 0 a 100 o velikostech $\frac{1}{2}$ a $\frac{1}{6}$. V ostatních bodech je distribuční funkce spojitá, přičemž

$$\mathbb{P}(X \in (0; 100)) = \mathbb{P}(Y \in (5; 15)) = 1 - \left(\frac{1}{2} + \frac{1}{6} \right) = \frac{1}{3}$$

a tudíž

$$F_X(x) = \begin{cases} 0 & \text{pro } x < 0, \\ \frac{1}{2} + \frac{x}{300} & \text{pro } 0 \leq x < 100, \\ 1 & \text{pro } x \geq 100. \end{cases}$$

Pravděpodobnost, že budova bude zničena alespoň z 50 %, je tedy

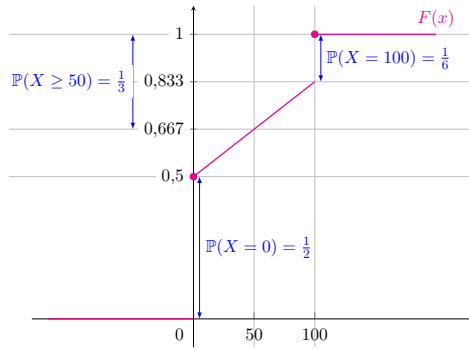
$$\mathbb{P}(X \geq 50) = \mu_F([50, 100]) = \int_{50-}^{100+} d\mu_F = \int_{50}^{100} \frac{1}{300} dx + \frac{1}{6} = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$$

Kompletní výpočet je ilustrován na obrázku 5.8. △

Nyní si ještě uvedme existenční větu pro distribuční funkci, tj. větu, která nám bude říkat, pro jakou funkci reálné proměnné existuje pravděpodobnostní prostor a na něm náhodná veličina tak, aby daná funkce byla její distribuční funkcí.

Věta 5.3. *Nechť funkce $F: \mathbb{R} \rightarrow \mathbb{R}$ má vlastnosti a), b), c) z věty 5.1. Pak existuje pravděpodobnostní prostor $(\Omega, \mathcal{A}, \mathbb{P})$ a na něm definovaná náhodná veličina X taková, že*

$$F_X(x) = F(x).$$



Obrázek 5.8: Distribuční funkce k příkladu 5.10

Důkaz. Položme

$$\Omega = \mathbb{R}, \quad \mathcal{A} = \mathcal{B}, \quad \mathbb{P} = \mu_F,$$

kde μ_F je Lebesgueův–Stieltjesova míra indukovaná funkcí F . Ještě položíme $X(\omega) = \omega$. Nyní máme

$$\begin{aligned} F_X(x) &= \mathbb{P}(X \leq x) = \mu_F((-\infty, x]) = \lim_{a \rightarrow -\infty} \mu_F((a, x]) \\ &= \lim_{a \rightarrow -\infty} (F(x) - F(a)) = F(x), \quad x \in \mathbb{R}, \end{aligned}$$

kde jsme použili spojitost konečné míry μ_F a vlastnost (b) distribuční funkce. $\square$

5.7 Úlohy

5.1 Necht množina elementárních jevů je $\Omega = \{\omega_1, \omega_2\}$, σ -algebru $\mathcal{A}$ definujeme na množině Ω takto:

$$\mathcal{A} = \{\emptyset, \{\omega_1\}, \{\omega_2\}, \Omega\}.$$

Zjistěte, zda funkce X daná předpisem $X(\omega_1) = 1$, $X(\omega_2) = 0$ je náhodná veličina.

5.2 Hodme jedenkrát kostkou. Množina elementárních jevů je

$$\Omega = \{\square, \begin{smallmatrix} \square & \square \\ \square & \square \end{smallmatrix}, \begin{smallmatrix} \square & \square \\ \square & \square \end{smallmatrix}, \begin{smallmatrix} \square & \square \\ \square & \square \end{smallmatrix}, \begin{smallmatrix} \square & \square \\ \square & \square \end{smallmatrix}, \begin{smallmatrix} \square & \square \\ \square & \square \end{smallmatrix}\}, \quad (5.7)$$

σ -algebru podmnožin množiny Ω definujeme takto:

$$\mathcal{A} = \{\emptyset, \{\square\}, \{\begin{smallmatrix} \square & \square \\ \square & \square \end{smallmatrix}, \begin{smallmatrix} \square & \square \\ \square & \square \end{smallmatrix}, \begin{smallmatrix} \square & \square \\ \square & \square \end{smallmatrix}, \begin{smallmatrix} \square & \square \\ \square & \square \end{smallmatrix}\}, \Omega\}.$$

Najděte funkci, která bude náhodnou veličinou vzhledem k $\mathcal{A}$.

5.3 V USA je známá populární hra *Chuck-a-luck*, ve které se hraje se třemi běžnými kostkami s puntíky na stěnách od 1 do 6. Michal hraje s Honzou hru, ve které Michal vsadí 1 € a vybere si určitý výsledek hodu, například, že na kostce padne stěna se třemi puntíky (padne číslo tři). Pokud se tak po hodu třemi kostkami nestane na žádné z kostek, ztratí Michal svůj vklad. Pokud padne číslo tři na právě jedné kostce, vyplatí Honza Michalovi 2 €, pokud padne číslo tři na právě dvou kostkách, dostane Michal 3 € a pokud padne číslo tři na všech třech kostkách, dostane Michal 4 €. Popište náhodnou veličinu, která vyjadřuje výši čistého zisku s ohledem na počet padnutých čísel tři, a určete rozdělení pravděpodobnosti této náhodné veličiny.⁹

5.4 Náhodná veličina X má distribuční funkci

$$F_X(x) = \begin{cases} 0 & \text{pro } x < 1, \\ (x-1)^2 & \text{pro } 1 \leq x < 2, \\ 1 & \text{pro } x \geq 2. \end{cases}$$

- Určete hustotu náhodné veličiny X .
- Znáznorněte graficky hustotu a distribuční funkci náhodné veličiny X .
- Určete $\mathbb{P}(1,5 < X < 1,75)$.

⁹Podle [6], str. 281, 282.

5.5 Je dána distribuční funkce

$$F(x) = a + b \cdot \arctan \frac{x}{2}, \quad -\infty < x < \infty.$$

Určete:

- a) pro jaké hodnoty a, b je F distribuční funkce,
- b) hustotu pravděpodobnosti f ,
- c) $\mathbb{P}(\alpha < X < \beta)$.

5.6 Graf hustoty náhodné veličiny X je lomená čára tvořená třemi body $A = [0; 0]$, $B = [5; y_B]$ a $C = [10; 0]$. Najděte souřadnici y_B bodu B tak, aby šlo o hustotu pravděpodobnosti. Určete vzorec a sestrojte graf hustoty pravděpodobnosti $f(x)$ a distribuční funkce $F(x)$ určené danou lomenou čarou.

5.7 Náhodná veličina X má hustotu

$$f(x) = \begin{cases} 0 & \text{pro } x < -4, \\ \frac{1}{\pi} \cdot \frac{1}{\sqrt{4^2 - x^2}} & \text{pro } -4 \leq x < 4, \\ 0 & \text{pro } x \geq 4. \end{cases}$$

Určete distribuční funkci F_X .

5.8 Hustota pravděpodobnosti náhodné veličiny X má tvar

$$f(x) = \begin{cases} 0 & \text{pro } x \leq -1, \\ \frac{a}{\sqrt{1-x^2}} & \text{pro } -1 < x < 1, \\ 0 & \text{pro } x \geq 1. \end{cases}$$

Určete:

- a) koeficient a ,
- b) $\mathbb{P}(-\frac{1}{2} < X < \frac{1}{2})$.

5.9 Zjistěte, pro jaká a, b je funkce

$$F(x) = \frac{1}{1 + e^{-(a+bx)}}, \quad x \in \mathbb{R}$$

distribuční funkcí.

Řešení úloh

5.1 Ano

5.2 Například $X(\blacksquare) = b$, $X(\blacksquare) = X(\blacksquare) = X(\blacksquare) = X(\blacksquare) = X(\blacksquare) = c$, kde b, c jsou libovolné reálné konstanty.

5.3 $X: \Omega \rightarrow \mathbb{R}$, $\Omega = \{(a, b, c) \in \{\blacksquare, \blacksquare, \dots, \blacksquare\}^3\}$, $|\Omega| = 6^3 = 216$.

$$X(a, b, c) = \begin{cases} 3 & \text{pro } a = b = c = \blacksquare, \\ 2 & \text{pokud právě dvě z } a, b, c \text{ jsou } \blacksquare, \\ 1 & \text{pokud právě jedno z } a, b, c \text{ je } \blacksquare, \\ -1 & \text{jinak.} \end{cases}$$

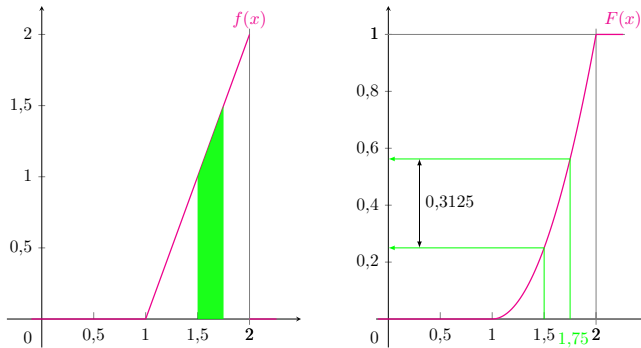
$$p(x) = \begin{cases} \frac{1}{216} \doteq 0,005 & \text{pro } x = 3, \\ \frac{15}{216} \doteq 0,069 & \text{pro } x = 2, \\ \frac{75}{216} \doteq 0,347 & \text{pro } x = 1, \\ \frac{125}{216} \doteq 0,579 & \text{jinak.} \end{cases}$$

5.4

a) $f(x) = \begin{cases} 2x - 2 & \text{pro } 1 \leq x \leq 2, \\ 0 & \text{jinak,} \end{cases}$

b) viz obrázek 5.9,

c) $\mathbb{P}(1,5 < X < 1,75) = 0,3125$.

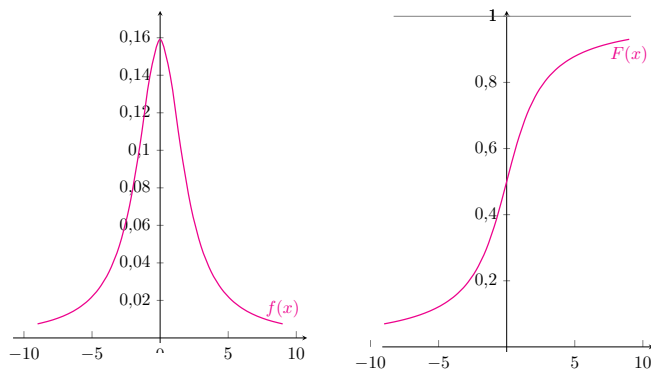


Obrázek 5.9: Graf hustoty a distribuční funkce z úlohy 5.4

5.5 a) $a = \frac{1}{2}$, $b = \frac{1}{\pi}$, b) $f_X(x) = \frac{2}{\pi} \cdot \left(\frac{1}{4+x^2} \right)$,

c) $\mathbb{P}(\alpha < X < \beta) = \frac{1}{\pi} \cdot \left(\arctan \frac{\beta}{2} - \arctan \frac{\alpha}{2} \right)$.

Jedná se o tzv. Cauchyho¹⁰ rozdělení, které se používá například ve financích pro modelování extrémních investičních rizik. Graf hustoty a distribuční funkce je na obrázku 5.10.



Obrázek 5.10: Graf hustoty a distribuční funkce z úlohy 5.5

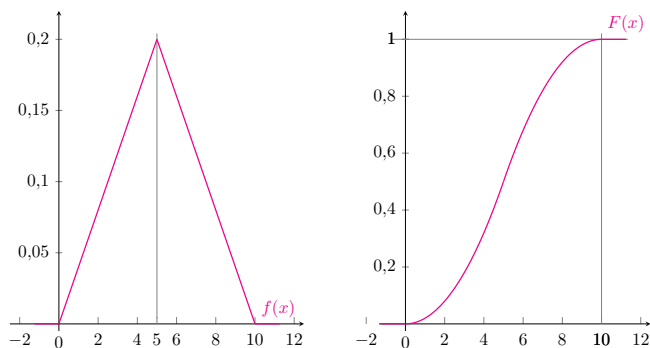
5.6 $y_B = 0,2$,

$$f(x) = \begin{cases} 0,04x & \text{pro } 0 \leq x < 5, \\ 0,4 - 0,04x & \text{pro } 5 \leq x < 10, \\ 0 & \text{jinak.} \end{cases}$$

$$F(x) = \begin{cases} 0 & \text{pro } x < 0, \\ 0,02x^2 & \text{pro } 0 \leq x < 5, \\ 0,4x - 0,02x^2 - 1 & \text{pro } 5 \leq x < 10, \\ 1 & \text{pro } x \geq 10. \end{cases}$$

Jedná se o tzv. trojúhelníkové rozdělení. Používá se například v simulačních studiích pro veličiny, u kterých známe pouze rozsah (dolní a horní mez) hodnot, jichž může nabývat, a její nejpravděpodobnější hodnotu (modus). Graf hustoty a distribuční funkce je na obrázku 5.11.

¹⁰Augustin-Louis Cauchy (1789–1857), francouzský matematik.

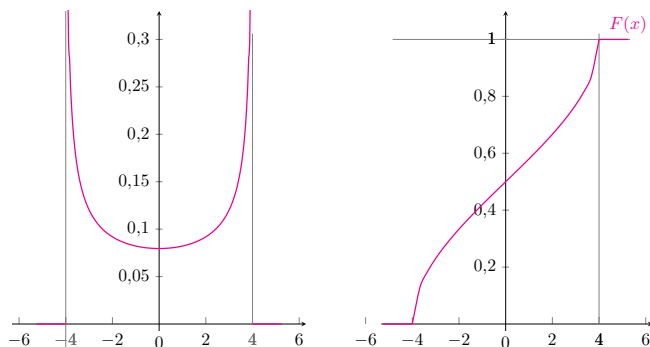


Obrázek 5.11: Graf hustoty a distribuční funkce z úlohy 5.6

5.7

$$F_X(x) = \begin{cases} 0 & \text{pro } x < -4, \\ \frac{2}{\pi} \arcsin \sqrt{\frac{x+4}{8}} & \text{pro } -4 \leq x < 4, \\ 1 & \text{pro } x \geq 4. \end{cases}$$

Jde o tzv. rozdělení arcsin (na obecném intervalu, viz obrázek 5.12), které má užití například v signální technice.



Obrázek 5.12: Graf hustoty a distribuční funkce z úlohy 5.7

5.8 a) $a = \frac{1}{\pi}$, b) $\frac{1}{3}$. Toto je rovněž rozdělení arcsin.

5.9 $a \in \mathbb{R}$, $b > 0$. Jedná se o tzv. logistické rozdělení, používané při modelování náhodných veličin reprezentujících kategorické charakteristiky.

Kapitola 6

Charakteristiky náhodných veličin

Pravděpodobnostní chování náhodné veličiny je úplně popsáno pravděpodobnostní funkcí, hustotou pravděpodobnosti, nebo distribuční funkcí. Úplná charakterizace náhodné veličiny může být ovšem někdy strukturně složitá nebo v některých případech dokonce nedostupná (např. se špatně nebo velmi nepřesně odhaduje). V praktických aplikacích však často postačuje shrnout informaci o rozdělení náhodné veličiny do několika vhodných číselných charakteristik, které dostatečně výstižně popisují základní vlastnosti tohoto rozdělení. V takových případech pak úplná charakterizace náhodné veličiny pomocí pravděpodobnostní funkce či hustoty a distribuční funkcí není ani potřebná. V této kapitole se budeme zabývat pouze tzv. charakteristikami polohy (střední hodnota, kvantil, medián, modus) a charakteristikami variability (rozptyl, směrodatná odchylka) a způsoby jejich výpočtu.

6.1 Střední hodnota náhodné veličiny a její vlastnosti

Definice 6.1 (Střední hodnota náhodné veličiny)

Nechť X je náhodná veličina definovaná na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, \mathbb{P})$ a F_X její distribuční funkce. Střední hodnotou $\mathbb{E}X$ náhodné veličiny X nazveme integrál

$$\mathbb{E}X = \int_{-\infty}^{\infty} x \, dF_X(x), \quad (6.1)$$

pokud tento integrál existuje.

Střední hodnota $\mathbb{E}X$ náhodné veličiny X se často také nazývá očekávaná hodnota a představuje číslo, které charakterizuje polohu hodnot náhodné veličiny na číselné ose s ohledem na jejich pravděpodobnosti. Integrál ve vzorci (6.1) je Lebesgueova–Stieltjesova typu (viz odstavec 5.6). Dále uvedeme speciální tvar definice 6.1 v případě, že náhodná veličina má diskrétní, resp. absolutně spojitě rozdělení.

Věta 6.1.

a) *Nechť X je diskrétní náhodná veličina nabývající pro $i = 1, 2, \dots$ hodnot x_i s pravděpodobnostmi $\mathbb{P}(X = x_i) = p_i$. Pak střední hodnota $\mathbb{E}X$ náhodné veličiny X je tvaru*

$$\mathbb{E}X = \sum_{i=1}^{\infty} x_i \cdot p_i, \quad (6.2)$$

pokud řada v (6.2) konverguje.

b) *Nechť X je absolutně spojitá náhodná veličina s hustotou f_X . Pak střední hodnota náhodné veličiny X je*

$$\mathbb{E}X = \int_{-\infty}^{\infty} x f_X(x) dx, \quad (6.3)$$

pokud integrál konverguje.

Důkaz. Viz odstavec 5.6. □

Všimněte si, že střední hodnota je váženým součtem hodnot náhodné veličiny. Roli vah hraje pravděpodobnostní funkce v diskrétním, resp. hustota pravděpodobnosti ve spojitém případě.

Příklad 6.1. Pro diskrétní náhodnou veličinu X , která udává počet padnutých líců v pěti hodech mincí (viz příklad 5.7), vypočítejte její střední hodnotu $\mathbb{E}X$.

Řešení. Střední hodnota se vypočítá dle vzorce (6.2)

$$\mathbb{E}X = \sum_{i=1}^5 x_i \cdot p_i = 0 \cdot 0,031\,25 + 1 \cdot 0,156\,25 + \dots + 5 \cdot 0,031\,25 = 2,5.$$

Tento výsledek lze snadno spočítat i v Excelu pomocí funkce skalárního součinu. Jsou-li například hodnoty náhodné veličiny (první řádek tabulky 5.2) umístěny v buňkách B1 až G1 a příslušející pravděpodobnosti (druhý řádek tabulky 5.2) v buňkách B2 až G2, potom je střední hodnota $\mathbb{E}X$ výsledkem funkce = SOUČIN.SKALÁRNÍ(B1:G1;B2:G2). △

Příklad 6.2. Uvažujme absolutně spojitou náhodnou veličinu X , která má hustotu $f(x) = \ln x$ pro $x \in [1; e]$ a 0 jinde. Vypočítejte střední hodnotu $\mathbb{E}X$.

Řešení. Nejprve ověříme, že $f(x)$ je hustota, tj. zda platí $\int_{-\infty}^{\infty} f(x) dx = 1$:

$$\int_{-\infty}^{\infty} f(x) dx = \int_1^e \ln x dx = [x \ln x - x]_1^e = 1.$$

Střední hodnota $\mathbb{E}X$ je dle vzorce (6.3) rovna hodnotě integrálu

$$\int_1^e x \cdot \ln x dx = \left[\frac{x^2}{2} \cdot \left(\ln x - \frac{1}{2} \right) \right]_1^e = \frac{e^2}{4} + \frac{1}{4} \doteq 2,097.$$

△

Příklad 6.3. Uvažujme absolutně spojitou náhodnou veličinu X , která má hustotu $f(x) = \frac{1}{\pi(1+x^2)}$ (jde o tzv. Cauchyho rozdělení). Vypočtěte $\mathbb{E}X$.

Řešení. Nejprve je vhodné ověřit, že funkce f je skutečně hustotou pravděpodobnosti, tj. že platí

$$\int_{-\infty}^{\infty} f(x) dx = \int_{-\infty}^{\infty} \frac{1}{\pi} \frac{1}{1+x^2} dx = \frac{1}{\pi} [\arctan x]_{-\infty}^{\infty} = 1.$$

Vidíme tedy, že f je hustota. Pokusíme se vypočítat $\mathbb{E}X$ dle vzorce (6.3):

$$\mathbb{E}X = \int_{-\infty}^{\infty} \frac{1}{\pi} \frac{x}{1+x^2} dx.$$

Integrál na pravé straně vzorce ovšem diverguje. Toto pravděpodobnostní rozdělení je tedy příkladem rozdělení, pro které střední hodnota neexistuje.

△

Střední hodnota náhodné veličiny je podle definice 6.1 Lebesgueovým–Stieltjesovým integrálem, tudíž sdílí všechny jeho vlastnosti a platí tedy tvrzení následující věty, kterou uvádíme jinak bez důkazu.

Věta 6.2 (o vlastnostech střední hodnoty). *Nechť X, Y, X_n , $n = 1, 2, \dots$ jsou náhodné veličiny na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, \mathbb{P})$ a a, b jsou reálné konstanty. Potom*

a) střední hodnota konstanty je konstanta:

$$\mathbb{E}a = a,$$

b) střední hodnota je absolutně integrovatelná:

$$\mathbb{E}X < \infty \text{ právě tehdy, když } \mathbb{E}|X| < \infty,$$

c) střední hodnota je lineárním operátorem:

$$\mathbb{E}(aX + bY) = a\mathbb{E}X + b\mathbb{E}Y,$$

d) jestliže existuje střední hodnota součtu náhodných veličin X_i , pak je dána jako součet středních hodnot náhodných veličin X_i :

$$\mathbb{E}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathbb{E}X_i,$$

e) střední hodnota je nezáporná pro nezáporné náhodné veličiny:

$$\text{je-li } X \geq 0 \text{ s. j., pak } \mathbb{E}X \geq 0,$$

f) střední hodnota je monotónní:

$$\text{jestliže } X_1 \leq X \leq X_2 \text{ s. j., pak } \mathbb{E}X_1 \leq \mathbb{E}X \leq \mathbb{E}X_2,$$

g)

$$|\mathbb{E}X| \leq \mathbb{E}|X|,$$

h)

$$\text{jestliže } |X| \leq Y \text{ s. j., } \mathbb{E}Y < \infty, \text{ pak } \mathbb{E}X < \infty,$$

i) integrace člen po členu:

$$\text{jestliže } \sum_{n=1}^{\infty} \mathbb{E}|X_n| < \infty, \text{ pak } \mathbb{E}\left(\sum_{n=1}^{\infty} X_n\right) = \sum_{n=1}^{\infty} \mathbb{E}X_n,$$

j) platí Fatouovo¹ lemma:

$$\text{jestliže } X_n \geq 0 \text{ s. j., pak } \mathbb{E}(\liminf_{n \rightarrow \infty} X_n) \leq \liminf_{n \rightarrow \infty} \mathbb{E}X_n.$$

I když vlastnosti a)–j) vyplývají z obecných vlastností Lebesgueova–Stieltjesova integrálu, lze je v případě diskretních a absolutně spojitých veličin dokázat i přímo. Je-li X diskretní náhodná veličina, využijeme vlastností sumy. V případě, že X je absolutně spojitá náhodná veličina, využijeme vlastností Riemannova či Lebesgueova integrálu.

¹Pierre Fatou (1878–1929), francouzský matematik.

Příklad 6.4. Podle úmrtnostních tabulek USA (1978–1979) je pravděpodobnost úmrtí 32leté ženy během jednoho roku rovna 0,001 819. Pojišťovna nabízí ženám tohoto věku, že při ročním pojistném 100 USD vyplatí pozůstalým v případě úmrtí pojištěnce 25 000 USD. Jaký zisk může pojišťovna očekávat, jestliže takovou pojistku uzavře 5 000 žen uvedeného věku?

Řešení. Zisk (či ztrátu) pojišťovny v případě uzavření jedné pojistky označíme jako náhodnou veličinu X_i , $i = 1, \dots, 5\,000$. Její střední hodnota je

$$\mathbb{E}X_i = 100 \cdot 0,998\,181 - 24\,900 \cdot 0,001\,819 = 99,818\,1 - 45,293\,1 = 54,525.$$

Uzavře-li pojišťovna 5 000 takových pojistek, je její očekávaný zisk roven střední hodnotě náhodné veličiny

$$Y = \sum_{i=1}^{5\,000} X_i,$$

tedy

$$\mathbb{E}Y = \mathbb{E}\left(\sum_{i=1}^{5\,000} X_i\right) = 5\,000 \cdot \mathbb{E}X_i = 272\,625.$$

Očekávaný zisk pojišťovny je 272 625 USD. △

Jestliže definujeme novou náhodnou veličinu Y jako funkci nějaké existující náhodné veličiny X , tj. $Y = \phi(X)$, je možné k výpočtu střední hodnoty Y využít rozdělení X , jak uvádí následující věta. Není tudíž v tomto případě nutné odvozovat rozdělení náhodné veličiny Y , abychom mohli střední hodnotu spočítat.

Věta 6.3. *Nechť X je náhodná veličina a nechť $\phi: \mathbb{R} \rightarrow \mathbb{R}$. Pak platí*

$$\mathbb{E}\phi(X) = \int_{-\infty}^{\infty} \phi(x) \, dF_X(x),$$

pokud jeden z integrálů konverguje.

Má-li náhodná veličina X diskrétní rozdělení $\{x_n, p_n\}_{n \in N_0}$, pak

$$\mathbb{E}\phi(X) = \sum_{n \in N_0} \phi(x_n) p_n,$$

pokud jedna ze stran rovnosti konverguje.

Má-li náhodná veličina X absolutně spojitě rozdělení s hustotou f , pak

$$\mathbb{E}\phi(X) = \int_{-\infty}^{\infty} \phi(x) f(x) \, dx,$$

pokud jeden z integrálů konverguje.

6.2 Další charakteristiky polohy náhodných veličin

Střední hodnotu náhodné veličiny řadíme mezi vlastnosti náhodných veličin, kterým souhrnně říkáme charakteristiky polohy. Další charakteristiky tohoto typu jsou odvozeny od tzv. kvantilové funkce.

Definice 6.2 (Kvantilová funkce, kvantil)

a) *Nechť F je distribuční funkce náhodné veličiny X . Zavedme funkci F^{-1} předpisem*

$$F^{-1}(u) = \inf\{x : F(x) \geq u\}, \quad 0 < u < 1. \quad (6.4)$$

Pak F^{-1} nazýváme kvantilová funkce odpovídající distribuční funkci F .

b) *Hodnoty funkce $F^{-1}(u)$ se nazývají kvantily. Konkrétně α -kvantilem nazýváme hodnotu $F^{-1}(\alpha)$.*

Z definice vyplývá, že pokud je F rostoucí a spojitá, pak kvantilová funkce F^{-1} je inverzní funkcí k F . Odtud pochází i označení F^{-1} , které se ovšem používá i pro zbylé případy². Kvantily také často značíme pomocí procent, protože α je číslo mezi nulou a jedničkou. V tomto kontextu se také označují jako *percentily*.

Nejdůležitějším kvantilem je tzv. *medián*; označujeme jej $\tilde{X}$ nebo $\text{Me}(X)$. Medián je 50% kvantil, tj. $F^{-1}(0,50)$. Hodnoty náhodné veličiny rozdělují na dvě části, které mají stejnou 50% pravděpodobnost:

$$\mathbb{P}(X \leq \tilde{X}) \geq \frac{1}{2} \quad \text{a} \quad \mathbb{P}(X \geq \tilde{X}) \geq \frac{1}{2}. \quad (6.5)$$

Medián se užívá všude tam, kde vážení hodnot pomocí pravděpodobnosti (tedy užití středních hodnot) selhává. Typicky tomu je v případech rozdělení s těžkými chvosty, jako je třeba Cauchyho rozdělení z příkladu 6.3 (ve kterém střední hodnota neexistuje vůbec), anebo rozdělení modelujících situace s výskyty extrémních hodnot – má-li tato extrémní hodnota nezanedbatelnou pravděpodobnost, vychýlí střední hodnotu směrem k sobě. Ukázkovým příkladem je střední hodnota mzdy náhodně vybraného zaměstnance nějaké firmy: jedna vysoká manažerská mzda velmi významně zkreslí střední hodnotu mezd ostatních zaměstnanců směrem k hodnotě, které většina zaměstnanců firmy se svoji mzdou nedosahuje. V takových případech je medián

²V případě, že je distribuční funkce na nějakém intervalu $(a; b)$ konstantní, inverzní funkce není pro příslušnou hodnotu definována, neboť ji není možné určit jednoznačně: je principiálně možné za kvantil volit kterýkoliv bod tohoto intervalu. Infimální vzorec (6.4) definuje jako kvantil dolní mez, ale v učebnicích je možné najít i supremální verzi.

lepší charakteristikou polohy náhodné veličiny: v případě mezd udává hodnotu, na kterou dosáhne přesně polovina zaměstnanců firmy. Poznamenejme ještě, že medián není podmínkou (6.5) určen jednoznačně (viz úlohu 6.2).

Hodnoty náhodné veličiny je samozřejmě možné pomocí kvantilů dělit i na díly nestejné velikosti: například 25% kvantil $F^{-1}(0,25)$ rozdělí hodnoty náhodné veličiny na dolní čtvrtinu a horní tři čtvrtiny (z hlediska pravděpodobnosti dosažení). Tento kvantil nazýváme *dolním kvantilem*. Obdobně *horní kvantil* je 75% kvantil $F^{-1}(0,75)$. Menší, resp. větší kvantily zase najdou uplatnění ve statistice, konkrétně při určování intervalových odhadů (viz kapitolu 14) a při testování hypotéz (viz kapitolu 15).

Příklad 6.5. Označme dobu čekání rybáře na úlovek (v minutách) jako náhodnou veličinu X . Určete střední hodnotu, dolní kvartil, medián a horní kvartil doby čekání rybáře na úlovek, má-li náhodná veličina X hustotu pravděpodobnosti ve tvaru

$$f(x) = \begin{cases} e^{-x} & \text{pro } 0 < x < \infty, \\ 0 & \text{jinak.} \end{cases} \quad (6.6)$$

Řešení. Rozdělení, jehož hustota je dána rovnicí (6.6), se nazývá exponenciální (více viz odstavec 8.2 na straně 145). Střední hodnotu určíme integrací per partes

$$\mathbb{E}X = \int_0^{\infty} x \cdot e^{-x} dx = [x \cdot (-e^{-x})]_0^{\infty} - \int_0^{\infty} (-e^{-x}) dx = 0 + [-e^{-x}]_0^{\infty} = 1.$$

Pro výpočet dolního kvartilu, mediánu a horního kvartilu nejdříve určíme distribuční funkci, stačí pro $x > 0$:

$$F(x) = \int_0^x e^{-t} dt = [-e^{-t}]_0^x = 1 - e^{-x}$$

(pro $x \leq 0$ je $F(x)$ nulová). Následně odvodíme i kvantilovou funkci jako funkci inverzní k $F(x)$:

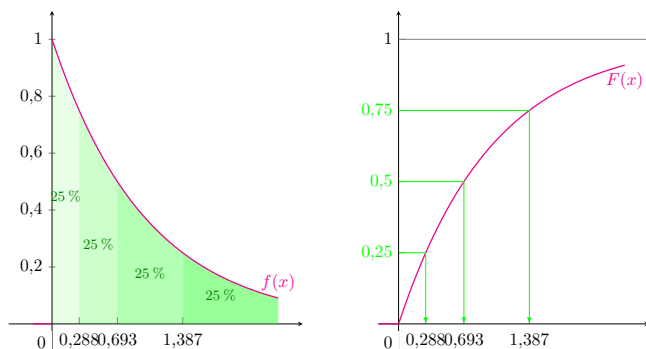
$$F^{-1}(u) = -\ln(1 - u).$$

Dosadíme za u postupně 0,25, 0,5 a 0,75. Dostáváme

$$F^{-1}(0,25) \doteq 0,288, \quad F^{-1}(0,5) \doteq 0,693, \quad F^{-1}(0,75) \doteq 1,387.$$

Hodnoty jsou v minutách, vynásobením výsledků hodnotou 60 dostaneme časy ve vteřinách. Význam těchto kvantilů je ilustrován na obrázku 6.1.

△



Obrázek 6.1: První, druhý a třetí kvartil náhodné veličiny X z příkladu 6.5

Poslední charakteristikou polohy rozdělení náhodné veličiny X , kterou zde zmíníme, je tzv. *modus*, který se obvykle značí $\hat{X}$, případně $\text{Mod}(x)$ nebo $\text{Mo}(X)$. Modus zjednodušeně řečeno představuje „nejpravděpodobnější“ hodnotu, které může náhodná veličina nabýt. Je-li diskrétní rozdělení soustředěno v bodech $x_1, x_2, \dots$, je modem $\hat{X}$ ta hodnota, pro kterou platí

$$\mathbb{P}(X = \hat{X}) \geq \mathbb{P}(X = x_i)$$

pro libovolné $i = 1, 2, \dots$. Je to tedy hodnota této veličiny, ve které nabývá pravděpodobnostní funkce $p(x)$ svého maxima. Pro absolutně spojitou náhodnou veličinu je modem $\hat{X}$ ta hodnota, ve které nabývá svého maxima hustota pravděpodobnosti $f(x)$, tj.

$$f(\hat{X}) \geq f(x)$$

pro libovolné $x \in (-\infty, \infty)$. Ani modus nemusí být určen jednoznačně, náhodná veličina může mít dva i více modů (viz úlohu 6.3).

Příklad 6.6. Určete modus $\hat{X}$ následujících náhodných veličin:

a) diskrétní veličiny X s rozdělením pravděpodobnosti

$$p_n = \begin{cases} \left(\frac{1}{2}\right)^n & \text{pro } n = 1, 2, \dots, \\ 0 & \text{jinak,} \end{cases}$$

b) spojitě náhodné veličiny s hustotou

$$f(x) = \begin{cases} \frac{x^2 e^{-x}}{2} & \text{pro } x \in (0; \infty), \\ 0 & \text{jinak.} \end{cases}$$

Řešení.

a) Vidíme, že

$$p_1 = \frac{1}{2}, \quad p_2 = \frac{1}{4}, \dots$$

Jelikož pro rostoucí n pravděpodobnost p_n geometricky klesá, je modus $\hat{X} = 1$.

b) Ke stanovení $\hat{X}$ v případě spojitého rozdělení je nutné najít maximum hustoty. Vyřešíme tedy průběh funkce hustoty $f(x)$. Jelikož v krajních bodech intervalu $(0; \infty)$ je limita $f(x)$ rovna nule, vyšetříme body, kde $f'(x) = 0$.

$$f'(x) = xe^{-x} - \frac{x^2 e^{-x}}{2} = xe^{-x} \left(1 - \frac{x}{2}\right) = 0.$$

Jediným takovým bodem uvnitř intervalu $(0; \infty)$ je $x = 2$. Jelikož

$$f''(2) = -e^{-2} < 0,$$

jedná se skutečně o maximum, a tedy modus $\hat{X} = 2$.

△

6.3 Rozptyl náhodné veličiny

Střední hodnota, medián, kvantily a ostatní charakteristiky polohy poskytují informaci, kde můžeme očekávat hodnoty náhodné veličiny, avšak neříkají samy o sobě nic o tom, jak daleko či blízko charakteristice polohy se tyto hodnoty vyskytují. Tuto informaci poskytnou tzv. charakteristiky měnlivosti neboli variability. Základní charakteristikou tohoto typu je rozptyl náhodné veličiny.

Definice 6.3 (Rozptyl náhodné veličiny)

Střední kvadratická odchylka od střední hodnoty náhodné veličiny, tj.

$$\text{var } X = \mathbb{E}(X - \mathbb{E}X)^2,$$

se nazývá rozptyl náhodné veličiny X .

Rozptyl je druhou nejdůležitější charakteristikou náhodné veličiny. Z jeho definice vidíme, že existence střední hodnoty je nutnou podmínkou k existenci rozptylu, takže například Cauchyho rozdělení z příkladu 6.3 nemůže mít vedle střední hodnoty ani rozptyl. Rozptyl náhodné veličiny je vždy nezáporný a rovná se nule právě tehdy, když $\mathbb{P}(X = c) = 1$, kde c je nějaká konstanta. Rozptyl má řadu pěkných matematických vlastností, které si shrneme v následující větě.

Věta 6.4 (o vlastnostech rozptylu). *Nechť je X náhodná veličina, a, c reálné konstanty. Potom*

a) *platí tzv. výpočetní vzorec rozptylu*

$$\text{var } X = \mathbb{E}(X^2) - (\mathbb{E}X)^2, \quad (6.7)$$

b) $\text{var } c = 0$,

c) $\text{var}(aX) = a^2 \text{var } X$,

d) $\text{var}(X + c) = \text{var } X$.

e) *Jestliže X má konečnou střední hodnotu a konečný nenulový rozptyl a*

$$Y = \frac{X - \mathbb{E}X}{\sqrt{\text{var } X}},$$

pak $\mathbb{E}Y = 0$ a $\text{var } Y = 1$.

Důkaz.

$$\begin{aligned} \text{a)} \quad \text{var } X &= \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}[X^2 - 2X\mathbb{E}X + (\mathbb{E}X)^2] = \\ &= \mathbb{E}X^2 - 2(\mathbb{E}X)^2 + (\mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2. \end{aligned}$$

b) Jelikož $\mathbb{E}c = c$, je nutně $\text{var}(c - \mathbb{E}c)^2 = \text{var}(c - c)^2 = 0$.

$$\begin{aligned} \text{c)} \quad \text{var}(aX) &= \mathbb{E}(aX - \mathbb{E}aX)^2 = \mathbb{E}[a(X - \mathbb{E}X)]^2 = \\ &= a^2 \mathbb{E}(X - \mathbb{E}X)^2 = a^2 \text{var } X. \end{aligned}$$

Části d) a e) ponecháme bez důkazu. □

Nevýhodou rozptylu je to, že udává variabilitu náhodné veličiny ve čtvrcích jejích jednotek a číselné hodnoty rozptylu jsou tak velmi obtížně interpretovatelné. Z tohoto důvodu se v praktických úlohách používá pro charakterizaci variability náhodných veličin tzv. *směrodatná odchylka*, což je druhá odmocnina rozptylu,

$$\sigma = \sqrt{\text{var } X}.$$

Směrodatná odchylka měří variabilitu náhodné veličiny v jejích původních jednotkách. Nesdílí však některé dobré vlastnosti rozptylu, se kterými se seznámíme v kapitole 9 (například směrodatná odchylka součtu nezávislých náhodných veličin *není* součtem směrodatných odchylek).

Příklad 6.7 (pokračování příkladu 6.5). Vraťme se k řešení příkladu o čekání rybáře na úlovek ze strany 99. Dobu čekání rybáře na úlovek (v minutách) jsme označili jako náhodnou veličinu X . Vypočítejte rozptyl a směrodatnou odchylku této náhodné veličiny X .

Řešení. Připomeňme, že náhodná veličina X má hustotu pravděpodobnosti

$$f(x) = \begin{cases} e^{-x} & \text{pro } 0 < x < \infty, \\ 0 & \text{jinak.} \end{cases}$$

Již jsme odvodili, že $\mathbb{E}X = 1$ minuta. Rozptyl vypočítáme podle vzorce (6.7) na straně 102. Nejdříve určíme $\mathbb{E}X^2$.

$$\begin{aligned} \mathbb{E}X^2 &= \int_0^\infty x^2 \cdot e^{-x} \, dx = [x^2 \cdot (-e^{-x})]_0^\infty - \int_0^\infty -2xe^{-x} \, dx = \\ &= 0 + 2 \cdot \int_0^\infty xe^{-x} \, dx = 2 \cdot \mathbb{E}X = 2. \end{aligned}$$

Odtud

$$\text{var } X = 2 - 1^2 = 1.$$

Rozptyl náhodné veličiny X je 1 min^2 , směrodatná odchylka je tedy 1 minuta. $\triangle$

Věta 6.5 (Čebyševova nerovnost³). *Nechť X je náhodná veličina s konečným rozptylem. Pak pro libovolné $\varepsilon > 0$ platí*

$$\mathbb{P}[|X - \mathbb{E}X| \geq \varepsilon] \leq \frac{\text{var } X}{\varepsilon^2}. \quad (6.8)$$

Větu 6.5 dokážeme v odstavci 6.4. Vzorec (6.8) můžeme přepsat do alternativní podoby: jestliže položíme $\varepsilon = k \cdot \sigma$, kde $k > 0$ a $\sigma = \sqrt{\text{var } X}$ je směrodatná odchylka náhodné veličiny X , potom

$$\mathbb{P}[|X - \mathbb{E}X| \geq k \cdot \sigma] \leq \frac{1}{k^2}. \quad (6.9)$$

Nerovnost (6.9) se nazývá klasická Čebyševova nerovnost a platí pro libovolné $k > 0$. Vzorec (6.9) poskytuje hrubý symetrický intervalový odhad pravděpodobnosti (říká se mu interval spolehlivosti), že náhodná veličina s konečným rozptylem bude v předem určené vzdálenosti od své střední hodnoty. Tato vzdálenost je vyjádřena jako k -násobek směrodatné odchylky náhodné veličiny. Pro účely odhadu nemá smysl volit $k \leq 1$. Nerovnost je nezávislá na rozdělení náhodné veličiny X , jedná se tedy o maximálně konzervativní odhad (fakticky horní mez pravděpodobnosti). Její podstatný teoretický význam spočívá v tom, že poslouží k důkazu Bernoulliho zákona velkých čísel (viz odstavec 11.2 na straně 221).

³Pafnutij Lvovič Čebyšev (1821–1894), ruský matematik.

Příklad 6.8. Předpokládejme, že spravujeme investiční portfolio s průměrným výnosem 8000 Kč a směrodatnou odchylkou 500 Kč. Jaká je nejvyšší možná pravděpodobnost, že se skutečný výnos bude od střední hodnoty lišit o více než $\pm 10\%$ průměrného výnosu?

Řešení. Rozdělení výnosu portfolia není známo, pro odhad tedy použijeme Čebyševovu nerovnost, která na rozdělení nezávisí. 10% z 8000 Kč je 800 Kč, tedy chceme nalézt

$$\mathbb{P}[|X - 8000| \geq 800].$$

Dle vzorce (6.9) odhadujeme

$$\mathbb{P}[|X - 8000| \geq 1,6 \cdot 500] \leq \frac{1}{1,6^2} \doteq 0,391.$$

Pokud bychom však znali rozdělení náhodné veličiny X , mohli bychom tento odhad významně snížit. $\triangle$

6.4 Momenty vyšších řádů

Definice střední hodnoty a rozptylu lze zobecnit prostřednictvím následující definice.

Definice 6.4 (n -tý moment)

Pro přirozené číslo n je n -tý moment náhodné veličiny X definován jako $\mathbb{E}(X^n)$; n -tý absolutní moment jako $\mathbb{E}(|X|^n)$; n -tý centrální moment jako $\mathbb{E}[(X - \mathbb{E}X)^n]$.

Využijeme-li Lebesgueovy–Stieltjesovy míry (viz odstavec 5.6), potom právě definované momenty můžeme vyjádřit obecně pomocí integrálů

$$\begin{aligned}\mathbb{E}(X^n) &= \int_{-\infty}^{\infty} x^n dF_X(x), \\ \mathbb{E}(|X|^n) &= \int_{-\infty}^{\infty} |x|^n dF_X(x), \\ \mathbb{E}[(X - \mathbb{E}X)^n] &= \int_{-\infty}^{\infty} (x - \mathbb{E}X)^n dF_X(x).\end{aligned}$$

Vzorce pro speciální případy získáme nahrazením $dF_X(x)$ výrazem $f(x) dx$ v absolutně spojitém případě a nahrazením integrálu váženou sumou s váhami p_n v diskrétním případě, podobně jako ve větě 6.1.

Z definice 6.4 vidíme, že střední hodnota je první moment. První centrální moment je vždy roven nule, neboť

$$\mathbb{E}(X - \mathbb{E}X) = \mathbb{E}X - \mathbb{E}(\mathbb{E}X) = \mathbb{E}X - \mathbb{E}X = 0.$$

Rozptyl náhodné veličiny je druhý centrální moment. V praktických aplikacích se ještě rutinně používají následující dvě charakteristiky založené na centrálních momentech třetího a čtvrtého řádu.

Definice 6.5

Označme $\mu = \mathbb{E}X$ střední hodnotu a $\sigma = \sqrt{\text{var } X}$ směrodatnou odchylku náhodné veličiny X . Potom

a) šikmost náhodné veličiny X je

$$\mathbb{E}\left(\frac{X - \mu}{\sigma}\right)^3 = \frac{\mathbb{E}(X - \mu)^3}{\sigma^3},$$

b) špičatost náhodné veličiny X je

$$\mathbb{E}\left(\frac{X - \mu}{\sigma}\right)^4 = \frac{\mathbb{E}(X - \mu)^4}{\sigma^4}.$$

Koeficient šikmosti popisuje asymetrii hustoty náhodné veličiny. Náhodné veličiny s hustotou symetrickou podle vertikální osy (procházející mediánem) mají šikmost nulovou. Náhodné veličiny, jejichž modus je vpravo od mediánu, mají šikmost zápornou a označujeme je zešikmené zleva. Jestliže je modus vlevo od mediánu, šikmost je kladná a náhodnou veličinu označujeme jako zešikmenou zprava.

Koeficient špičatosti charakterizuje koncentraci hustoty kolem mediánu. Špičatost normované náhodné veličiny, jejíž hustotou je Gaussova křivka (viz odstavec 8.5 na straně 159), má koeficient špičatosti roven hodnotě 3. Náhodné veličiny, jejichž pravděpodobnost je více soustředěna kolem mediánu (mají hustotu kolem mediánu větší než Gaussova křivka), mají hodnotu špičatosti větší než 3 a označujeme je jako *leptokurtické* (*špičaté*). Náhodné veličiny, jejichž pravděpodobnost je oproti Gaussově křivce rozložena více do chvostů, mají hodnotu špičatosti menší než 3 a nazýváme je *platykurtické* (*ploché*). Náhodné veličiny s koeficientem špičatosti 3 jsou označovány jako *mesokurtické*.

Někdy se koeficient špičatosti definuje s posunem -3 oproti výše uvedené definici (tj. platykurtické veličiny mají tento posunutý koeficient záporný). Anglicky se tato charakteristika nazývá *excess kurtosis*, v doslovném překladu „nadměrná špičatost“. Například Microsoft Excel pracuje s pojmem špičatosti ve smyslu této upravené definice.

Čebyševovu nerovnost lze zobecnit pro libovolný absolutní moment, což formalizujeme následujícím lemmatem. Lemma rovněž využijeme pro odložený důkaz Čebyševovy nerovnosti.

Lemma 6.6. *Nechť existuje n -tý moment náhodné veličiny X , $n > 0$. Pak pro libovolné $\varepsilon > 0$ platí*

$$\mathbb{P}[|X| \geq \varepsilon] \leq \frac{\mathbb{E}|X|^n}{\varepsilon^n}.$$

Důkaz.

$$\begin{aligned} \mathbb{E}|X|^n &= \int_{-\infty}^{\infty} |x|^n \, dF_X(x) \geq \int_{|x| \geq \varepsilon} |x|^n \, dF_X(x) \geq \\ &\geq \varepsilon^n \int_{|x| \geq \varepsilon} dF_X(x) = \varepsilon^n \mathbb{P}[|X| \geq \varepsilon]. \end{aligned}$$

□

Důkaz věty 6.5. V předchozím lemmatu položíme $n = 2$ a uvažujeme místo náhodné veličiny X náhodnou veličinu $X - \mathbb{E}X$. □

6.5 Úlohy

6.1 Škody na majetku v tisících Kč představují náhodnou veličinu X , která se řídí rozdělením s hustotou

$$f(x) = \begin{cases} \frac{c}{x^3} & \text{pro } x > 3, \\ 0 & \text{jinak.} \end{cases}$$

- Vypočítejte hodnotu c tak, aby šlo o hustotu pravděpodobnosti.
- Existuje střední hodnota a rozptyl? Pokud ano, vypočítejte je.
- Jaká je pravděpodobnost, že škody budou do 10 tisíc Kč?

6.2 Najděte takové diskrétní a spojitě rozdělení, kde není medián určen jednoznačně.

6.3 Najděte takové diskrétní a spojitě rozdělení, kde není modus určen jednoznačně.

6.4 Vyšetřete, pro které hodnoty μ má hustota

$$f(x) = \frac{1}{2\sqrt{2\pi}} \left(e^{-\frac{(x-\mu)^2}{2}} e^{-\frac{(x+\mu)^2}{2}} \right)$$

maximum.

6.5 Zkoušený přístroj je složen z pěti prvků. n -tý prvek se porouchá s pravděpodobností

$$p_n = 0,2 + 0,1(n - 1).$$

Určete střední hodnotu počtu porouchaných prvků.

6.6 Zkouší se n přístrojů. Pravděpodobnost poruchy je u všech přístrojů stejná a rovná se p . Určete střední hodnotu počtu přístrojů, které se během zkoušky porouchají.

6.7 V loterii je m_1 výher o hodnotě k_1 , m_2 výher o hodnotě $k_2, \dots, m_n$ výher o hodnotě k_n . Celkem je N losů. Určete cenu losu tak, aby střední hodnota výhry na jeden los byla rovna polovině jeho ceny.

6.8 V urně je m bílých a n černých koulí. Koule se vytahují tak dlouho, dokud není vytažena bílá koule. Koule se po vytažení vrací zpět. Určete střední hodnotu a rozptyl počtu vytažených koulí.

6.9 Určete střední hodnotu a rozptyl náhodné veličiny X s hustotou

$$f(x) = \frac{1}{\pi\sqrt{a^2 - x^2}}, \quad -a \leq x \leq a.$$

Řešení úloh

6.1 a) $c = 18$, b) $\mathbb{E}X = 6$, $\text{var } X$ neexistuje, c) $\mathbb{P}(X \leq 10) = 0,91$. Jde o tzv. Paretovo rozdělení typu I s parametrem tvaru $\alpha = 2$, používané ve financích právě pro modelování výskytu extrémních jevů.

6.2 Diskrétní: jakékoliv rozdělení se sudým počtem možných hodnot. Spojité: rozdělení s hustotou, která je symetrická kolem nuly a zároveň je hustota na intervalu $(-c; c)$ pro nějakou konstantu $c > 0$ nulová. Mediánem může být kterékoliv číslo z intervalu $(-c; c)$.

6.3 Diskrétní: rozdělení, které má pro dvě různé hodnoty stejnou pravděpodobnost, která je zároveň maximální hodnotou pravděpodobnostní funkce. Spojité: rozdělení s hustotou, která má dva vrcholy stejné hodnoty, která je zároveň maximální hodnotou hustoty.

6.4 Pro $\mu = 0$ je maximum pouze jedno.

6.5 Označme $Y = X_1 + \dots + X_5$ počet porouchaných prvků, kde X_n je náhodná veličina nabývající hodnoty 1, pokud se n -tý prvek porouchá, a 0 v ostatních případech.

$$EY = EX_1 + \dots + EX_5 = \sum_{n=1}^5 [0,2 + 0,1(n-1)] \cdot 1 = 2.$$

6.6 $\mathbb{E}X = np$

6.7

$$\frac{2}{N} (m_1 k_1 + \dots + m_n k_n)$$

6.8

$$\mathbb{E}X = \frac{m+n}{m}, \quad \text{var } X = \frac{1 - \frac{m}{m+n}}{\left(\frac{m}{n+m}\right)^2} = \frac{n}{m^2}.$$

Jedná se o tzv. geometrické rozdělení s parametrem úspěchu $p = \frac{m}{n+m}$.

6.9 $\mathbb{E}X = 0$, $\text{var } X = \frac{a^2}{2}$. Jedná se o tzv. rozdělení arcsin.

Kapitola 7

Vybraná rozdělení diskrétních náhodných veličin

7.1 Rovnoměrné diskrétní rozdělení

Nechť D je množina o n hodnotách ($n \in \mathbb{N}$). Náhodná veličina X , jejíž každá hodnota z množiny D nastane se stejnou pravděpodobností p , má rovnoměrné diskrétní rozdělení $\text{Rd}(n)$ s parametrem n , píšeme $X \sim \text{Rd}(n)$. Rovnoměrné diskrétní rozdělení je známé také jako „rozdělení se stejně pravděpodobnými výsledky“, jde tedy o model klasické pravděpodobnosti v pojmech náhodných veličin.

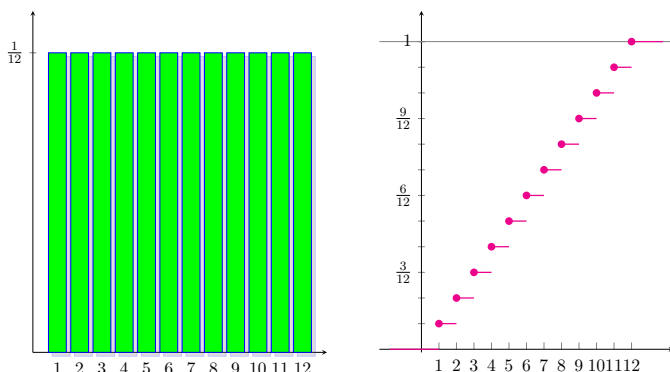
Pro další úvahy budeme pracovat s množinou výsledků speciálního tvaru¹ $D = \{1, 2, \dots, n\}$, pro kterou explicitně odvodíme základní vlastnosti náhodné veličiny X . Pravděpodobnostní a distribuční funkce mají pro tuto množinu tvar

$$p(x) = \begin{cases} \frac{1}{n} & \text{pro } x \in D, \\ 0 & \text{jinak.} \end{cases} \quad F(x) = \begin{cases} 0 & \text{pro } x < 1, \\ [x] \cdot \frac{1}{n} & \text{pro } 1 \leq x < n \\ 1 & \text{pro } x \geq n, \end{cases}$$

kde $[x]$ je zaokrouhlení čísla x na celé číslo směrem dolů. Střední hodnota a rozptyl jsou tvaru

$$\mathbb{E}X = \frac{n+1}{2} \quad \text{a} \quad \text{var } X = \frac{n^2-1}{12},$$

¹Pro jiné množiny D bychom mohli učinit totéž, jen budou formulace vzorců složitější.



Obrázek 7.1: Graf pravděpodobnostní a distribuční funkce rovnoměrného diskrétního rozdělení $Rd(12)$

(viz úlohu 7.1).

Příklad 7.1. Uvažujme hod kostkou ve tvaru pravidelného dvanáctistěnu, jehož každá ze stěn je označena postupně číslem od 1 do 12.

- Popište rozdělení pravděpodobnosti náhodné veličiny X , která udává hozené číslo (tj. číslo na svrchní stěně dvanáctistěnu po jeho dopadu), pomocí pravděpodobnostní a distribuční funkce.
- Sestavte grafy pravděpodobnostní a distribuční funkce.
- Vypočítejte střední hodnotu a rozptyl.
- Spočítejte pravděpodobnost, že
 - na horní stěně bude po dopadu sudé číslo (jev A),
 - na horní stěně bude po dopadu prvočíslo (jev B).

Řešení. Všechny možné výsledky hodu kostkou jsou reprezentovány množinou $D = \{1, 2, \dots, 12\}$ (tedy $n = 12$) a jsou všechny stejně pravděpodobné, s hodnotou pravděpodobnosti $\frac{1}{12}$.

- Pravděpodobnostní a distribuční funkce mají tvar

$$p(x) = \begin{cases} \frac{1}{12} & \text{pro } x \in D, \\ 0 & \text{jinak,} \end{cases} \quad F(x) = \begin{cases} 0 & \text{pro } x < 1, \\ \lfloor x \rfloor \cdot \frac{1}{12} & \text{pro } 1 \leq x < 12, \\ 1 & \text{pro } x \geq 12. \end{cases}$$

- Graf pravděpodobnostní a distribuční funkce ukazuje obrázek 7.1.
- $\mathbb{E}X = 6,5$, $\text{var } X = \frac{143}{12}$.

d) Příslušné pravděpodobnosti jsou

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(X = 2) + \mathbb{P}(X = 4) + \cdots + \mathbb{P}(X = 12) = 0,5, \\ \mathbb{P}(B) &= \mathbb{P}(X = 2) + \mathbb{P}(X = 3) + \mathbb{P}(X = 5) + \mathbb{P}(X = 7) + \mathbb{P}(X = 11) = \\ &= \frac{5}{12}.\end{aligned}$$

△

7.2 Degenerované rozdělení

Náhodná veličina X , která nabývá nějaké konstantní hodnoty c s pravděpodobností 1, tedy $\mathbb{P}(X = c) = 1$, a každou jinou hodnotu nabude s pravděpodobností 0, tedy $\mathbb{P}(X \neq c) = 0$, má degenerované rozdělení $\text{Dg}(c)$ s parametrem c , píšeme $X \sim \text{Dg}(c)$. Degenerované rozdělení se nazývá také Diracova² míra (centrovaná) v bodě c a značí se δ_c . Degenerované rozdělení má svůj význam především v tom, že umožňuje pracovat s nenáhodnou veličinou, u které je předem znám výsledek, jako by šlo o veličinu náhodnou. Degenerované rozdělení má například náhodná veličina vyjadřující počet puntíků padlých při házení s kostkou, která má na všech stěnách šest puntíků.

7.3 Alternativní rozdělení

V odstavci 3.3 jsme se zabývali speciálním typem náhodného pokusu, jež jsme nazvali Bernoulliho pokusem. Bernoulliho pokus nabývá pouze dvou možných výsledků, které obvykle označujeme jako „úspěch“ a „neúspěch“. Příkladem může být úspěch při hození mincí, úspěch (hození šestky) ve hře Člověče, nezlob se!, objevení vadného kusu výrobku při výstupní kontrole, pohlaví chlapec právě narozeného dítěte (nebo dívka, dle preferencí čtenáře-rodiče), vybavení náhodně vybrané domácnosti vysokorychlostní datovou přípojkou atd. V praxi se hodí tento pokus vyjádřit v termínech náhodných veličin pomocí tzv. *alternativního*, nazývaného také *Bernoulliho* nebo *nula-jedničkové* rozdělení³.



²Paul Dirac (1902–1984), britský matematik a fyzik.

³Mezinárodně je rozšířen název Bernoulliho rozdělení, který má přímou návaznost na Bernoulliho pokus. V Polsku a některých dalších zemích se užívá názvu nula-jedničkové rozdělení podle hodnot, kterých náhodná veličina nabývá. V naší učebnici jsme se přidrželi označení alternativní rozdělení, které je tradiční v českém, slovenském a německém prostředí.

Předpokládejme, že pravděpodobnost úspěchu v Bernoulliho pokusu je dána parametrem p , přičemž předpokládáme $0 < p < 1$ ⁴. Zavedme náhodnou veličinu X se dvěma hodnotami 0 a 1, reprezentující neúspěch, resp. úspěch v Bernoulliho pokusu takto:

$$X = \begin{cases} 0 & \text{jestliže nastal neúspěch,} \\ 1 & \text{jestliže nastal úspěch v Bernoulliho pokusu.} \end{cases}$$

Snadno odvodíme, že pravděpodobnostní a distribuční funkce tohoto rozdělení mají tvar

$$p(x) = \begin{cases} 1-p & \text{pro } x=0, \\ p & \text{pro } x=1, \\ 0 & \text{jinak,} \end{cases} \quad F(x) = \begin{cases} 0 & \text{pro } x < 0, \\ 1-p & \text{pro } 0 \leq x < 1, \\ 1 & \text{pro } x \geq 1. \end{cases}$$

V úloze 7.3 čtenář sám dokáže, že střední hodnota a rozptyl jsou tvaru

$$\mathbb{E}X = p \quad \text{a} \quad \text{var } X = p(1-p).$$

Alternativní rozdělení budeme označovat $A(p)$ ⁵.

Příklad 7.2. V osudí je jedna bílá a dvě černé koule. Najděte pravděpodobnostní a distribuční funkci náhodné veličiny X , která hodnotou 1 udává, zda jsme při jednorázovém tažení právě jedné koule z osudí vytáhli bílou kouli, a hodnotou 0, že jsme vytáhli kouli černou. Jaká je pravděpodobnost, že vytáhneme bílou kouli? Vypočítejte střední hodnotu a rozptyl.

Řešení. $X \sim A(p)$, kde $p = \frac{1}{3}$ je pravděpodobnost vytažení bílé koule. Dále tedy

$$p(x) = \begin{cases} \frac{2}{3} & \text{pro } x=0, \\ \frac{1}{3} & \text{pro } x=1, \\ 0 & \text{jinak,} \end{cases} \quad F(x) = \begin{cases} 0 & \text{pro } x < 0, \\ \frac{2}{3} & \text{pro } 0 \leq x < 1, \\ 1 & \text{pro } x \geq 1. \end{cases}$$

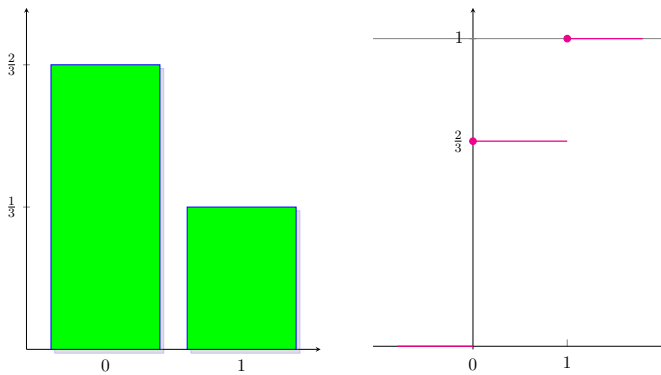
Grafy $p(x)$ a $F(x)$ pro $X \sim A(\frac{1}{3})$ jsou na obrázku 7.2. Střední hodnota a rozptyl jsou

$$\mathbb{E}X = \frac{1}{3}, \quad \text{var } X = \frac{1}{3} \cdot \left(1 - \frac{1}{3}\right) = \frac{2}{9}.$$

△

⁴Pokud by bylo $p=0$ nebo $p=1$, jednalo by se o degenerované rozdělení.

⁵Někdy se označuje také $\text{Alt}(p)$, v anglicky psaných publikacích pak $B(p)$.



Obrázek 7.2: Graf pravděpodobnostní a distribuční funkce alternativního rozdělení $A(1/3)$

7.4 Binomické rozdělení

Mějme posloupnost n nezávislých Bernoulliho pokusů s konstantní pravděpodobností úspěchu $p \in (0; 1)$. Ty můžeme reprezentovat pomocí posloupnosti náhodných veličin $X_1, \dots, X_n$ s identickým alternativním rozdělením $X_i \sim A(p)$, $i = 1, \dots, n$. Rozdělení náhodné veličiny X představující počet úspěchů v této posloupnosti nezávislých Bernoulliho pokusů se nazývá *binomické rozdělení*. Z podstaty této definice vyplývá, že náhodná veličina X může být reprezentována jako součet

$$X = \sum_{i=1}^n X_i \quad (7.1)$$

a může nabývat hodnot z množiny $\{0, 1, \dots, n\}$. Binomické rozdělení budeme značit $\text{Bi}(n; p)$, kde n je počet Bernoulliho pokusů a p pravděpodobnost úspěchu.

Pravděpodobnostní funkci binomického rozdělení jsme již odvodili v příkladu 3.8 jako vzorec (3.10) a distribuční funkce je kumulativním součtem těchto pravděpodobností. Pravděpodobnost, že v řadě n Bernoulliho pokusů nastane právě k úspěchů, je tedy

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{pro } k = 0, 1, \dots, n. \quad (7.2)$$

Distribuční funkce $F(x)$ je tvaru

$$F(x) = \begin{cases} 0 & \text{pro } x < 0, \\ \sum_{k=0}^{\lfloor x \rfloor} \binom{n}{k} p^k (1-p)^{n-k} & \text{pro } 0 \leq x < n, \\ 1 & \text{pro } x \geq n. \end{cases}$$

Pro výpočet hodnot pravděpodobnostní funkce, resp. distribuční funkce můžeme využít funkce Excelu `=BINOM.DIST`, jak je ukázáno v příkladu 7.3. Poslední parametr této funkce udává typ výpočtu: použijeme hodnotu NEPRAVDA pro výpočet hodnot pravděpodobnostní funkce, a PRAVDA pro hodnoty funkce distribuční. První tři parametry funkce jsou postupně hodnoty parametrů k , n a p .

Příklad 7.3. Sestavte graf pravděpodobnostní funkce náhodné veličiny X s binomickým rozdělením

- $\text{Bi}(16; 0,1)$,
- $\text{Bi}(16; 0,5)$,
- $\text{Bi}(16; 0,8)$

a k nim grafy distribučních funkcí. Pro případ c) určete pravděpodobnost, že náhodná veličina X nabude hodnoty

- právě 12,
- nejvýše 12,
- méně než 12,
- alespoň 12,
- více než 12,
- alespoň 12 a nejvýše 15,
- více než 12 a méně než 15?

Řešení. Grafy pravděpodobnostních a distribučních funkcí jsou na obrázku 7.3. Výpočty pravděpodobností následují.

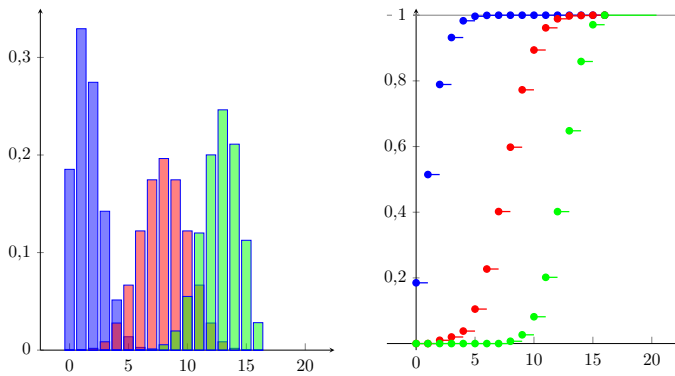
$$1) \mathbb{P}(X = 12) = p(12) = \binom{16}{12} 0,8^{12} 0,2^4 \doteq 0,2,$$

$$= \text{BINOM.DIST}(12; 16; 0,8; \text{NEPRAVDA})$$

$$2) \mathbb{P}(X \leq 12) = p(0) + \dots + p(12) = F(12) =$$

$$= \binom{16}{0} 0,8^0 0,2^{16} + \dots + \binom{16}{12} 0,8^{12} 0,2^4 \doteq 0,402,$$

$$= \text{BINOM.DIST}(12; 16; 0,8; \text{PRAVDA})$$



Obrázek 7.3: Graf pravděpodobnostních a distribučních funkcí binomického rozdělení $\text{Bi}(16; 0,1)$ (modrá barva), $\text{Bi}(16; 0,5)$ (červená barva), $\text{Bi}(16; 0,8)$ (zelená barva)

$$\begin{aligned}
 3) \quad \mathbb{P}(X < 12) &= p(0) + \dots + p(11) = F(11) = \\
 &= \binom{16}{0} 0,8^0 0,2^{16} + \dots + \binom{16}{11} 0,8^{11} 0,2^5 \doteq 0,202, \\
 &= \text{BINOM.DIST}(11; 16; 0,8; \text{PRAVDA})
 \end{aligned}$$

$$\begin{aligned}
 4) \quad \mathbb{P}(X \geq 12) &= p(12) + \dots + p(16) = 1 - F(11) = \\
 &= 1 - \left[\binom{16}{0} 0,8^0 0,2^{16} + \dots + \binom{16}{11} 0,8^{11} 0,2^5 \right] \doteq 0,798, \\
 &= 1 - \text{BINOM.DIST}(11; 16; 0,8; \text{PRAVDA})
 \end{aligned}$$

$$\begin{aligned}
 5) \quad \mathbb{P}(X > 12) &= p(13) + \dots + p(16) = 1 - F(12) = \\
 &= 1 - \left[\binom{16}{0} 0,8^0 0,2^{16} + \dots + \binom{16}{12} 0,8^{12} 0,2^4 \right] \doteq 0,598, \\
 &= 1 - \text{BINOM.DIST}(12; 16; 0,8; \text{PRAVDA})
 \end{aligned}$$

$$\begin{aligned}
 6) \quad \mathbb{P}(12 \leq X \leq 15) &= p(12) + \dots + p(15) = F(15) - F(11) = \\
 &= \binom{16}{12} 0,8^{12} 0,2^4 + \dots + \binom{16}{15} 0,8^{15} 0,2^1 \doteq 0,770, \\
 &= \text{BINOM.DIST}(15; 16; 0,8; \text{PRAVDA}) - \text{BINOM.DIST}(11; 16; 0,8; \text{PRAVDA})
 \end{aligned}$$

$$\begin{aligned}
 7) \quad \mathbb{P}(12 < X < 15) &= p(13) + p(14) = F(14) - F(12) = \\
 &= \binom{16}{13} 0,8^{13} 0,2^3 + \binom{16}{14} 0,8^{14} 0,2^2 \doteq 0,457, \\
 &= \text{BINOM.DIST}(14; 16; 0,8; \text{PRAVDA}) - \text{BINOM.DIST}(12; 16; 0,8; \text{PRAVDA})
 \end{aligned}$$

nebo také jako součet

```
= BINOM.DIST(14;16;0,8;NEPRAVDA) a
= BINOM.DIST(13;16;0,8;NEPRAVDA).
```

△

Název binomické rozdělení vychází ze skutečnosti, že pravděpodobnosti $p_k = \mathbb{P}(X = k)$ jsou členy binomického rozvoje výrazu

$$1^n = [p + (1 - p)]^n = \sum_{k=0}^n \binom{n}{k} p^k (1 - p)^{n-k} = \sum_{k=0}^n p_k. \quad (7.3)$$

Ze vzorce (7.3) je i okamžitě vidět, že pravděpodobnosti p_k splňují podmínky pro pravděpodobnostní rozdělení, totiž

a) nezápornost $p_k \geq 0$ pro libovolné k a

b) normovanost $\sum_{k=0}^n p_k = 1$.

Náhodná veličina X s rozdělením $\text{Bi}(n; p)$ má střední hodnotu np a rozptyl $np(1 - p)$. Můžeme to odvodit přímo z definic:

$$\begin{aligned} \mathbb{E}X &= \sum_{k=0}^n k p_k = \sum_{k=0}^n k \binom{n}{k} p^k (1 - p)^{n-k} = \\ &= np \sum_{k=1}^n \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} (1 - p)^{n-k} = \\ &= np \sum_{j=0}^{n-1} \binom{n-1}{j} p^j (1 - p)^{n-1-j} = \\ &= np(p + (1 - p))^{n-1} = np. \end{aligned}$$

Pro výpočet rozptylu uijeme výpočetního vzorce (6.7), odkud

$$\text{var } X = \mathbb{E}X(X - 1) + \mathbb{E}X - (\mathbb{E}X)^2.$$

Z definic dopočteme potřebnou hodnotu

$$\begin{aligned} \mathbb{E}X(X - 1) &= \sum_{k=0}^n k(k-1) \binom{n}{k} p^k (1 - p)^{n-k} = \\ &= n(n-1)p^2 \sum_{k=2}^n \frac{(n-2)!}{(k-2)!(n-k)!} p^{k-2} (1 - p)^{n-k} = \\ &= n(n-1)p^2 \sum_{j=0}^{n-2} \binom{n-2}{j} p^j (1 - p)^{n-2-j} = \\ &= n(n-1)p^2. \end{aligned}$$

Dosazením konečně dostáváme

$$\text{var } X = n(n-1)p^2 + np - (np)^2 = np(1-p).$$

Při odvození střední hodnoty lze rovněž využít vlastnosti d) věty 6.2 o součtu středních hodnot a faktu, že náhodná veličina X je dle (7.1) takovým vhodným součtem. Pro rozptyl platí obdobná vlastnost (věta 9.2), ta je však založena na nezávislosti náhodných veličin, o které budeme hovořit až v odstavci 9.4.

Příklad 7.4. Víme-li, že pravděpodobnost narození chlapce je 0,515, jaká je pravděpodobnost, že mezi čtyřmi po sobě narozenými dětmi budou

- první dva chlapci, další dvě dívky,
- právě dva chlapci?
- Zjistěte, kolik se musí narodit dětí, aby pravděpodobnost, že mezi nimi bude alespoň jeden chlapec, byla větší nebo rovna 0,99.

Pro řešení všech úloh připustíme zjednodušující předpoklad, že pohlaví po sobě narozených dětí jsou nezávislé jevy⁶.

Řešení.

- Jelikož je pořadí narození chlapců a dívek určeno pevně, je hledaná pravděpodobnost

$$0,515^2 \cdot 0,485^2 \doteq 0,062.$$

- V této situaci na pořadí narození nezáleží, musíme uvažovat všechny možnosti, kterých je $\binom{4}{2}$. Výsledná pravděpodobnost je

$$p_2 = \binom{4}{2} \cdot 0,515^2 \cdot 0,485^2 \doteq 0,374.$$

Tuto pravděpodobnost můžeme nalézt také zavedením náhodné veličiny X reprezentující počet narozených chlapců; řídí se binomickým rozdělením $\text{Bi}(4; 0,515)$. Pravděpodobnost $p_2 = \mathbb{P}(X = 2) \doteq 0,374$ vypočítáme v MS Excel.

= BINOM.DIST(2;4;0,515;NEPRAVDA)

- Opačným jevem k požadovanému jevu „narozen alespoň jeden chlapec“ je „nenarozen žádný chlapec“ neboli „narozeno n dívek“, jehož pravděpodobnost je

$$\mathbb{P}(\text{narozeno } n \text{ dívek}) = (1 - 0,515)^n = 0,485^n.$$

⁶To v případě narození dvojčat obvykle neplatí.

Počet narozených dětí tedy musí dle požadavků zadání úlohy splňovat nerovnost

$$1 - (0,485)^n \geq 0,99,$$

odkud již snadno odvodíme

$$n \geq \log_{0,485} 0,01 = \frac{\log 0,01}{\log 0,485} \doteq 6,47.$$

Potřebujeme tedy alespoň sedm narozených dětí.

△

Příklad 7.5. Víme, že pravděpodobnost vypěstování zdravé sazenice ze semena je 0,62. Za náhodnou veličinu X budeme považovat počet zdravých sazenic vypěstovaných z 27 semen. Určete,

- jaký je nejpravděpodobnější počet zdravých rostlin a jaká je jeho pravděpodobnost,
- střední hodnotu a rozptyl náhodné veličiny X .

Řešení. Předpokládejme, že semena klíčí a rostliny rostou nezávisle na sobě. Pak je ze zadání příkladu zřejmé, že náhodná veličina X má $\text{Bi}(27; 0,62)$.

- Máme nalézt nejpravděpodobnější hodnotu náhodné veličiny X , tzn. modus $\hat{X}$. Z definice $\hat{X}$ plyne, že

$$\frac{\mathbb{P}(X = \hat{X} - 1)}{\mathbb{P}(X = \hat{X})} \leq 1 \quad \text{a} \quad \frac{\mathbb{P}(X = \hat{X} + 1)}{\mathbb{P}(X = \hat{X})} \leq 1.$$

Protože X má $\text{Bi}(27; 0,62)$, dostáváme odtud, že

$$\begin{aligned} \frac{\binom{27}{\hat{X}-1} 0,62^{\hat{X}-1} (1-0,62)^{27-\hat{X}+1}}{\binom{27}{\hat{X}} 0,62^{\hat{X}} (1-0,62)^{27-\hat{X}}} &\leq 1, \\ \frac{\binom{27}{\hat{X}+1} 0,62^{\hat{X}+1} (1-0,62)^{27-\hat{X}-1}}{\binom{27}{\hat{X}} 0,62^{\hat{X}} (1-0,62)^{27-\hat{X}}} &\leq 1, \end{aligned}$$

neboli

$$\frac{\hat{X}}{27 - \hat{X} + 1} \cdot \frac{1 - 0,62}{0,62} \leq 1 \quad \text{a} \quad \frac{27 - \hat{X}}{\hat{X} + 1} \cdot \frac{0,62}{1 - 0,62} \leq 1.$$

Z obou nerovnic vyjádříme $\hat{X}$ a dostáváme

$$16,36 \leq \hat{X} \leq 17,36.$$

Nejpravděpodobnější počet sazenic, které vypěstujeme z 27 semen, je 17. Alternativní možností nalezení modu rozdělení je spočítat si hodnoty pravděpodobnostní funkce v MS Excel pomocí `= BINOM.DIST` pro všechny argumenty $k = 0, 1, \dots, 27$ a nalézt mezi nimi tu s nejvyšší hodnotou $\mathbb{P}(X = k)$. Tou je pravděpodobnost, že vypěstujeme právě 17 sazenic,

$$\mathbb{P}(X = 17) = \binom{27}{17} 0,62^{17} 0,38^{10} \doteq 0,156\,5.$$

`= BINOM.DIST(17;27;0,62;NEPRAVDA)`

- b) Dle vzorců pro výpočet střední hodnoty a rozptylu binomického rozdělení dostáváme

$$\begin{aligned}\mathbb{E}X &= 27 \cdot 0,62 = 16,74, \\ \text{var } X &= 27 \cdot 0,62 \cdot 0,38 = 6,361\,2.\end{aligned}$$

△

7.5 Geometrické rozdělení

Náhodná veličina X má asymetrické geometrické rozdělení $\text{Ge}(p)$, píšeme $X \sim \text{Ge}(p)$, jestliže X označuje počet neúspěšných pokusů před prvním úspěšným pokusem v Bernoulliho procesu. Nabývá hodnot $r = 0, 1, 2, \dots$ s pravděpodobnostmi

$$P(X = r) = p_r = p(1 - p)^r,$$

kde parametr $p \in (0; 1)$ je pravděpodobnost úspěchu v jednom Bernoulliho pokusu a je stejná pro všechny nezávislé pokusy. Všimněte si, že tyto pravděpodobnosti tvoří geometrickou posloupnost (geometricky klesají s rostoucí hodnotou r), odkud vychází i název tohoto rozdělení. Je zřejmé, že všechna

$p_r \geq 0$ a $\sum_{r=0}^{\infty} p_r = 1$, neboť

$$\sum_{r=0}^{\infty} p_r = \sum_{r=0}^{\infty} p(1 - p)^r = p \sum_{r=0}^{\infty} (1 - p)^r = p \frac{1}{1 - (1 - p)} = 1.$$

Pravděpodobnostní funkce geometrického rozdělení má tedy tvar

$$p(x) = \begin{cases} p(1 - p)^x & \text{pro } x = 0, 1, \dots, \\ 0 & \text{jinak.} \end{cases}$$

Distribuční funkce geometrického rozdělení je tvaru

$$F(x) = \begin{cases} 0 & \text{pro } x < 0, \\ \sum_{r=0}^{\lfloor x \rfloor} p(1-p)^r & \text{pro } x \geq 0. \end{cases}$$

Příklad 7.6. Náhodná veličina X udává počet úspěchů v sérii nezávislých pokusů. Pravděpodobnost úspěchu je 0,2. Vypočítejte pravděpodobnost, se kterou před prvním úspěchem nastanou

- právě čtyři neúspěchy,
- nejvýše čtyři neúspěchy,
- alespoň čtyři neúspěchy,
- více než dva a nejvýše čtyři neúspěchů.

Řešení.

- $\mathbb{P}(X = 4) = p(4) = 0,2(1 - 0,2)^4 = 0,081\,92.$
- $\mathbb{P}(X \leq 4) = p(0) + \dots + p(4) = F(4) =$
 $= 0,2(1 - 0,2)^0 + \dots + 0,2(1 - 0,2)^4 = 0,672\,32.$
- $\mathbb{P}(X \geq 4) = 1 - \mathbb{P}(X < 4) = 1 - \mathbb{P}(X \leq 3) = 1 - F(3) =$
 $= 1 - [0,2(1 - 0,2)^0 + \dots + 0,2(1 - 0,2)^3] = 0,409\,6.$
- $\mathbb{P}(2 < X \leq 4) = p(3) + p(4) = F(4) - F(2) =$
 $= 0,2(1 - 0,2)^3 + 0,2(1 - 0,2)^4 = 0,184\,32.$

△

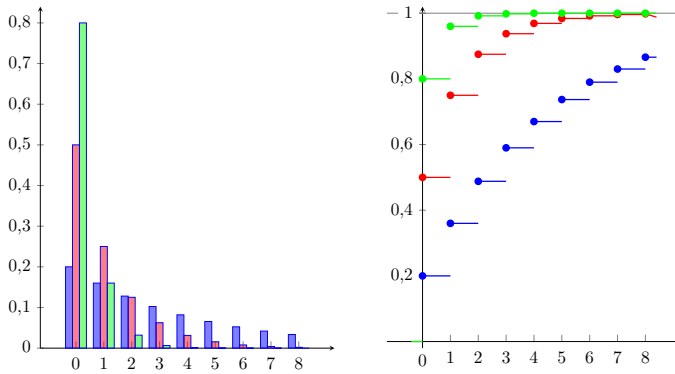
Příklad 7.7. Sestrojte graf pravděpodobnostní a distribuční funkce náhodné veličiny X s geometrickým rozdělením

- $\text{Ge}(0,2),$
- $\text{Ge}(0,5),$
- $\text{Ge}(0,8).$

Řešení. Grafy pravděpodobnostních a distribučních funkcí pro uvedená rozdělení ukazuje obrázek 7.4. △

Před výpočtem střední hodnoty a rozptylu si připomeneme vzorce pro součet řad

$$\sum_{r=1}^{\infty} r q^{r-1} = \frac{1}{(1-q)^2} \quad \text{a} \quad \sum_{r=2}^{\infty} r(r-1) q^{r-2} = \frac{2}{(1-q)^3},$$



Obrázek 7.4: Graf pravděpodobnostních a distribučních funkcí geometrického rozdělení $\text{Ge}(0,2)$ (modrá barva), $\text{Ge}(0,5)$ (červená barva), $\text{Ge}(0,8)$ (zelená barva)

které se odvodí derivováním podle q vzorce pro součet geometrické řady

$$\sum_{r=0}^{\infty} q^r = \frac{1}{1-q}.$$

Využitím těchto vzorců dostáváme

$$\begin{aligned} \mathbb{E}X &= \sum_{r=0}^{\infty} rp(1-p)^r = p(1-p) \sum_{r=0}^{\infty} r(1-p)^{r-1} = \\ &= p(1-p) \frac{1}{(1-(1-p))^2} = \frac{1-p}{p}. \end{aligned}$$

$$\begin{aligned} \mathbb{E}X(X-1) &= \sum_{r=2}^{\infty} r(r-1)p(1-p)^r = \\ &= p(1-p)^2 \sum_{r=2}^{\infty} r(r-1)(1-p)^{r-2} = \\ &= p(1-p)^2 \cdot \frac{2}{(1-(1-p))^3} = \frac{2(1-p)^2}{p^2}. \end{aligned}$$

Odtud plyne, že

$$\begin{aligned} \text{var } X &= \mathbb{E}X(X-1) + \mathbb{E}X - (\mathbb{E}X)^2 = \\ &= \frac{2(1-p)^2}{p^2} + \frac{1-p}{p} - \frac{(1-p)^2}{p^2} = \frac{1-p}{p^2}. \end{aligned}$$

Geometrické rozdělení se uplatňuje jako jednoduchý pravděpodobnostní model například pro diskrétní délku „života“ (dobu trvanlivosti, bezporuchovosti) výrobku ve dnech, měsících či letech, než se rozbije, dobu čekání na určitou událost, například než zákazník nahlásí pojišťovně škodní událost, nebo dobu pracovní neschopnosti, než dotyčný bude opět schopen práce.

V některých publikacích se jako geometrické rozdělení označují i rozdělení příbuzných náhodných veličin, konkrétně

- počet pokusů $X' = X + 1$ potřebných k prvnímu úspěchu s pravděpodobnostní funkcí

$$\mathbb{P}(X' = n) = \mathbb{P}(X = n - 1) = p(1 - p)^{n-1}, \text{ nebo}$$

(označováno též jako *posunuté geometrické rozdělení*),

- počet úspěchů X'' před prvním dosaženým neúspěchem s pravděpodobnostní funkcí

$$\mathbb{P}(X'' = k) = p^k(1 - p)$$

(význam parametrů je otočen).

V tomto ohledu není názvosloví ani označení ustálené, vždy je třeba se ujistit, která z parametrizací byla pro tu kterou náhodnou veličinu autorem textu zvolena.

Příklad 7.8. S jakou pravděpodobností padne při hodu kostkou šestka

- až v osmém hodu,
- nejpozději ve třetím hodu,
- v jiných než ve druhém a třetím hodu?

Řešení. Nechť náhodná veličina X udává počet hodů, kdy šestka nepadne. Pravděpodobnost úspěchu, tj. že v jednom hodu padne šestka, je $p = \frac{1}{6}$.

- Jev „šestka padne až v osmém hodu“ znamená, že úspěšnému osmému hodu předchází sedm hodů neúspěšných, tedy

$$\mathbb{P}(X = 7) = \frac{1}{6} \left(1 - \frac{1}{6}\right)^7 \doteq 0,0465.$$

- Jev „šestka padne nejpozději ve třetím hodu“ znamená, že padne hned v prvním hodu (tedy nepředchází ji žádný neúspěch), nebo ve druhém hodu (předchází ji jeden hod neúspěšný), nebo šestka padne při třetím

hodu (předchází ji dva neúspěšné pokusy). Proto

$$\begin{aligned}\mathbb{P}(X \leq 2) &= p(0) + p(1) + p(2) = F(2) = \sum_{r=0}^2 p(1-p)^r = \\ &= \frac{1}{6} + \frac{1}{6} \cdot \frac{5}{6} + \frac{1}{6} \left(\frac{5}{6}\right)^2 \doteq 0,421.\end{aligned}$$

$$\text{c)} \quad \mathbb{P}(X \neq 2; X \neq 3) = 1 - p(1) - p(2) \doteq 0,7454.$$

△

Příklad 7.9. Předpokládejme, že dravá ryba přežije, pokud uloví rybu alespoň jednou za dva dny. Během dvou dnů podnikne osm zápasů s pravděpodobností ulovení $p = 0,25$. Jaká je pravděpodobnost, že dravá ryba nezemře?

Řešení. Dravá ryba přežije, jestliže poprvé zvítězí během osmi zápasů. Hledaná pravděpodobnost je tedy součet

$$\sum_{n=1}^8 p_{n-1},$$

kde p_{n-1} je pravděpodobnost, že dravá ryba poprvé zvítězí v n -tém zápase ($n-1$ je tedy počet předchozích neúspěchů). Tyto pravděpodobnosti se řídí geometrickým rozdělením s parametrem p :

$$\begin{aligned}\sum_{n=1}^8 p_{n-1} &= \sum_{r=0}^7 p_r = \sum_{r=0}^7 p(1-p)^r = \sum_{r=0}^7 0,25 \cdot 0,75^r = \\ &= \frac{1}{4} + \frac{3}{16} + \frac{9}{64} + \frac{27}{256} + \frac{81}{1024} + \frac{243}{4096} + \frac{729}{16384} + \frac{2187}{65536} \doteq 0,90.\end{aligned}$$

Tedy dravá ryba má 10% šanci, že zemře, než se jí podaří kořist ulovit.

Porovnejme pro zajímavost ještě střední hodnotu $\mathbb{E}X$ počtu neúspěšných soubojů $X \sim \text{Ge}(0,25)$ s nejpravděpodobnějším počtem neúspěšných soubojů, neboli modem $\hat{X}$:

$$\mathbb{E}X = \frac{1-p}{p} = \frac{0,75}{0,25} = 3, \quad \hat{X} = 0,$$

protože $p_0 = \frac{1}{4} > p_k = \frac{1}{4} \cdot \left(\frac{3}{4}\right)^r$ pro libovolné $r = 1, 2, \dots$. Vidíme, že střední hodnota geometrické náhodné veličiny může být značně vzdálená od nejpravděpodobnější hodnoty (tím více, čím je pravděpodobnost úspěchu menší). △

7.6 Negativně-binomické rozdělení

Náhodná veličina X s negativně binomickým rozdělením udává počet $r \in \mathbb{N}$ neúspěšných pokusů, které předcházejí k -tému úspěšnému pokusu ($k \in \mathbb{N}$) v posloupnosti celkem $n = k + r$ nezávislých Bernoulliho pokusů, ve kterých úspěch nastává s pravděpodobností $p \in (0; 1)$. Toto rozdělení označujeme $\text{NB}(k; p)$ nebo $\text{NeBi}(k; p)$. Občas se nazývá také Pascalovo či Pólyovo⁷ rozdělení.

Pravděpodobnostní funkce negativně binomického rozdělení má při výše uvedeném parametrizaci tvar

$$p(r) = \mathbb{P}(X = r) = \begin{cases} \binom{k+r-1}{r} p^k (1-p)^r & \text{pro } r = 0, 1, \dots, \\ 0 & \text{jinak,} \end{cases}$$

kde p je pravděpodobnost úspěchu. Distribuční funkce má tvar

$$F(x) = \begin{cases} 0 & \text{pro } x < 0, \\ \sum_{r=0}^{\lfloor x \rfloor} \binom{k+r-1}{r} p^k (1-p)^r & \text{pro } x \geq 0. \end{cases}$$

Vztahy pro střední hodnotu a rozptyl mají tvar

$$\mathbb{E}X = \frac{k(1-p)}{p}, \quad \text{var } X = \frac{k(1-p)}{p^2}.$$

Negativně-binomické rozdělení je speciálně pro $k = 1$ geometrickým rozdělením (viz odstavec 7.5). Pro výpočet geometrických pravděpodobností v příkladech 7.8 a 7.9 je tedy kromě přímého výpočtu možno využít funkci `= NEGBINOM.DIST` určenou pro výpočet pravděpodobnostní a distribuční funkce negativně binomického rozdělení. Například pravděpodobnost v příkladu 7.9 dostaneme jednodušeji jako výsledek funkce

`= NEGBINOM.DIST(7; 1; 0,25; PRAVDA)`.

Příklad 7.10. Nechť náhodná veličina $X \sim \text{NB}(2; 0,8)$ vyjadřuje počet neúspěchů před 2. úspěchem v sérii nezávislých Bernoulliho pokusů. Vypočítejte pravděpodobnost, že

- nenastane žádný neúspěch,
- nastanou čtyři neúspěchy,
- nastanou nejvýše čtyři neúspěchy,
- nastane více než pět a nejvýše osm neúspěchů.

⁷György Pólya (1887–1985), maďarský matematik.

Řešení. V nezávislých pokusech je pravděpodobnost úspěchu 0,8.

a) $X = 0$ znamená, že před 2. úspěchem nenastane žádný neúspěch. Tedy

$$\mathbb{P}(X = 0) = p(0) = F(0) = \binom{0+2-1}{0} 0,8^2(1-0,8)^0 = 0,8^2 = 0,64,$$

$$= \text{NEGBINOM.DIST}(0;2;0,8;\text{NEPRAVDA})$$

$$\text{b) } \mathbb{P}(X = 4) = p(4) = \binom{4+2-1}{4} 0,8^2(1-0,8)^4 = 0,005\,12,$$

$$= \text{NEGBINOM.DIST}(4;2;0,8;\text{NEPRAVDA})$$

$$\text{c) } \mathbb{P}(X \leq 4) = p(0) + \dots + p(4) = F(4) =$$

$$= \binom{0+2-1}{0} 0,8^2(1-0,8)^0 + \dots + \binom{4+2-1}{4} 0,8^2(1-0,8)^4 =$$

$$= 0,998\,4,$$

$$= \text{NEGBINOM.DIST}(4;2;0,8;\text{PRAVDA})$$

$$\text{d) } \mathbb{P}(5 < X \leq 8) = p(6) + p(7) + p(8) = F(8) - F(5) =$$

$$= \binom{6+2-1}{6} 0,8^2(1-0,8)^6 + \binom{7+2-1}{7} 0,8^2(1-0,8)^7 +$$

$$+ \binom{8+2-1}{8} 0,8^2(1-0,8)^8 \doteq 0,000\,367.$$

$$= \text{NEGBINOM.DIST}(8;2;0,8;\text{PRAVDA}) - \text{NEGBINOM.DIST}(5;2;0,8;\text{PRAVDA})$$

△

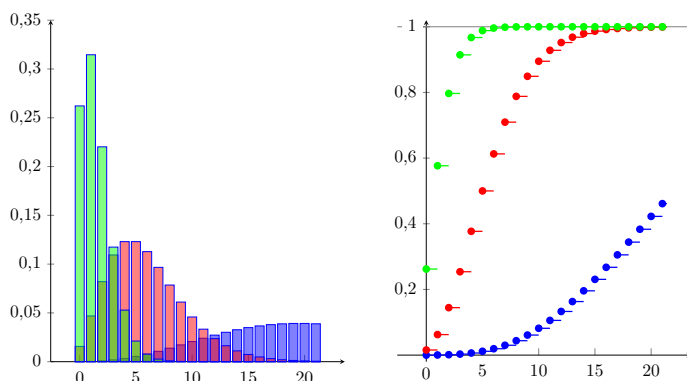
Příklad 7.11. Sestrojte grafy pravděpodobnostních funkcí a distribučních funkcí náhodné veličiny X s negativně-binomickým rozdělením NB(6; p) pro

- a) $p = 0,2$,
- b) $p = 0,5$,
- c) $p = 0,8$.

Řešení. Grafy pravděpodobnostních a distribučních funkcí jsou na obrázku 7.5. △

Stejnou informaci jako negativně binomicky rozdělená náhodná veličina X (vyjadřující počet neúspěchů před k -tým úspěchem) nese náhodná veličina W reprezentující dobu čekání na k -tý úspěch, tj. vyjadřující počet pokusů potřebných k dosažení k -tého úspěchu. V tomto případě

$$\mathbb{P}(W = n) = \mathbb{P}(X = n - k) = \binom{n-1}{k-1} p^k (1-p)^{n-k}.$$



Obrázek 7.5: Graf pravděpodobnostních a distribučních funkcí negativně binomického rozdělení $\text{NB}(6; 0,2)$ (modrá barva), $\text{NB}(6; 0,5)$ (červená barva), $\text{NB}(6; 0,8)$ (zelená barva)

Případně lze situaci charakterizovat pomocí náhodné veličiny Y reprezentující počet úspěchů před r -tým neúspěchem, pak

$$\mathbb{P}(Y = k) = \binom{k+r-1}{k} p^k (1-p)^r.$$

Rozdělení všech tří náhodných veličin X , W , Y bývají v literatuře označována shodně za negativně binomická rozdělení, i když se matematicky jedná o odlišná rozdělení.

Příklad 7.12. Jaká je pravděpodobnost, že „pro hosení dvou šestek při hodu kostkou budeme muset házení opakovat“:

- právě 10krát,
- více než 20krát,
- alespoň 12krát a nejvýše 17krát,
- 100 a vícekrát?

Řešení. Předpokládejme, že kostka je spravedlivá, tedy všechny výsledky hodu kostkou nastanou se stejnou pravděpodobností $p = \frac{1}{6}$. Za náhodný pokus budeme považovat hosení jedné šestky s dvěma možnými výsledky: „padne šestka“ (úspěch), „nepadne šestka“ (neúspěch). Nechť W udává počet hodů, které musíme provést, abychom hodili šestku dvakrát, tj. $k = 2$. Pro $2 \leq n < \infty$ tedy platí

$$\mathbb{P}(W = n) = \binom{n-1}{2-1} \left(\frac{1}{6}\right)^2 \left(1 - \frac{1}{6}\right)^{n-2} = \binom{n-1}{1} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^{n-2}.$$

$$\text{a) } \mathbb{P}(W = 10) = \binom{10-1}{1} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^{10-2} = \binom{9}{1} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^8 \doteq 0,058.$$

Pro výpočet pravděpodobnosti pomocí Excelu musíme náhodnou veličinu reparametrizovat jako počet neúspěchů před 2. úspěchem, tedy jako $X = W - 2$ s rozdělením $X \sim \text{NB}(2; \frac{1}{6})$. V tom případě $\mathbb{P}(W = 10) = \mathbb{P}(X = 8)$, a tedy můžeme nechat vypočítat

$$= \text{NEGBINOM.DIST}(8; 2; 1/6; \text{NEPRAVDA})$$

$$\text{b) } \mathbb{P}(W > 20) = p(21) + p(22) + p(23) + \dots = 1 - \mathbb{P}(W \leq 20) =$$

$$= 1 - \sum_{n=2}^{20} \binom{n-1}{1} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^{n-2} \doteq 0,13042,$$

$$= 1 - \text{NEGBINOM.DIST}(18; 2; 1/6; \text{PRAVDA})$$

$$\text{c) } \mathbb{P}(12 \leq W \leq 17) = p(12) + p(13) + p(14) + p(15) + p(16) + p(17) =$$

$$= \sum_{n=12}^{17} \binom{n-1}{2-1} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^{n-2} \doteq 0,232,$$

$$\begin{array}{|l} \mathbf{x} \\ 1 \\ 2 \end{array} \begin{array}{l} \text{NEGBINOM.DIST}(10; 2; 1/6; \text{NEPRAVDA}) + \\ \dots + \text{NEGBINOM.DIST}(15; 2; 1/6; \text{NEPRAVDA}) \end{array}$$

Nebo také: $\mathbb{P}(12 \leq W \leq 17) = F(17) - F(11) \doteq 0,232,$

$$= \text{NEGBINOM.DIST}(15; 2; 1/6; \text{PRAVDA}) - \text{NEGBINOM.DIST}(9; 2; 1/6; \text{PRAVDA})$$

$$\text{d) } \mathbb{P}(W \geq 100) = p(100) + p(101) + p(102) + \dots =$$

$$= 1 - \mathbb{P}(W < 100) = 1 - \mathbb{P}(X \leq 99) = 1 - F(99) =$$

$$= \sum_{n=2}^{99} \binom{n-1}{1} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^{99-2} \doteq 3,014 \cdot 10^{-7}.$$

$$= 1 - \text{NEGBINOM.DIST}(97; 2; 1/6; \text{PRAVDA})$$

△

7.7 Hypergeometrické rozdělení

Diskrétní rozdělení uvedená v předchozích odstavcích vycházela z Bernoulliho procesu, který můžeme názorně reprezentovat jako opakovaný experiment tahů s vracením, kde pravděpodobnost úspěchu p zůstává pro všechny tahy neměnná. Pokud však tažený prvek do populace nevracíme, pravděpodobnost úspěchu se bude v závislosti na výsledcích jednotlivých tahů měnit. Mějme náhodnou veličinu X reprezentující počet úspěchů v takové posloupnosti n tahů bez vracení z osudí o celkovém počtu N prvků, z nichž M prvků

představuje úspěch⁸. Tato náhodná veličina má tzv. *hypergeometrické rozdělení* a budeme ji značit $\text{HGeom}(n; M; N)$ ⁹.

Pravděpodobnostní funkci hypergeometrického rozdělení snadno odvodíme pomocí kombinatorických pravidel (srovnej s příkladem 2.6.b))

$$p(x) = \begin{cases} \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}} & \text{pro } x = \max\{0, M - N + n\}, \dots, \min\{n, M\}, \\ 0 & \text{jinak.} \end{cases}$$

Distribuční funkce je pak ve tvaru

$$F(x) = \begin{cases} 0 & \text{pro } x < [M - N + n]^+, \\ \sum_{k=[M-N+n]^+}^{\lfloor x \rfloor} \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}} & \text{pro } [M - N + n]^+ \leq x \leq \min\{n, M\}, \end{cases}$$

kde $[M - N + n]^+ = \max\{0, M - N + n\}$ představuje tzv. kladnou část čísla $M - N + n$. Vztahy pro střední hodnotu a rozptyl jsou

$$\mathbb{E}X = n \frac{M}{N}, \quad \text{var } X = n \frac{M}{N} \left(1 - \frac{M}{N}\right) \frac{N - n}{N - 1}.$$

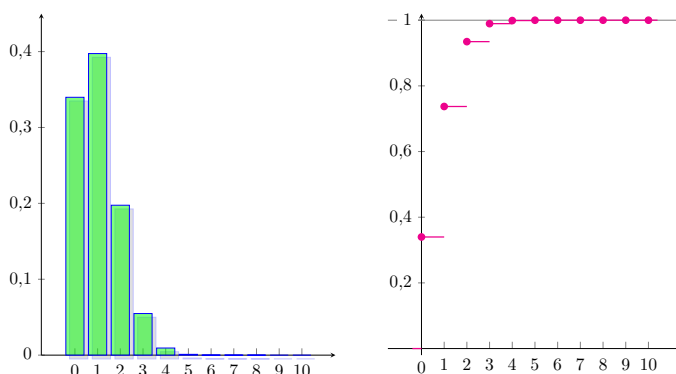
Příklad 7.13. Balení obsahuje 200 kusů zboží. Při převzetí zboží probíhá kontrola deseti náhodně vybraných kusů zboží z celého balení, neboť během přepravy může dojít k jeho poškození, což se stává pravidelně v 10 % všech výrobků. Zboží je převzato, jestliže v kontrolovaném vzorku jsou nejvýše dva rozbité výrobky. Určete pravděpodobnostní a distribuční funkci rozdělení náhodné veličiny, která udává počet kusů rozbitého zboží, a sestrojte jejich grafy. S jakou pravděpodobností bude balení s výrobky převzato?

Řešení. Všech výrobků v balení je $N = 200$, rozbitých výrobků $M = 20$ (10 % všech výrobků), nerozbitých výrobků $N - M = 200 - 20 = 180$, počet náhodně vybíraných výrobků je $n = 10$. Počet kusů rozbitého zboží je náhodná veličina s hypergeometrickým rozdělením, tj. $X \sim \text{HGeom}(10; 180; 200)$. Pravděpodobnostní funkce má pro $k = 0, 1, \dots, 10$ tvar

$$p(k) = \frac{\binom{20}{k} \binom{180}{10-k}}{\binom{200}{10}}.$$

⁸V případě experimentu s vracením by tato veličina měla binomické rozdělení s parametrem $p = \frac{M}{N}$.

⁹Označuje se také $\text{Hg}(n; M; N)$ a někdy i s jiným pořadím jednotlivých parametrů.



Obrázek 7.6: Graf pravděpodobnostní a distribuční funkce hypergeometrického rozdělení $HGeom(10; 180; 200)$

Grafy pravděpodobnostní a distribuční funkce ukazuje obrázek 7.6. Pravděpodobnost, že v balení jsou nejvýše dva rozbité výrobky, je

$$\begin{aligned} \mathbb{P}(X \leq 2) &= F(2) = p(0) + p(1) + p(2) = \\ &= \frac{\binom{20}{0} \binom{180}{10-0}}{\binom{200}{10}} + \frac{\binom{20}{1} \binom{180}{10-1}}{\binom{200}{10}} + \frac{\binom{20}{2} \binom{180}{10-2}}{\binom{200}{10}} \doteq 0,935. \end{aligned}$$

```
x
1 (KOMBINACE(180;10)+20*KOMBINACE(180;9)
2 +KOMBINACE(20;2)*KOMBINACE(180;8))/KOMBINACE(200;10)
```

nebo jednodušeji

```
= HYPGEOM.DIST(2;10;20;200;PRAVDA).
```

Zboží bude převzato s pravděpodobností asi 93,5 %.

△

Všimněme si, že binomické rozdělení je limitním případem hypergeometrického rozdělení: jestliže $N \rightarrow \infty$ a $\frac{M}{N} \rightarrow p \in (0; 1)$, potom pro $k = 0, 1, \dots, n$ platí

$$\lim_{N \rightarrow \infty} \frac{\binom{M}{k} \binom{N-M}{n-k}}{\binom{N}{n}} = \binom{n}{k} p^k (1-p)^{n-k}.$$

Jsou-li tedy počty prvků N a M výrazně vyšší než počet tahů n , pravděpodobnosti v modelu s vrácením a bez vrácení se nebudou od sebe příliš odlišovat. V příkladu 7.13 by například pravděpodobnost převzetí zboží vypočtená přibližně pomocí binomického rozdělení činila 0,930.

```
= BINOM.DIST(2;10;20/200;PRAVDA)
```


7.8 Poissonovo rozdělení

V Bernoulliho procesu jsme sledovali výsledek náhodného pokusu v předem daných diskrétních časových okamžicích: každý takový okamžik odpovídá jednomu Bernoulliho pokusu. Jestliže v nějakém Bernoulliho pokusu pozorujeme úspěch, řekjeme, že nastala *událost*. V takovém kontextu se Bernoulliho proces také alternativně nazývá jako *proces událostí v diskrétním čase*. Nyní uvolníme podmínku na diskrétní časové okamžiky: mějme časový interval, prozatím pro zjednodušení délky jedné časové jednotky¹⁰, a na tomto intervalu sledujeme výskyt určité události. Jestliže jednotlivé události přicházejí náhodně, jednotlivě a vzájemně nezávisle (níže vysvětlíme, co to znamená přesně), potom hovoříme o tzv. *procesu událostí ve spojitém čase* nebo též o *Poissonově procesu*. Poissonův proces je charakterizován následujícími axiomy:

- 1) počty událostí v libovolných dvou nepřekrývajících se časových intervalech jsou vzájemně nezávislé jevy;
- 2) pravděpodobnost výskytu jedné události v krátkém časovém intervalu $(t; t + h]$ (pro $t \in [0; 1)$ a malé $h > 0$) je přímo úměrná délce tohoto intervalu, s konstantním koeficientem úměry λ . Matematicky

$$\lim_{h \rightarrow 0} \frac{1}{h} \mathbb{P}(\text{jedna událost v intervalu } (t; t + h]) = \lambda;$$

- 3) pravděpodobnost výskytu dvou nebo více událostí v tomtéž časovém okamžiku je zanedbatelná. Matematicky

$$\lim_{h \rightarrow 0} \frac{\mathbb{P}(\text{dvě či více událostí v intervalu } (t; t + h])}{\mathbb{P}(\text{jedna událost v intervalu } (t; t + h])} = 0.$$

Parametr $\lambda > 0$ se nazývá *intenzitou událostí* Poissonova procesu a představuje průměrný počet sledovaných událostí za zvolenou časovou jednotku.

Uvažujme nyní náhodnou veličinu X udávající počet událostí za zvolenou časovou jednotku. Tato náhodná veličina nabývající hodnot $k = 0, 1, 2, \dots$ má tzv. *Poissonovo rozdělení* s parametrem intenzity λ , které budeme označovat $\text{Po}(\lambda)$. Pravděpodobnostní funkci lze odvodit pomocí řešení diferenciální rovnice vycházející z axiomů Poissonova procesu; má pro $k = 0, 1, 2, \dots$ tvar

$$\mathbb{P}(X = k) = p_k = \frac{1}{k!} \lambda^k e^{-\lambda}.$$

Zde pouze ověříme podmínky na pravděpodobnosti p_k :

- a) $p_k \geq 0$ pro libovolné $k = 0, 1, 2, \dots$,

¹⁰V odstavci 8.2 tento předpoklad zobecníme.

$$\text{b) } \sum_{k=0}^{\infty} p_k = e^{-\lambda} \sum_{k=0}^{\infty} \frac{1}{k!} \lambda^k = e^{-\lambda} \cdot e^{\lambda} = 1.^{11}$$

Distribuční funkce je tvaru

$$F(x) = \begin{cases} 0 & \text{pro } x < 0, \\ \sum_{k=0}^{\lfloor x \rfloor} \frac{1}{k!} \lambda^k e^{-\lambda} & \text{pro } x \geq 0. \end{cases}$$

Střední hodnota $\mathbb{E}X$ Poissonova rozdělení je rovna parametru λ , neboť

$$\begin{aligned} \mathbb{E}X &= \sum_{k=0}^{\infty} k p_k = \sum_{k=1}^{\infty} k \frac{1}{k!} \lambda^k e^{-\lambda} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{1}{(k-1)!} \lambda^{k-1} = \\ &= \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{1}{j!} \lambda^j = \lambda e^{-\lambda} e^{\lambda} = \lambda. \end{aligned}$$

Vidíme, že je to v souladu s pojmenováním parametru λ jako intenzity, tj. průměrného počtu událostí za časovou jednotku. Pro výpočet rozptylu použijeme vztah

$$\text{var } X = \mathbb{E}X(X-1) + \mathbb{E}X - (\mathbb{E}X)^2,$$

kde

$$\begin{aligned} \mathbb{E}X(X-1) &= \sum_{k=0}^{\infty} k(k-1) \frac{1}{k!} \lambda^k e^{-\lambda} = e^{-\lambda} \lambda^2 \sum_{k=2}^{\infty} \frac{1}{(k-2)!} \lambda^{k-2} = \\ &= e^{-\lambda} \lambda^2 \sum_{j=0}^{\infty} \frac{1}{j!} \lambda^j = e^{-\lambda} \lambda^2 e^{\lambda} = \lambda^2. \end{aligned}$$

Odtud již

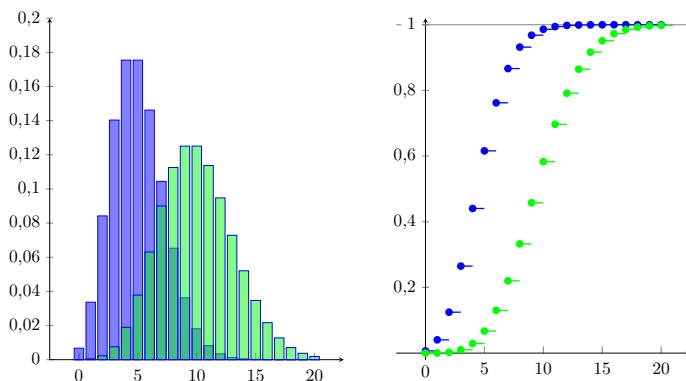
$$\text{var } X = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Vidíme tedy, že u Poissonova rozdělení je také rozptyl roven λ .

Příklad 7.14. Je dána náhodná veličina $X \sim \text{Po}(5)$. Vypočítejte:

- $\mathbb{P}(X = 6)$,
- $\mathbb{P}(X \leq 6)$,
- $\mathbb{P}(X < 6)$,
- $\mathbb{P}(3 < X \leq 5)$,
- $\mathbb{P}(X > 7)$.

¹¹Pro součet byla využita Taylorova řada $e^{\lambda} = \sum_{k=0}^{\infty} \frac{1}{k!} \lambda^k$.



Obrázek 7.7: Graf pravděpodobnostních a distribučních funkcí Poissonova rozdělení $Po(5)$ (modrá barva), $Po(10)$ (zelená barva)

Řešení.

a) $\mathbb{P}(X = 6) = p(6) = \frac{1}{6!} 5^6 e^{-5} \doteq 0,146,$

= `POISSON.DIST(6;5;NEPRAVDA)`

b) $\mathbb{P}(X \leq 6) = p(0) + \dots + p(6) = F(6) = \sum_{k=0}^6 \frac{1}{k!} 5^k e^{-5} \doteq 0,762,$

= `POISSON.DIST(6;5;PRAVDA)`

c) $\mathbb{P}(X < 6) = p(0) + \dots + p(5) = F(5) = \sum_{k=0}^5 \frac{1}{k!} 5^k e^{-5} \doteq 0,616,$

= `POISSON.DIST(5;5;PRAVDA)`

d) $\mathbb{P}(3 < X \leq 5) = p(4) + p(5) = F(5) - F(3) = \sum_{k=4}^5 \frac{1}{k!} 5^k e^{-5} \doteq 0,351,$

= `POISSON.DIST(4;5;NEPRAVDA)+POISSON.DIST(5;5;NEPRAVDA)`

nebo také

= `POISSON.DIST(5;5;PRAVDA)-POISSON.DIST(3;5;PRAVDA)`

e) $\mathbb{P}(X > 7) = 1 - \mathbb{P}(X \leq 7) = 1 - F(7) \doteq 0,133.$

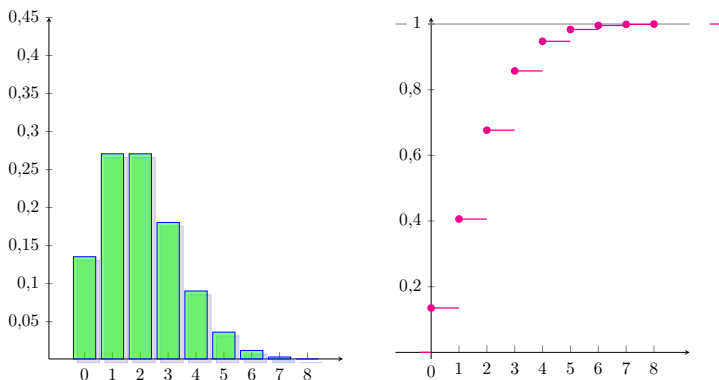
= `1-POISSON.DIST(7;5;PRAVDA)`

△

Příklad 7.15. Sestrojte grafy pravděpodobnostních a distribučních funkcí náhodných veličin s Poissonovým rozdělením $Po(5)$ a $Po(10)$.

Tabulka 7.1: Teoretické pravděpodobnosti výskytu daného počtu telefonních výzev za 1 minutu k příkladu 7.16

x	0	1	2	3	4	5	6	7	8
$p(x)$	0,135	0,271	0,271	0,180	0,090	0,036	0,012	0,003	0,001



Obrázek 7.8: Graf pravděpodobnostní a distribuční funkce k příkladu 7.16

Řešení. Grafy pravděpodobnostních a distribučních funkcí Poissonova rozdělení ukazují obrázek 7.7. $\triangle$

Poissonovým rozdělením se řídí náhodná veličina, kterou je počet výskytů sledovaného jevu v určitém časovém intervalu. Předpokládáme, že jev může nastat v kterémkoliv okamžiku a počet výskytů během tohoto časového intervalu závisí jen na jeho délce a ne na jeho počátku ani na tom, kolikrát jev nastal před jeho počátkem. Náhodnou veličinou, která má Poissonovo rozdělení, je tedy například počet vadných výrobků ve velké sérii (je-li pravděpodobnost vadného výrobku velmi malá), počet zákazníků, kteří přišli do prodejny v určitém časovém intervalu, počet telefonních zavolaní během nějakého časového intervalu, počet hlášení škodní události během roku v rámci uzavřeného pojištění, počet případů onemocnění vzácnou chorobou během jednoho měsíce nebo počet zákazníků, kteří během jednoho měsíce uzavřou smlouvu o poskytnutí úvěru. Často se Poissonovo rozdělení označuje jako „zákon řídkých jevů“.

Příklad 7.16. Uvažujme náhodnou veličinu X , která udává počet telefonních výzev za 1 minutu s intenzitou průměrně dvou hovorů za uvedenou dobu. Sestrojte graf pravděpodobnostní funkce $p(x)$ a distribuční funkce $F(x)$ rozdělení náhodné veličiny X .

Řešení. Náhodná veličina X má rozdělení $Po(2)$. Tabulka 7.1 obsahuje část hodnot a pravděpodobností náhodné veličiny X . Ve skutečnosti může veličina X nabývat spočetně mnoha hodnot, tj. během jedné minuty se může udát libovolně velký (přirozený) počet výzev, byť s velmi malými pravděpodobnostmi. Graf pravděpodobnostní funkce $p(x)$ a distribuční funkce $F(x)$ ukazují obrázek 7.8. $\triangle$

Příklad 7.17. Na nádraží mají být instalovány automaty na prodej jízdenek, které po vhození příslušné mince vydají během deseti sekund žádanou jízdenku. Předpokládejme, že v době největší frekvence cestujících bude chtít použít automat v průměru šest osob za minutu. Kolik automatů je nutné instalovat, aby s pravděpodobností větší než 0,95 byl v době největší frekvence obslužen každý zájemce okamžitě?



Řešení. Náhodnou veličinou X bude počet zákazníků za časovou jednotku deseti sekund v době největší frekvence cestujících. V této době je intenzita zájmu o odbavení 6 osob za minutu. Veličina X má zřejmě Poissonovo rozdělení s parametrem $\lambda = \frac{10}{60} \cdot 6 = 1$ cestující za 10 sekund. Hledáme nyní nejmenší hodnotu k , pro kterou platí, že ji náhodná veličina X nepřekročí s pravděpodobností větší než 0,95, tzn. nejmenší hodnotu k , pro kterou platí $\mathbb{P}(X \leq k) > 0,95$. Protože

$$\mathbb{P}(X \leq 0) = e^{-1} \doteq 0,368$$

$$\mathbb{P}(X \leq 1) = e^{-1} + e^{-1} \doteq 0,736,$$

$$\mathbb{P}(X \leq 2) = e^{-1} + e^{-1} + \frac{e^{-1}}{2} \doteq 0,920,$$

$$\mathbb{P}(X \leq 3) = e^{-1} + e^{-1} + \frac{e^{-1}}{2} + \frac{e^{-1}}{6} \doteq 0,981,$$

musíme instalovat nejméně tři automaty, aby s pravděpodobností větší než 0,95 byl každý zájemce obslužen okamžitě. $\triangle$

Pro dostatečně velké n (např. alespoň $n > 30$) a dostatečně malé p (např. $p < 0,1$) lze binomické rozdělení aproximovat Poissonovým rozdělením s parametrem $\lambda = np$. Ve skutečnosti platí, že Poissonovo rozdělení $Po(\lambda)$ je limitním případem binomického rozdělení $Bi(n; p)$, jestliže $n \rightarrow \infty$, $p \rightarrow 0$ a $np \rightarrow \lambda$. Vyšetřeme k -tý člen binomického rozdělení pro speciální případ

$n \rightarrow \infty$ a $p \rightarrow 0$, kde $np = \lambda$:

$$\begin{aligned} \lim_{n \rightarrow \infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} &= \\ &= \lim_{n \rightarrow \infty} \frac{n(n-1) \cdots (n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} = \\ &= \frac{\lambda^k}{k!} \lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right) \left(1 - \frac{\lambda}{n}\right)^{-k} \lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^n. \end{aligned}$$

První limita je rovna 1, druhá limita je $e^{-\lambda}$, čímž dostáváme přímo předpis pravděpodobnostní funkce Poissonova rozdělení. Tuto skutečnost demonstujeme na příkladu 7.18.

Příklad 7.18. Předpokládejme, že pojišťovna zaznamená za určité období 1 000 pojistných událostí. Pravděpodobnost, že během jedné minuty nastane pojistná událost, je 0,006. S jakou pravděpodobností nastane během jedné minuty deset pojistných událostí? Řešte užitím Poissonova a binomického rozdělení pravděpodobnosti. Výsledky zaokrouhlete na šest desetinných míst a porovnejte je.

Řešení. $\lambda = np = 1\,000 \cdot 0,006 = 6$, tedy během jedné minuty očekáváme průměrně šest pojistných událostí. Tedy, pomocí Poissonova rozdělení,

$$\mathbb{P}(X = 10) = e^{-6} \frac{6^{10}}{10!} \doteq 0,041\,303.$$

= `POISSON.DIST(10;6;NEPRAVDA)`

Pomocí binomického rozdělení dostáváme

$$\mathbb{P}(X = 10) = \binom{1000}{10} 0,006^{10} 0,994^{990} \doteq 0,041\,178.$$

= `BINOM.DIST(10;1000;0,006;NEPRAVDA)`

Při zvolené aproximaci se výsledky liší až na čtvrtém desetinném místě.

△

7.9 Úlohy

7.1 Dokažte, že náhodná veličina X řídící se rovnoměrným diskrétním rozdělením na množině $D = \{1, 2, \dots, n\}$ má střední hodnotu $\mathbb{E}X$ rovnou $\frac{n+1}{2}$ a rozptyl $\text{var } X = \frac{n^2-1}{12}$.

7.2 Vytvořte grafy pravděpodobnostní a distribuční funkce náhodné veličiny X s degenerovaným rozdělením $\text{Dg}(1)$.

7.3 Dokažte, že náhodná veličina X řídící se alternativním rozdělením má střední hodnotu $\mathbb{E}X$ rovnou p a rozptyl $\text{var } X = p(1-p)$.

7.4 Výrobce dodává výrobky v baleních po 15 kusech. Za každé balení, v němž je alespoň jeden výrobek vadný, budou zákazníkovi vráceny peníze za celý balík. Je známo, že pravděpodobnost vyrobení kvalitního výrobku je 0,95 a že náklady na výrobu jednoho celého balení jsou 2 \$. Jakou cenu jednoho výrobku musí výrobce stanovit, aby mohl očekávat zisk 23 %? (Ziskem rozumíme rozdíl příjmů z prodeje výrobků a nákladů na veškerou výrobu včetně výrobků reklamovaných.)

7.5 Nechť X je diskrétní náhodná veličina s oborem hodnot $\{1, 2, 3, 4, 5\}$ a pravděpodobnostmi $\mathbb{P}(X = i) = \frac{c}{i}$ pro $i = 1, 2, \dots, 5$.

- a) Určete konstantu c .
- b) Určete $\mathbb{E}X$ a $\text{var } X$.

7.6 Uvažujte osm známých krevních skupin $A+$, $A-$, $B+$, $B-$, $AB+$, $AB-$, $0+$ a $0-$ a předpokládejme, že se vyskytují se stejnou pravděpodobností. Jaká je pravděpodobnost, že pro nalezení třech dárců krevní skupiny $A+$ budeme muset vyšetřit:

- a) právě deset osob neznajících svou krevní skupinu (A),
- b) více než devět osob neznajících svou krevní skupinu (B),
- c) více než sedm a méně než dvanáct osob neznajících svou krevní skupinu (C)?

7.7 Mezi sty vejci je 5 % prasklých. Jaká je pravděpodobnost, že když vybereme 20 vajec, budou z nich:

- a) „právě dvě prasklá“ (A),
- b) „alespoň dvě prasklá“ (B),

c) „nejvýše dvě prasklá“ (C)?

7.8 Telefonní ústředna zapojí během hodiny průměrně 15 hovorů. Jaká je pravděpodobnost, že během čtyř minut zapojí ústředna

- a) „právě jeden hovor“ (jev A),
- b) „alespoň dva hovory“ (jev B),
- c) „alespoň dva hovory a nejvýše pět hovorů“ (jev C)?

7.9 Účastník volá telefonní ústřednu v době největšího zatížení linky, kdy pravděpodobnost, že linka nebude obsazena, je 0,25. Jednotlivé pokusy o spojení opakuje po několika minutách tak dlouho, dokud není spojen. Určete

- a) pravděpodobnost, že „dosáhne spojení až při pátém pokusu“ (jev A),
- b) střední hodnotu a rozptyl počtu neúspěšných pokusů do okamžiku, kdy je navázáno spojení.

7.10 Hráč hází kostkou tak dlouho, dokud nepadnou tři šestky. Jaká je pravděpodobnost, že hráč bude muset hodit kostkou desetkrát?

7.11 Střední hodnota počtu poruch vysílače za 10 000 hodin jeho činnosti je rovna 10. Předpokládejme, že jednotlivé výskyty poruch v čase jsou na sobě nezávislé. Určete pravděpodobnost poruchy zařízení za 100 hodin činnosti.

7.12 Na telefonní ústřednu je napojeno 326 účastníků. Každý z nich bude ústřednu volat během hodiny s pravděpodobností 0,01. Předpokládejme, že časy hovorů jednotlivých účastníků jsou na sobě nezávislé. Jaká je pravděpodobnost, že během hodiny zavolají

- a) 4 účastníci,
- b) alespoň 6 účastníků,
- c) nejvýše 2 účastníci?

7.13 Na 500 stranách knihy je 500 tiskových chyb. Předpokládejme, že jednotlivé výskyty chyb jsou vzájemně nezávislé. Určete pravděpodobnost toho, že na náhodně vybrané stránce jsou

- a) alespoň tři chyby,
- b) právě jedna chyba,
- c) nejvýše dvě chyby.

7.14 Během hodiny přijde do výčepu průměrně 60 hostů. Jaká je pravděpodobnost, že během půl minuty, ve které nikdo neobsluhuje, neprijde žádný zákazník?

7.15 Určete pravděpodobnost toho, že mezi 136 výrobky je/jsou

- a) právě sedm vadných výrobků,
- b) alespoň tři vadné výrobky,
- c) nejvýše deset vadných výrobků,

jestliže víme, že vadné výrobky tvoří průměrně 2,6 % produkce.

7.16 Z důvodu úspory peněz na drahých krevních testech při testování černého kašle během druhé světové války přišla armáda s následujícím plánem: místo samostatného testování vzorku krve každého vojáka smíchali krev deseti vojáků a otestovali směs. Jestliže byl test negativní, věděli, že všichni vojáci ve skupině jsou negativní. (Uvažujme, že smíchání více vzorků krve nemá vliv na identifikaci nákazy.) Jestliže byl test pozitivní, museli otestovat každého vojáka z dané skupiny samostatně. Je jasné, že pokud je mezi všemi vojáky počet těch nakažených vysoký, tento přístup se nevyplatí. Za jakých podmínek ušetří tento postup peníze oproti okamžitému testování každého vojáka samostatně? Tj., určete nejvyšší možnou pravděpodobnost, s jakou může být testovaný voják pozitivní, aby se testování deseti vojáků současně vyplatilo.

7.17 Restaurace dává ke každému jídlu kartičku s obrázkem jednoho hráče místního týmu. Při každé návštěvě restaurace obdržíte náhodně jednu kartičku.

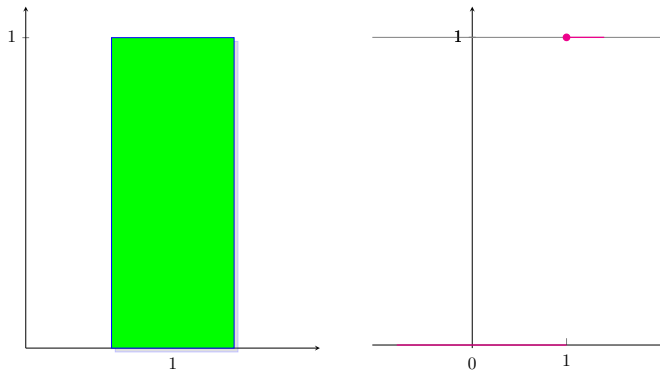
- a) Jestliže karty zobrazují basketbalisty (pět hráčů základní sestavy), kolikrát musíte jít v průměru do restaurace, abyste získali od všech hráčů alespoň jednu kartičku?
- b) Jestliže karty zobrazují baseballisty (devět hráčů základní sestavy), kolikrát musíte jít v průměru do restaurace, abyste získali od všech hráčů alespoň jednu kartičku?

Řešení úloh

7.1

$$\begin{aligned}
 \mathbb{E}X &= \sum_{i=1}^n i \cdot \frac{1}{n} = \frac{1}{n} \cdot 1 + \frac{1}{n} \cdot 2 + \cdots + \frac{1}{n} \cdot n = \\
 &= \frac{1}{n} \cdot (1 + \cdots + n) = \frac{1}{n} \cdot \frac{n}{2} \cdot (n+1) = \frac{n+1}{2}, \\
 \mathbb{E}X^2 &= \frac{1}{n} \cdot 1^2 + \frac{1}{n} \cdot 2^2 + \cdots + \frac{1}{n} \cdot n^2 = \frac{1}{n} \cdot (1^2 + 2^2 + \cdots + n^2) = \\
 &= \frac{1}{n} \cdot \frac{n \cdot (n+1) \cdot (2n+1)}{6} = \frac{(n+1) \cdot (2n+1)}{6}, \\
 \text{var } X &= \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{(n+1) \cdot (2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{n^2 - 1}{12}.
 \end{aligned}$$

7.2



Obrázek 7.9: Graf pravděpodobnostní a distribuční funkce pro úlohu 7.2

7.3

$$\begin{aligned}
 \mathbb{E}X &= 0 \cdot (1-p) + 1 \cdot p = p, \\
 \mathbb{E}X^2 &= 0^2(1-p) + 1^2p = p, \\
 \text{var } X &= \mathbb{E}X^2 - (\mathbb{E}X)^2 = p - p^2 = p(1-p).
 \end{aligned}$$

7.4 Pravděpodobnost, že balení nebude obsahovat vadný výrobek a bude z něj plynout zisk, je $0,95^{15} \doteq 0,463\,291$. Například pro 100 výrobků budou výdaje 200 \$ a příjmy musí být 246 \$, aby byl zisk 23 %. Cena za jeden výrobek tedy musí být $246 : 46,3291 = 5,31$ \$.

7.5

$$\text{a) } 1 = \sum_{i=1}^5 \mathbb{P}(X = i) = c \cdot \sum_{i=1}^5 \frac{1}{i} = c \cdot \left(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5}\right) = c \cdot \frac{137}{60},$$

$$\Rightarrow c = \frac{60}{137}.$$

$$\text{b) } \mathbb{E}X = \sum_{i=1}^5 i \cdot \mathbb{P}(X = i) = \frac{60}{137} \cdot \sum_{i=1}^5 1 = \frac{300}{137} \doteq 2,190;$$

$$\mathbb{E}(X^2) = \sum_{i=1}^5 i^2 \cdot \mathbb{P}(X = i) = \frac{60}{137} \cdot \sum_{i=1}^5 i = \frac{60}{137} \cdot \frac{5 \cdot 6}{2} = \frac{900}{137},$$

$$\text{var } X = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \frac{900}{137} - \left(\frac{300}{137}\right)^2 \doteq 1,774.$$

$$\text{7.6 a) } \mathbb{P}(A) \doteq 0,028; \text{ b) } \mathbb{P}(B) \doteq 0,908; \text{ c) } \mathbb{P}(C) \doteq 0,103.$$

$$\text{7.7 a) } \mathbb{P}(A) \doteq 0,207; \text{ b) } \mathbb{P}(B) \doteq 0,261; \text{ c) } \mathbb{P}(C) \doteq 0,947.$$

$$\text{7.8 a) } \mathbb{P}(A) \doteq 0,3679; \text{ b) } \mathbb{P}(B) \doteq 0,2642; \text{ c) } \mathbb{P}(C) \doteq 0,2636.$$

$$\text{7.9 a) } \mathbb{P}(A) \doteq 0,079; \text{ b) } 3; 12.$$

$$\text{7.10 } 0,0465$$

$$\text{7.11 } 0,095$$

$$\text{7.12 } \lambda = 3,26; \text{ a) } 0,1807; \text{ b) } 0,1124; \text{ c) } 0,3675.$$

$$\text{7.13 a) } 0,080; \text{ b) } 0,368; \text{ c) } 0,920.$$

$$\text{7.14 } 0,607$$

$$\text{7.15 a) } 0,039; \text{ b) } 0,689; \text{ c) } 0,999091.$$

$$\text{7.16 } p < 0,004087$$

$$\text{7.17 a) } 11,41, \text{ b) } 24,46.$$

Kapitola 8

Vybraná rozdělení spojitých náhodných veličin

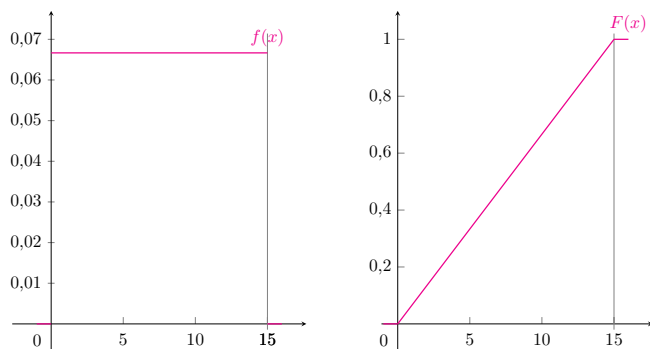
8.1 Rovnoměrné spojité rozdělení

Uvažujme interval $[a; b]$ reálných hodnot, ve kterém není žádná z hodnot z hlediska výskytu preferována. V takovém případě můžeme pro charakterizaci pravděpodobnostního chování využít model geometrické pravděpodobnosti diskutovaný v odstavci 2.3. Alternativně je možné zavést nad pravděpodobnostním prostorem náhodnou veličinu X s hodnotami v intervalu $[a; b]$, o které budeme říkat, že má na tomto intervalu tzv. *rovnoměrné spojité rozdělení* $U(a; b)$ ¹. Protože se jedná o spojité rozdělení, pravděpodobnostní chování je charakterizováno hustotou, která je vzhledem k zadanému chování na zvoleném intervalu $[a; b]$ konstantní. Dále rovněž snadno odvodíme distribuční funkci pomocí integrace – distribuční funkce je primitivní funkcí k hustotě a hustota je derivací distribuční funkce (viz též úlohu 8.2). Hustota rovnoměrného rozdělení a jeho distribuční funkce mají tedy tvar

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{pro } a \leq x \leq b, \\ 0 & \text{jinak,} \end{cases} \quad F(x) = \begin{cases} 0 & \text{pro } x < a, \\ \frac{x-a}{b-a} & \text{pro } a \leq x \leq b, \\ 1 & \text{pro } x > b. \end{cases}$$

Hustota rozdělení je symetrická kolem středu úsečky $[a; b]$, tento střed je tudíž střední hodnotou i mediánem náhodné veličiny X . Rozptyl náhodné

¹Označení U pochází z anglického slova *uniform*, česky jednotný, rovnoměrný, stejný, konstantní apod., což jsou mj. vlastnosti grafu hustoty pravděpodobnosti tohoto rozdělení, jde totiž o úsečku rovnoběžnou s osou x . V české literatuře se rovnoměrné spojité rozdělení označuje $Rs(a; b)$ nebo jen $R(a; b)$.



Obrázek 8.1: Hustota a distribuční funkce pro příklad 8.1

veličiny X dopočítáme pomocí výpočetního vzorce (6.7) (viz úlohu 8.1), odkud dostaneme

$$\mathbb{E}X = \tilde{X} = \frac{a+b}{2}, \quad \text{var } X = \frac{1}{12}(b-a)^2. \quad (8.1)$$

Příklad 8.1. Vytvořte graf hustoty pravděpodobnosti a graf distribuční funkce rovnoměrného spojitého rozdělení náhodné veličiny X v intervalu $[0; 15]$. Vypočítejte střední hodnotu a rozptyl náhodné veličiny X .

Řešení. $f(x) = \frac{1}{15}$, $F(x) = \frac{x}{15}$, $x \in [0; 15]$. Graf hustoty a distribuční funkce ukazuje obrázek 8.1. Střední hodnota a rozptyl mají hodnoty

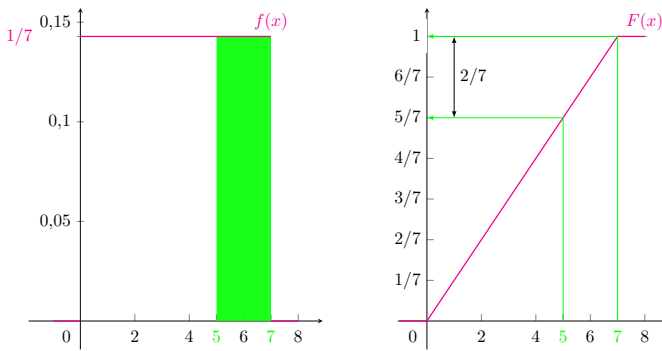
$$\mathbb{E}X = \frac{0+15}{2} = 7,5, \quad \text{var } X = \frac{1}{12}(15-0)^2 = 18,75.$$

△

S rovnoměrným rozdělením $U(a; b)$ se setkáváme například při vyšetřování chyb ze zaokrouhlování v numerických výpočtech. Jsou-li čísla vstupující do výpočtů nekonečné desetinné zlomky, jež se zaokrouhlují na k desetinných míst, pak lze chybu ze zaokrouhlení považovat za náhodnou veličinu s rovnoměrným rozdělením na intervalu $[-5 \cdot 10^{-k-1}; 5 \cdot 10^{-k-1}]$.

U rovnoměrného spojitého rozdělení je k dispozici velmi názorná vizualizace výpočtů pravděpodobnosti, počítáme totiž obsah obdélníka pod hustotou pravděpodobnosti (viz příklad 8.2 a obrázek 8.2).

Příklad 8.2. Autobusy městské dopravy odjíždějí ze zastávky v sedmiminutových intervalech. Cestující může přijít na zastávku v libovolném okamžiku. Jaká je pravděpodobnost, že bude náhodně příchozí muset čekat více než 5 minut? Jaká je střední hodnota a rozptyl doby jeho čekání na odjezd autobusu ze stanice?



Obrázek 8.2: Hustota, distribuční funkce a výpočet pravděpodobnosti k příkladu 8.2

Řešení. Doba čekání na odjezd autobusu je náhodná veličina s rovnoměrným rozdělením v intervalu $[0; 7]$, tudíž

$$f(x) = \frac{1}{7}$$

na intervalu $[0; 7]$. Střední hodnota a rozptyl mají podle vzorce (8.1) hodnoty

$$\mathbb{E}X = \frac{0 + 7}{2} = 3,5 \text{ min} \quad \text{a} \quad \text{var } X = \frac{(7 - 0)^2}{12} = \frac{49}{12} = 4,08\bar{3} \text{ min}^2.$$

Pravděpodobnost, že náhodně příchozí bude muset čekat na zastávce více jak pět minut, je

$$\mathbb{P}(X \geq 5) = \int_5^7 \frac{1}{7} dx = \frac{1}{7}(7 - 5) = \frac{2}{7} \doteq 0,286.$$

Výpočet této pravděpodobnosti je ilustrován na obrázku 8.2.

△

8.2 Exponenciální rozdělení

V odstavci 7.8 jsme si představili Poissonův proces, tj. proces událostí ve spojitém čase s parametrem intenzity λ událostí za předepsanou časovou jednotku. Uvažujme náhodnou veličinu T představující dobu, která uplyne mezi dvěma po sobě následujícími událostmi Poissonova procesu. Nyní odvodíme pravděpodobnostní rozdělení této náhodné veličiny. Definujme nejprve pomocnou náhodnou veličinu X , reprezentující počet událostí, které nastanou za předepsanou dobu $t > 0$ od poslední události. Tato veličina

má Poissonovo rozdělení s parametrem λt . Pravděpodobnost, že doba čekání na následující událost překročí čas t , je tedy shodná s pravděpodobností, že za dobu t nenastane žádná událost, tj.

$$\mathbb{P}(T > t) = \mathbb{P}(X = 0) = \frac{1}{0!}(\lambda t)^0 e^{-\lambda t} = e^{-\lambda t} \quad (8.2)$$

pro $t > 0$. Odtud přímo dostaneme předpis pro distribuční funkci náhodné veličiny T ,

$$F_T(t) = \begin{cases} 0 & \text{pro } t \leq 0, \\ 1 - \mathbb{P}(T > t) = 1 - e^{-\lambda t} & \text{pro } t > 0 \end{cases}$$

a derivováním $F_T(t)$ též předpis hustoty náhodné veličiny T ,

$$f_T(t) = \begin{cases} 0 & \text{pro } t \leq 0, \\ \lambda e^{-\lambda t} & \text{pro } t > 0. \end{cases}$$

Rozdělení náhodné veličiny T se nazývá *exponenciální rozdělení* a značí se $\text{Exp}(\lambda)^2$. Snadno ověříme, že funkce $f(t)$ je skutečně hustota. Především $f(t) \geq 0$ pro libovolné $t \in \mathbb{R}$ a dále

$$\int_{-\infty}^{\infty} f(t) dt = \int_0^{\infty} \lambda e^{-\lambda t} dt = 1,$$

podle pravidla pro integrování exponenciální funkce

$$\int e^{ct} dt = \frac{1}{c} e^{ct} \quad \text{pro libovolné } c \in \mathbb{R}.$$

Opakovaným integrováním rovněž odvodíme střední hodnotu a rozptyl náhodné veličiny T , které jsou rovny

$$\mathbb{E}X = \frac{1}{\lambda} \quad \text{a} \quad \text{var } X = \frac{1}{\lambda^2}.$$

Exponenciální rozdělení $\text{Exp}(\lambda)$ je vhodným modelem pravděpodobnostní charakterizace „doby čekání na událost“ (doby čekání na realizaci určitého

²Někdy se označuje $\text{Ex}(\lambda)$ nebo jen $\text{E}(\lambda)$. Poznamenejme, že značení parametru exponenciálního rozdělení není v literatuře zcela jednotné: v některých publikacích se namísto intenzity λ používá parametr střední hodnoty $\beta = 1/\lambda$, tedy například $\text{Exp}(\beta)$. Mnohdy se navíc i pro parametr střední hodnoty využívá písmeno λ . Čtenář proto musí být při čtení literatury pozorný, aby si ujasnil, jakou formu parametrizace autor publikace pro značení exponenciálního rozdělení zvolil. V této knize používáme výhradně označení pomocí parametru intenzity s označením λ a písmeno β vyhradíme pro parametr střední hodnoty.

jevu), například doby životnosti určitého zařízení. V následující větě 8.1 ukážeme, že pravděpodobnostní rozdělení zbývajících doby nezávisí na tom, jak dlouho již čekáme. Říká se tomu, že exponenciální rozdělení „nemá paměť“.

Věta 8.1. *Má-li náhodná veličina T exponenciální rozdělení, pak*

$$\mathbb{P}(T > t_0 + t | T > t_0) = \mathbb{P}(T > t) \text{ pro libovolná } t_0 > 0, t > 0.$$

Důkaz. Nechť $T \sim \text{Exp}(\lambda)$, pak podle vzorce (8.2) a podle definice podmíněné pravděpodobnosti platí

$$\mathbb{P}(T > t_0 + t | T > t_0) = \frac{\mathbb{P}(X > t_0 + t)}{\mathbb{P}(X > y)} = \frac{e^{-\lambda(t_0+t)}}{e^{-\lambda t_0}} = e^{-\lambda t}.$$

□

Pro výpočet hodnot hustoty pravděpodobnosti, resp. distribuční funkce můžeme využít funkce Excelu `=EXPON.DIST`, jak je ukázáno v příkladu 8.3. Poslední parametr této funkce udává typ výpočtu: použijeme hodnotu `NEPRAVDA` pro výpočet hodnot hustoty pravděpodobnosti a `PRAVDA` pro výpočet hodnot distribuční funkce. První dva parametry funkce jsou postupně proměnná t (argument hustoty, resp. distribuční funkce) a hodnota parametru λ .

Příklad 8.3. Nechť náhodná veličina T má exponenciální rozdělení $\text{Exp}(\frac{1}{3})$. Vypočítejte:

- $\mathbb{P}(T = 1)$,
- $\mathbb{P}(T \leq 2)$,
- $\mathbb{P}(T > 5)$,
- $\mathbb{P}(T < 3)$,
- $\mathbb{P}(T \geq 5)$,
- $\mathbb{P}(2 < T \leq 4)$.

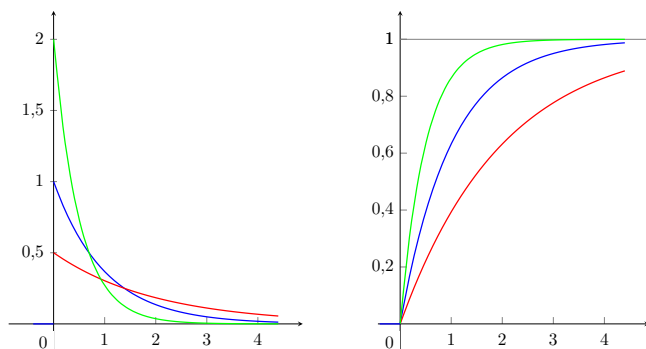
Řešení.

- $\mathbb{P}(T = 1) = 0$, neboť se jedná o spojité rozdělení,
- $\mathbb{P}(T \leq 2) = F(2) = 1 - e^{-\frac{1}{3} \cdot 2} \doteq 0,487$,

$$= \text{EXPON.DIST}(2; 1/3; \text{PRAVDA})$$

- $\mathbb{P}(T > 5) = 1 - \mathbb{P}(T \leq 5) = 1 - F(5) = 1 - \left(1 - e^{-\frac{1}{3} \cdot 5}\right) \doteq 0,189$.

$$= 1 - \text{EXPON.DIST}(5; 1/3; \text{PRAVDA})$$



Obrázek 8.3: Graf hustot a distribučních funkcí exponenciálního rozdělení: $\text{Exp}(\frac{1}{2})$ (červená čára), $\text{Exp}(1)$ (modrá čára), $\text{Exp}(2)$ (zelená čára)

d) $\mathbb{P}(T < 3) = \mathbb{P}(T \leq 3) = 1 - e^{-\frac{1}{3} \cdot 3} \doteq 0,632$,

= `EXPON.DIST(3;1/3;PRAVDA)`

e) $\mathbb{P}(T \geq 5) = \mathbb{P}(T > 5) \doteq 0,189$, viz c),

f) $\mathbb{P}(2 < T \leq 4) = F(4) - F(2) = 1 - e^{-\frac{1}{3} \cdot 4} - \left[1 - e^{-\frac{1}{3} \cdot 2}\right] \doteq 0,250$.

= `EXPON.DIST(4;1/3;PRAVDA) - EXPON.DIST(2;1/3;PRAVDA)`

△

Příklad 8.4. Vytvořte grafy hustot pravděpodobností a distribučních funkcí náhodné veličiny X s exponenciálním rozdělením

- $\text{Exp}(\frac{1}{2})$,
- $\text{Exp}(1)$,
- $\text{Exp}(2)$.

Řešení. Grafy hustot a distribučních funkcí jsou na obrázku 8.3. Všimněte si, že graf hustoty začíná vždy v hodnotě parametru intenzity λ na svislé ose a pak exponenciálně klesá. △

Příklad 8.5. Čas mezi příchody dvou po sobě jdoucích pacientů se řídí exponenciálním rozdělením se střední hodnotou 15 minut. Vypočítejte pravděpodobnost, že recepce bude čekat na dalšího pacienta:

- nejvýše 15 minut,
- nejméně 15 minut,
- právě 15 minut,
- mezi 5 a 20 minutami.

Řešení. Nechť T je doba mezi příchody dvou pacientů po sobě. Je dáno, že $\mathbb{E}T = 15 = \frac{1}{\lambda}$, tudíž $\lambda = \frac{1}{15}$ a $T \sim \text{Exp}(\frac{1}{15})$.

$$\begin{aligned} \text{a) } \mathbb{P}(T \leq 15) &= F(15) = \int_0^{15} \frac{1}{15} e^{-\frac{t}{15}} dt = \frac{1}{15} \int_0^{15} e^{-\frac{t}{15}} dt = \\ &= \frac{1}{15} \cdot (-15) \left[e^{-\frac{t}{15}} \right]_0^{15} = -e^{-1} + 1 \doteq 0,632, \\ &= \text{EXPON.DIST}(15; 1/15; \text{PRAVDA}) \end{aligned}$$

$$\begin{aligned} \text{b) } \mathbb{P}(T \geq 15) &= 1 - \mathbb{P}(T < 15) = 1 - F(15) \doteq 1 - 0,632 = 0,368, \\ &= 1 - \text{EXPON.DIST}(15; 1/15; \text{PRAVDA}) \end{aligned}$$

nebo přímo $\mathbb{P}(T \geq 15) = e^{-\frac{1}{15} \cdot 15} = e^{-1} \doteq 0,368$.

c) $\mathbb{P}(T = 15) = 0$, neboť se jedná o spojité rozdělení,

$$\begin{aligned} \text{d) } \mathbb{P}(5 < T < 20) &= \int_5^{20} \frac{1}{15} e^{-\frac{t}{15}} dt = \frac{1}{15} \int_5^{20} e^{-\frac{t}{15}} dt = \\ &= \frac{1}{15} \cdot (-15) \left[e^{-\frac{t}{15}} \right]_5^{20} = - \left(e^{-\frac{4}{3}} - e^{-\frac{1}{3}} \right) \doteq \\ &\doteq 0,7165 - 0,2636 \doteq 0,453, \end{aligned}$$

také ovšem

$$\mathbb{P}(5 < T < 20) = F(20) - F(5) \doteq 0,7364 - 0,2835 \doteq 0,453.$$

$$= \text{EXPON.DIST}(20; 1/15; \text{PRAVDA}) - \text{EXPON.DIST}(5; 1/15; \text{PRAVDA})$$

Recepce bude čekat na dalšího pacienta nejvýše 15 minut s pravděpodobností asi 63 %, nejméně 15 minut s pravděpodobností asi 37 %, právě 15 minut s nulovou pravděpodobností a mezi 5 a 20 minutami s pravděpodobností asi 45 %. $\triangle$

V programu MS Excel není k dispozici funkce na výpočet hodnot kvantilové funkce pro náhodnou veličinu s exponenciálním rozdělením. Proto je v případě potřeby provést manuální výpočet, jak je patrné v následujícím příkladu 8.6.

Příklad 8.6. Z dlouhodobých měření je známo, že notebook má poruchu v průměru jednou za 10 000 hodin. Předpokládejme, že „doba čekání na poruchu“ je náhodná veličina T s exponenciálním rozdělením. Stanovte hodnotu t tak, aby pravděpodobnost, že notebook bude pracovat delší dobu než t , byla 0,99.

Řešení. V tomto příkladu nejprve odvodíme předpis kvantilové funkce pro obecnou intenzitu λ a předepsanou pravděpodobnost α . Nechť

$$\alpha = F(t) = 1 - e^{-\lambda t},$$

a tedy

$$1 - \alpha = e^{-\lambda t},$$

odkud dostáváme

$$t = F^{-1}(\alpha) = -\frac{1}{\lambda} \ln(1 - \alpha). \quad (8.3)$$

Dosažením $\lambda = \frac{1}{10\,000}$ a $\alpha = 0,01$ (neboť $\mathbb{P}(T > t) = 1 - F(t) = 1 - \alpha = 0,99$) dostáváme

$$t = -10\,000 \cdot \ln 0,99 \doteq 100,5.$$

S pravděpodobností 99 % bude notebook pracovat déle než 100,5 hodin. $\triangle$

Příklad 8.7. Do obchodu přijde průměrně 30 zákazníků za 1 hodinu. Jaká je pravděpodobnost, že do obchodu nepřijde žádný zákazník během $\frac{1}{2}$ minuty, ve které je prodavač nepřítomen?

Řešení. Úlohu vyřešíme pomocí exponenciálního i Poissonova rozdělení.

- a) Necht náhodná veličina T udává dobu čekání na příchod dalšího zákazníka v minutách. T se řídí exponenciálním rozdělením s parametrem intenzity $\lambda = \frac{1}{2}$ zákazníka za minutu (neboť střední doba čekání na jednoho zákazníka je $\mathbb{E}T = \frac{1}{\lambda} = 2$ minuty). Hledaná pravděpodobnost je pak podle vzorce (8.2)

$$\mathbb{P}\left(T > \frac{1}{2}\right) = e^{-\frac{1}{2} \cdot \frac{1}{2}} = e^{-\frac{1}{4}} \doteq 0,778\,8.$$

= `1-EXPON.DIST(0,5;0,5;PRAVDA)`

- b) Necht náhodná veličina X udává počet zákazníků, kteří přijdou do obchodu během $\frac{1}{2}$ minuty. X má Poissonovo rozdělení s parametrem $\lambda' = \frac{1}{4}$. Všimněte si, že se parametr rozdělení liší od předchozího postupu, protože se změnila časová jednotka – vidíme, že

$$\lambda' = \lambda t = \frac{1}{2} \cdot \frac{1}{2}.$$

Hledaná pravděpodobnost je pak

$$\mathbb{P}(X = 0) = \frac{1}{0!} \left(\frac{1}{4}\right)^0 e^{-\frac{1}{4}} = e^{-\frac{1}{4}} \doteq 0,778\,8.$$

= `POISSON.DIST(0;0,25;NEPRAVDA)`

Pravděpodobnost, že v době nepřítomnosti prodavače v délce půl minuty nepřijde žádný zákazník, je asi 78 %. $\triangle$

8.3 Erlangovo³ rozdělení a rozdělení gama

Uvažujme nyní náhodnou veličinu W_k , jež udává dobu do výskytu k -té události v Poissonově procesu s intenzitou $\lambda > 0$ událostí za zvolenou časovou jednotku. To znamená, že v časovém intervalu $(0; t)$ nastane $(k-1)$ událostí s pravděpodobností

$$\frac{1}{(k-1)!}(\lambda t)^{k-1}e^{-\lambda t},$$

která se řídí Poissonovým rozdělením, a poslední k -tá událost nastane v krátkém časovém okamžiku $(t; t + dt)$ s pravděpodobností λdt , která vyplývá z 2. axiomu Poissonova procesu (viz odstavec 7.8). Protože se oba časové úseky nepřekrývají, jsou dle 1. axiomu Poissonova procesu oba jevy nezávislé. Pravděpodobnost toho, že nastanou současně, je tedy jejich součinem,

$$\frac{1}{(k-1)!}(\lambda t)^{k-1}e^{-\lambda t}\lambda dt.$$

Limitním přechodem $dt \rightarrow 0$ dostáváme předpis hustoty rozdělení náhodné veličiny W_k , jež je ve tvaru

$$f_{W_k}(t) = \begin{cases} \frac{1}{(k-1)!}\lambda^k t^{k-1}e^{-\lambda t} & \text{pro } t > 0, \\ 0 & \text{jinak.} \end{cases} \quad (8.4)$$

Toto rozdělení se nazývá *Erlangovo rozdělení*. Budeme jej v tomto odstavci značit $\text{Erlang}(k, \lambda)$, kde parametr k je počet požadovaných událostí a parametr λ je intenzita výskytu těchto událostí. Lze dokázat, že Erlangovo rozdělení je rozdělením součtu k nezávislých náhodných veličin s exponenciálním rozdělením s identickým parametrem λ . Speciálně je tedy $\text{Exp}(\lambda) = \text{Erlang}(1, \lambda)$. Distribuční funkce tohoto rozdělení má tvar

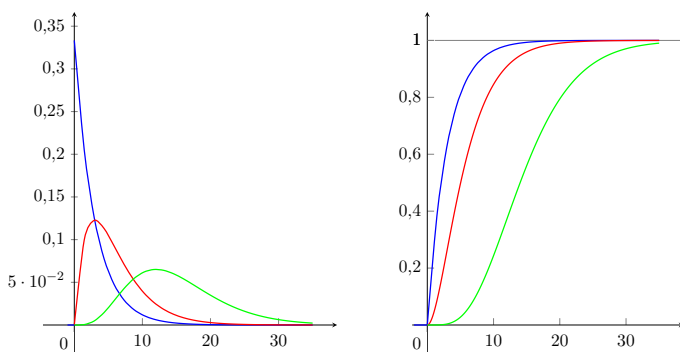
$$F(t) = \begin{cases} 1 - e^{-\lambda t} \cdot \sum_{j=0}^{k-1} \frac{(\lambda t)^j}{j!} & \text{pro } t > 0, \\ 0 & \text{jinak} \end{cases}$$

a střední hodnota a rozptyl jsou tvaru

$$\mathbb{E}W_k = \frac{k}{\lambda} \quad \text{a} \quad \text{var } W_k = \frac{k}{\lambda^2}.$$

Příklad 8.8. Vytvořte grafy hustot pravděpodobnosti a distribučních funkcí náhodné veličiny $X \sim \text{Erlang}(k, \frac{1}{3})$ a zkoumejte vliv hodnoty parametru k na průběh grafů. Za první parametr postupně zvolte:

³Agner Krarup Erlang (1878–1929), dánský matematik, statistik a inženýr.



Obrázek 8.4: Grafy hustot pravděpodobnosti a distribučních funkcí náhodné veličiny s Erlangovým rozdělením: $\text{Erlang}(1; \frac{1}{3})$ (modrá čára), $\text{Erlang}(2; \frac{1}{3})$ (červená čára), $\text{Erlang}(5; \frac{1}{3})$ (zelená čára)

- a) $k = 1$,
- b) $k = 2$,
- c) $k = 5$.

Řešení. Grafy hustot pravděpodobností a distribučních funkcí jsou na obrázku 8.4. △

Podíváme-li se pozorně na obrázek 8.4, vidíme, že parametr k určuje tvar hustoty, resp. distribuční funkce. Pokud opustíme koncept diskrétního počtu událostí Poissonova procesu, je možné pracovat (v jiných kontextech) s hustotami o parametru k , který nemusí být nutně celočíselný. Hustota takového rozdělení bude mít tvar podobný vzorci (8.4), ve kterém jen musíme nahradit faktoriál $(k-1)!$ ve jmenovateli nějakou jeho spojitou verzí. Označme tuto prozatím neznámou funkci $\Gamma(\alpha)$. Je-li argumentem α celé číslo, dejme tomu k , budeme vyžadovat, aby tato funkce měla shodné vlastnosti jako má faktoriál $(k-1)!$. Speciálně tedy vyžadujeme⁴

$$\begin{aligned}\Gamma(k) &= (k-1)!, \\ \Gamma(\alpha) &= (\alpha-1)\Gamma(\alpha-1), \\ \Gamma(1) &= 1.\end{aligned}$$

Jakmile takovou funkci nalezneme, potom hustota Erlangova rozdělení ve tvaru

$$f(t) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha t^{\alpha-1} e^{-\lambda t} \quad (8.5)$$

⁴Připomeňme, že pro faktoriál libovolného přirozeného čísla k platí rekursivní vztah $(k-1)! = (k-1)(k-2)!$ s počáteční podmínkou $0! = 1$.

bude dobře definována i pro neceločíselné hodnoty α . Má-li to ovšem být hustota, pak nutně musí splňovat rovnost

$$\int_0^{+\infty} \frac{1}{\Gamma(\alpha)} \lambda^\alpha t^{\alpha-1} e^{-\lambda t} dt = 1. \quad (8.6)$$

Speciálně (8.6) musí platit pro hodnotu $\lambda = 1$, odkud dostáváme předpis

$$\Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1} e^{-t} dt. \quad (8.7)$$

Pro obecnou hodnotu $\lambda > 0$ dostaneme jednoduchou substitucí

$$\int_0^{+\infty} t^{\alpha-1} e^{-\lambda t} dt = \frac{1}{\lambda^\alpha} \int_0^{+\infty} t^{\alpha-1} e^{-t} dt = \frac{\Gamma(\alpha)}{\lambda^\alpha},$$

tedy potřebná rovnost (8.6) platí i pro obecnou hodnotu $\lambda > 0$.

Funkce Γ definovaná předpisem (8.7) se nazývá *funkce gama*. Je obecně definována pro všechna komplexní čísla kromě nuly a celých záporných čísel a integrál v (8.7) konverguje, je-li $\alpha > 0$. Funkce gama se často také nazývá *Eulerův integrál druhého druhu* nebo *Eulerova funkce* a v teorii pravděpodobnosti slouží jako normalizační činitel ve vzorci (8.5), aby integrál z hustoty pravděpodobnosti zůstal jednotkový, jak bylo ukázáno výše. Rozdělení s hustotou ve tvaru

$$f(t) = \begin{cases} \frac{1}{\Gamma(\alpha)} \lambda^\alpha t^{\alpha-1} e^{-\lambda t} & \text{pro } t > 0, \\ 0 & \text{pro } t \leq 0, \end{cases}$$

se nazývá *rozdělení gama* se značením $\Gamma(\alpha; \lambda)$ ⁵, kde $\alpha > 0$ je parametr tvaru a $\lambda > 0$ parametr intenzity rozdělení. Distribuční funkce tohoto rozdělení je pro $x > 0$ dána integrálem⁶

$$F(x) = \int_0^x f(t) dt = \frac{1}{\Gamma(\alpha)} \int_0^x t^{\alpha-1} e^{-t} dt. \quad (8.8)$$

Střední hodnota a rozptyl náhodné veličiny X s rozdělením gama mají analogicky k Erlangovu rozdělení tvar

$$\mathbb{E}X = \frac{\alpha}{\lambda}, \quad \text{var } X = \frac{\alpha}{\lambda^2}.$$

Podobně jako exponenciální rozdělení lze i rozdělení gama parametrizovat pomocí parametru $\beta = 1/\lambda$ (viz poznámku pod čarou 2 na straně 146),

⁵Někdy se používá též značení Ga pro odlišení od značení Γ pro funkci gama.

⁶Integrál na pravé straně vzorce (8.8) se nazývá *dolní neúplná gama funkce*.

který se v případě rozdělení gama nazývá parametr měřítka. Tuto parametrizaci používá například software Microsoft Excel ve funkci `= GAMMA.DIST`, která slouží pro výpočet pravděpodobnosti Erlangova rozdělení, resp. rozdělení gama. Ukážeme to na následujících příkladech.

Příklad 8.9. Necht $X \sim \text{Erlang}(2; \frac{1}{3})$. Užitím vhodné funkce v MS Excel vypočítejte:

- $\mathbb{P}(X \leq 10)$,
- $\mathbb{P}(X \geq 20)$,
- $\mathbb{P}(10 < X < 20)$.

Řešení. K řešení využijeme funkci `= GAMMA.DIST`, do které dosadíme parametry $\alpha = k = 2$ a $\beta = 1/\lambda = 3$.

a) $\mathbb{P}(X \leq 10) = F(10) \doteq 0,845$,

`= GAMMA.DIST(10;2;3;PRAVDA)`

b) $\mathbb{P}(X \geq 20) = 1 - \mathbb{P}(X < 20) \doteq 0,009\,75$,

`= 1-GAMMA.DIST(20;2;3;PRAVDA)`

c) $\mathbb{P}(10 < X < 20) = F(20) - F(10) \doteq 0,145$.

`= GAMMA.DIST(20;2;3;PRAVDA)-GAMMA.DIST(10;2;3;PRAVDA)`

△

Příklad 8.10. Necht $X \sim \Gamma(2,5; \frac{1}{3})$. Užitím vhodné funkce v MS Excel vypočítejte:

- $\mathbb{P}(X \leq 10)$,
- $\mathbb{P}(X \geq 20)$,
- $\mathbb{P}(10 < X < 20)$.

Řešení. K řešení využijeme funkci `= GAMMA.DIST`, do které dosadíme parametry $\alpha = 2,5$ a $\beta = 1/\lambda = 3$.

a) $\mathbb{P}(X \leq 10) = F(10) \doteq 0,753$,

`= GAMMA.DIST(10;2,5;3;PRAVDA)`

b) $\mathbb{P}(X \geq 20) = 1 - \mathbb{P}(X < 20) = 1 - F(20) \doteq 0,02$,

`= 1-GAMMA.DIST(20;2,5;3;PRAVDA)`

c) $\mathbb{P}(10 < X < 20) = F(20) - F(10) \doteq 0,226$.

`= GAMMA.DIST(20;2,5;3;PRAVDA)-GAMMA.DIST(10;2,5;3;PRAVDA)`

△

Příklad 8.11. Střední životnost elektrického holicího strojku je dva roky. Jaká je pravděpodobnost, že průměrná životnost náhodně vybraných 150 strojků bude vyšší než 27 měsíců?

Řešení. Necht $X_1, X_2, \dots, X_{150}$ jsou náhodné veličiny sledující životnost 1., 2., ..., 150. holicího strojku. Každá tato náhodná veličina sleduje exponenciální rozdělení s identickým parametrem intenzity: $\lambda = \frac{1}{24}$ a $X_i \sim \text{Exp}\left(\frac{1}{24}\right)$, $i = 1, \dots, 150$, zvolíme-li za časovou jednotku 1 měsíc. Definujme náhodnou veličinu

$$Y = \sum_{i=1}^{150} X_i$$

za předpokladu, že životnosti jednotlivých holicích strojků jsou nezávislé náhodné veličiny, platí $Y \sim \text{Erlang}(150; \frac{1}{24})$.

Hledáme pravděpodobnost, že „průměrná životnost náhodně vybraných 150 strojků bude vyšší než 27 měsíců“, tj.

$$\mathbb{P}\left(\frac{1}{150}Y > 27\right) = \mathbb{P}(Y > 27 \cdot 150) = \mathbb{P}(Y > 4050).$$

K výpočtu $\mathbb{P}(Y > 4050)$ využijeme funkci

$$= 1 - \text{GAMMA.DIST}(4050; 150; 24; \text{PRAVDA})$$

$\mathbb{P}(Y > 4050) = 1 - \mathbb{P}(Y \leq 4050) \doteq 0,067$, tj. 6,7 %.

△

Erlangovo rozdělení se používá především v teorii spolehlivosti: jestliže exponenciální rozdělení modeluje dobu do poruchy komponenty, potom doba funkčnosti do k -té poruchy je popsána Erlangovým rozdělením. Do této kategorie spadá i předchozí příklad 8.11. Rozdělení gama s neceločíselným parametrem tvaru α se objevuje ve speciálních statistických metodách⁷ nebo jako model výše škody pojistné události. Speciálním případem rozdělení gama je také tzv. Pearsonovo χ^2 rozdělení, s kterým se podrobněji seznámíme v odstavci 8.7 na straně 166.

8.4 Obecné normální rozdělení

Obecné normální rozdělení, alternativně též *Gaussovo rozdělení* je definováno hustotou ve tvaru

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty, \quad (8.9)$$

⁷Například se objevuje při odhadování parametru λ Poissonova procesu pomocí bayesovské statistiky.

kde parametry rozdělení jsou reálné číslo μ reprezentující střední hodnotu rozdělení a kladné reálné číslo σ^2 reprezentující rozptyl rozdělení. Hustota pravděpodobnosti $f(x)$ má typický zvonovitý charakter, její graf je souměrný kolem osy vedené střední hodnotou μ kolmo k vodorovné souřadnicové ose, a v bodě μ též nabývá svého maxima, tedy μ je rovněž modem rozdělení.

Distribuční funkce má tvar

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^x e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt, \quad -\infty < x < \infty. \quad (8.10)$$

Distribuční funkce normálního rozdělení je jednou ze známých křivek, ke kterým neexistuje analytické vyjádření a je k dispozici jen integrální zápis (8.10). Střední hodnota a rozptyl mají tvary

$$\mathbb{E}X = \mu, \quad \text{var } X = \sigma^2.$$

Díky symetrii je střední hodnota μ též mediánem rozdělení.

Lze ukázat, že normální rozdělení je střední hodnotou μ a rozptylem σ^2 určeno jednoznačně. Je zvykem ho označovat jako $N(\mu; \sigma^2)$ ⁸. Speciální význam má tzv. normované (standardizované) normální rozdělení $N(0; 1)$, kterému věnujeme samostatný odstavec 8.5.

Příklad 8.12. Pro náhodnou veličinu $X \sim N(0; 2)$ vypočítejte pravděpodobnost, že náhodná veličina X bude:

- a) přesně rovna 0,
- b) menší nebo rovno -2 ,
- c) větší než 4,
- d) alespoň -1 a nejvýše 2.

Řešení. K výpočtu užijeme funkce `= NORM.DIST` programu MS Excel. Rozdělení parametrizuje pomocí směrodatné odchylky, proto dosazujeme jako třetí parametr hodnotu $\sigma = \sqrt{2}$.

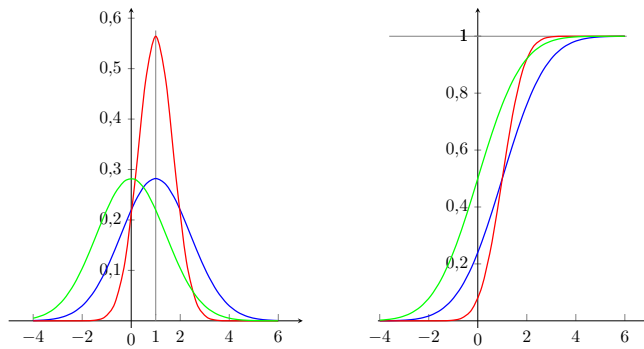
a) $\mathbb{P}(X = 0) = 0,$

b) $\mathbb{P}(X \leq -2) = \frac{1}{\sqrt{2\pi}\sqrt{2}} \int_{-\infty}^{-2} e^{-\frac{t^2}{4}} dt = 1 - F(-2) \doteq 0,078\,65,$

`= NORM.DIST(-2;0;ODMOCNINA(2);PRAVDA)`

c) $\mathbb{P}(X > 4) = \frac{1}{\sqrt{2\pi}\sqrt{2}} \int_4^{\infty} e^{-\frac{t^2}{4}} dt$
 $= 1 - \mathbb{P}(X \leq 4) = 1 - F(4) \doteq 0,002\,3,$

⁸Tato konvence opět není v literatuře zcela ustálená, někteří autoři používají jako druhý parametr směrodatnou odchylku σ a píší $N(\mu, \sigma)$. V této učebnici se však přidržíme parametrizace pomocí rozptylu σ^2 .



Obrázek 8.5: Grafy hustot pravděpodobnosti a distribučních funkcí náhodné veličiny s normálním rozdělením: $N(1; 2)$ (modrá čára), $N(1; \frac{1}{2})$ (červená čára), $N(0; 2)$ (zelená čára)

$$= 1 - \text{NORM.DIST}(4; 0; \text{ODMOCNINA}(2); \text{PRAVDA})$$

$$d) \mathbb{P}(-1 \leq X \leq 2) = \frac{1}{\sqrt{2\pi}\sqrt{2}} \int_{-1}^2 e^{-\frac{t^2}{4}} dt = F(2) - F(-1) \doteq 0,6816.$$

$$= \text{NORM.DIST}(2; 0; \text{ODMOCNINA}(2); \text{PRAVDA}) -$$

$$= -\text{NORM.DIST}(-1; 0; \text{ODMOCNINA}(2); \text{PRAVDA})$$

△

Příklad 8.13. Vytvořte grafy hustot pravděpodobnosti a grafy distribučních funkcí náhodných veličin, které mají rozdělení

- a) $N(1; 2)$,
- b) $N(1; \frac{1}{2})$,
- c) $N(0; 2)$.

Posuďte vliv hodnoty parametru σ^2 na průběh grafů.

Řešení. Grafy ukazuje obrázek 8.5. Parametr rozptylu σ^2 ovlivňuje „šířku hustoty“ (správněji její špičatost): hodnoty náhodné veličiny se vyskytují převážně v intervalu $(\mu - 3\sigma; \mu + 3\sigma)$, např. v případě a) leží v intervalu $(-5; 7)$. Podrobnější zdůvodnění najdete na straně 164 v odstavci 8.5 o standardním normálním rozdělení. △

Poznámka 8.1. Lze dokázat, že střední hodnota $\mathbb{E}X = \mu$, rozptyl $\text{var } X = \sigma^2$, že všechny liché centrální momenty (kromě prvního) jsou nulové, tj.

$$\mathbb{E}(X - \mathbb{E}X)^{2k-1} = 0, \quad k = 2, 3, \dots,$$

a že sudé centrální momenty jsou

$$\mathbb{E}(X - \mathbb{E}X)^{2k} = 1 \cdot 3 \cdot 5 \cdots (2k - 1)\sigma^{2k}, \quad k = 1, 2, \dots$$

Normální rozdělení má mimořádný význam v teorii pravděpodobnosti a matematické statistice. V odstavci 11.3 bude dokázáno, že součet velkého počtu náhodných veličin, o jejichž rozdělení se činí jen velmi obecné předpoklady, má přibližně normální rozdělení. Tím lze vysvětlit klíčovou roli tohoto rozdělení v teorii pravděpodobnosti a matematické statistice. Náhodné veličiny, s nimiž se v reálném světě setkáváme, lze velmi často považovat za výslednice působení velkého počtu drobných náhodných vlivů rozličných rozdělení. Díky výše uvedenému tvrzení lze očekávat, že normální rozdělení bude vhodným modelem pro takové náhodné veličiny. Nejběžnějším typem normálně rozdělených náhodných veličin jsou náhodné chyby, například chyby měření způsobené velkým počtem neznámých a vzájemně nezávislých příčin. Normální rozdělení je vhodným modelem také pro řadu fyzikálních, technických a biologických veličin, jako například tělesná výška jedinců homogenní populace, roční částka, kterou pojišťovna vyplátí za pojistné události atd.

Příklad 8.14. Ze statistik biologů je známo, že populace určitého druhu květin dorůstá výšky X cm s normálním rozdělením $N(20; 16)$. Vypočítejte pravděpodobnost, že náhodně vybraná květina má v centimetrech výšku

- menší než 16,
- větší než 20,
- v mezích od 12 do 28,
- menší než 12 nebo větší než 28,
- rovnu 22.

Řešení.

a) $\mathbb{P}(X < 16) = F(16) \doteq 0,159,$

$= \text{NORM.DIST}(16; 20; 4; \text{PRAVDA})$

b) $\mathbb{P}(X > 20) = 1 - \mathbb{P}(X \leq 20) = 1 - F(20) = 0,5.$

$= 1 - \text{NORM.DIST}(20; 20; 4; \text{PRAVDA})$

Všimněte si, že v tomto případě jsme nemuseli hodnotu pravděpodobnosti počítat, protože plyne z faktu, že 20 je mediánem rozdělení.

c) $\mathbb{P}(12 \leq X \leq 28) = F(28) - F(12) \doteq 0,955,$

$= \text{NORM.DIST}(28; 20; 4; \text{PRAVDA}) - \text{NORM.DIST}(12; 20; 4; \text{PRAVDA})$

d) $\mathbb{P}(X < 12) + \mathbb{P}(X > 28) = F(12) + 1 - F(28) \doteq 0,046,$

$= \text{NORM.DIST}(12; 20; 4; \text{PRAVDA}) + 1 - \text{NORM.DIST}(28; 20; 4; \text{PRAVDA})$

e) $\mathbb{P}(X = 22) = 0$.

△

Příklad 8.15. V bankovní regulaci se k ocenění rizikovosti investičního portfolia užívá míra, která se nazývá *hodnota v riziku*. Jde o výši ztráty portfolia, která bude překročena pouze s předepsanou malou mírou pravděpodobnosti, přičemž ztrátou portfolia obvykle rozumíme opačnou hodnotu k jeho výnosu.

Předpokládejme, že denní ztráta X nějakého portfolia (v procentních bodech) se při finanční krizi řídí normálním rozdělením se střední hodnotou 0,015 a rozptylem 0,01. Spočítejte 1% a 5% hodnotu v riziku tohoto portfolia.

Řešení. 1%, resp. 5% hodnota v riziku portfolia odpovídá z pravděpodobnostního pohledu 99%, resp. 95% kvantilu rozdělení ztráty $N(0,015; 0,01)$ (rizikem je pravý chvost⁹ rozdělení, který reprezentuje vysoké ztráty). K jeho výpočtu použijeme funkci `= NORM.INV`, do které dosadíme jako třetí parametr směrodatnou odchylku $\sigma = \sqrt{0,01} = 0,1$.

$$= \text{NORM.INV}(0,99; 0,015; 0,1), \quad = \text{NORM.INV}(0,95; 0,015; 0,1)$$

Hodnoty v riziku portfolia činí $F^{-1}(0,99) \doteq 0,248$, resp. $F^{-1}(0,95) \doteq 0,179$. Tyto hodnoty představují ztrátu portfolia v procentních bodech, jejíž překročení je jen málo pravděpodobné. △

8.5 Normované normální rozdělení

Tzv. *normované* (též standardizované nebo standardní) *normální rozdělení* je speciálním případem normálního rozdělení, pro které je $\mu = 0$ a $\sigma^2 = 1$. Je výsledkem procesu tzv. standardizace, matematicky popsané následující větou, kterou jinak uvádíme bez důkazu.

Věta 8.2 (Standardizace normálně rozdělené náhodné veličiny). *Nechť náhodná veličina $Z \sim N(0; 1)$ a*

$$X = \mu + \sigma Z.$$

Potom $X \sim N(\mu; \sigma^2)$. Obráceně, jestliže $X \sim N(\mu; \sigma^2)$, potom náhodná veličina

$$Z = \frac{X - \mu}{\sigma} \sim N(0; 1). \quad (8.11)$$

⁹Chvostem se nazývá část grafu hustoty pravděpodobnosti, která se ve srovnání se zbylou částí hustoty jen málo odlišuje od osy x .

Transformace (8.11) zjednodušuje formulaci mnoha vět v teorii pravděpodobnosti i statistice, neboť zjednodušeně říká, že normální rozdělení obecné a normované lze mezi sebou jednoduše převádět. Praktické využití transformace ukážeme na příkladu 8.16.

Příklad 8.16. Vraťte se k zadání příkladu 8.14, vypočítejte příslušné pravděpodobnosti pomocí transformace na normované normální rozdělení a porovnejte je s původními výsledky.

Řešení. Provedeme transformaci veličiny X na normovanou veličinu

$$Z = \frac{X - 20}{4}$$

a stejně upravíme i druhou stranu nerovnosti, které se říká z -skór. K výpočtu z -skóru lze využít funkci `= STANDARDIZE` v MS Excel, pokud známe hodnoty μ a σ nenormované náhodné veličiny. V našem příkladě je v otázce a) $\mu = 20$, $\sigma = 4$, a tedy

$$\mathbb{P}(X < 16) = \mathbb{P}\left(\frac{X - 20}{4} < \frac{16 - 20}{4}\right) = \mathbb{P}(Z < -1),$$

přičemž pravou stranu nerovnosti dostaneme jako výsledek

$$= \text{STANDARDIZE}(16; 20; 4)$$

Podobně spočítáme i zbývající případy. Distribuční funkci normovaného normálního rozdělení v následujících příkladech označujeme tradičním symbolem Φ .

$$\text{a) } \mathbb{P}(X < 16) = \mathbb{P}(Z < -1) = \Phi(-1) \doteq 0,1586,$$

$$= \text{NORM.S.DIST}(-1; \text{PRAVDA})$$

$$\text{b) } \mathbb{P}(X > 20) = \mathbb{P}\left(Z > \frac{20 - 20}{4}\right) = \mathbb{P}(Z > 0) = 1 - \Phi(0) = 0,5.$$

$$= \text{STANDARDIZE}(20; 20; 4)$$

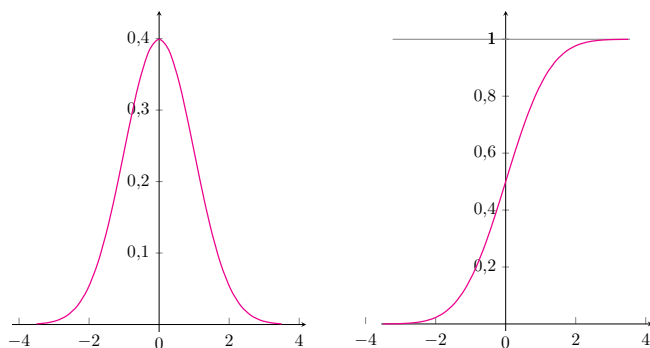
$$= 1 - \text{NORM.S.DIST}(0; \text{PRAVDA})^{10}$$

V tomto případě není ovšem potřeba hodnotu $\Phi(0)$ nechat počítat v jakémkoliv softwaru: můžeme vyjít z toho, že 0 je zároveň medián rozdělení $N(0; 1)$, odkud přímo plyne odpověď.

$$\begin{aligned} \text{c) } \mathbb{P}(12 \leq X \leq 28) &= \mathbb{P}\left(\frac{12 - 20}{4} \leq Z \leq \frac{28 - 20}{4}\right) = \\ &= \mathbb{P}(-2 \leq Z \leq 2) = \Phi(2) - \Phi(-2) \doteq 0,9545. \end{aligned}$$

¹⁰Funkci `= STANDARDIZE` lze vložit za první parametr funkce `= NORM.S.DIST`

a dostaneme `= NORM.S.DIST(STANDARDIZE(20; 20; 4), PRAVDA)`.



Obrázek 8.6: Grafy hustoty a distribuční funkce náhodné veličiny s normovaným normálním rozdělením

```

1 STANDARDIZE(12;20;4) STANDARDIZE(28;20;4)
2 NORM.S.DIST(2;PRAVDA)-NORM.S.DIST(-2;PRAVDA)

```

I v tomto případě lze výpočet zjednodušit využitím symetrie rozdělení. Platí totiž $\Phi(-2) = 1 - \Phi(2)$ (slovy, levý chvost rozdělení do -2 je stejné velikosti jako pravý chvost od 2), tudíž $\Phi(2) - \Phi(-2) = 2\Phi(2) - 1$. Z výše uvedených hodnot distribučních funkcí tedy fakticky stačí spočítat jen jednu.

$$= 2 * \text{NORM.S.DIST}(2; \text{PRAVDA}) - 1$$

- d) $\mathbb{P}(X < 12) + \mathbb{P}(X > 28) = 1 - \mathbb{P}(12 \leq X \leq 28) \doteq 0,045\,50$,
kde využijeme již spočteného výsledku úlohy c).

△

Dosadíme-li parametry $\mu = 0$ a $\sigma^2 = 1$ do (8.9), dostaneme předpis pro hustotu normovaného normálního rozdělení ve tvaru

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty.$$

Na obrázku 8.6 vlevo je graf hustoty pravděpodobnosti $\varphi(z)$, který má tvar zvonovité křivky, nazývané Gaussova¹¹ křivka, která je osově souměrná podle osy procházející číslem 0 kolmo k vodorovné ose a ve stejném čísle také dosahuje svého maxima $\frac{1}{\sqrt{2\pi}} \doteq 0,398\,94$.

Distribuční funkce normovaného normálního rozdělení opět nemá analytické vyjádření a zapisujeme ji pouze v integrálním tvaru

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{t^2}{2}} dt, \quad -\infty < z < \infty.$$

¹¹ Johann Carl Friedrich Gauss (1777–1855), německý matematik a fyzik.

Graf této distribuční funkce je na obrázku 8.6 vpravo. I distribuční funkce obecného a normálního rozdělení lze vzájemně převádět: důsledkem věty 8.2 je vztah

$$F_X(x) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

pro $X \sim N(\mu, \sigma^2)$.

Pro značení normovaného normálního rozdělení se někdy užívá symbolu U (bez uvedení parametrů) a pro označení příslušné náhodné veličiny se volí písmeno U (píše se pak $U \sim U$). Toto označení je třeba odlišovat od značení rovnoměrného rozdělení. My se tedy přidržíme označení $N(0; 1)$ a budeme psát $Z \sim N(0; 1)$. Poznamenejme ještě, že značení hustoty φ a distribuční funkce Φ normovaného normálního rozdělení (pro odlišení od ostatních rozdělení) je rovněž poměrně ustálenou zvyklostí, která podtrhuje význam tohoto rozdělení v teorii pravděpodobnosti.

Ověrme nejprve, že $\varphi(z)$ je hustota. Při výpočtech budeme užívat tzv. Gaussova integrálu (integrálu Gaussovy funkce)

$$\int_{-\infty}^{\infty} e^{-y^2} dy = \sqrt{\pi}. \quad (8.12)$$

Vidíme nejprve, že $\varphi(z) \geq 0$ pro libovolné $z \in \mathbb{R}$. Ověřme dále

$$\int_{-\infty}^{\infty} \varphi(z) dz = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sqrt{2} e^{-y^2} dy = 1,$$

a tedy $\varphi(z)$ je hustota. Střední hodnota a všechny liché momenty jsou rovný nule, tj.

$$\mathbb{E}Z = \mathbb{E}Z^{2k-1} = 0, \quad k = 1, 2, 3, \dots,$$

neboť funkce

$$\frac{1}{\sqrt{2\pi}} z^{2k-1} e^{-\frac{z^2}{2}}$$

je lichá a integrovatelná v $\mathbb{R}$. Níže ukážeme, že sudé momenty jsou rovny

$$\mathbb{E}Z^{2k} = 1 \cdot 3 \cdot 5 \cdots (2k-1) = \frac{(2k)!}{2^k k!}, \quad k = 1, 2, 3, \dots \quad (8.13)$$

Jelikož $\mathbb{E}Z = 0$, dosazením hodnoty $k = 1$ odtud plyne

$$\text{var } Z = \mathbb{E}Z^2 = 1.$$

Všimněte si též čtvrtého absolutního momentu $\mathbb{E}Z^4$, tedy špičatosti rozdělení¹², jež je konstantní hodnoty 3. Provedeme-li v (8.12) substituci $y = z\sqrt{\frac{t}{2}}$, dostáváme

$$\int_{-\infty}^{\infty} e^{-\frac{tz^2}{2}} dz = \sqrt{\frac{2\pi}{t}}. \quad (8.14)$$

Derivujme levou i pravou stranu (8.14) podle t . Dostáváme

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-\frac{tz^2}{2}} dz = t^{-\frac{3}{2}}.$$

Obecně po k opakování derivace dostáváme

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^{2k} e^{-\frac{tz^2}{2}} dz = 1 \cdot 3 \cdots (2k-1) \cdot t^{-\frac{2k+1}{2}}. \quad (8.15)$$

Výsledek (8.13) pak plyne z (8.15) položením $t = 1$.

Hodnoty hustoty pravděpodobnosti $\varphi(z)$ normovaného normálního rozdělení pro zadanou hodnotu z lze v MS Excel určit pomocí funkce `= PHI`, která je ekvivalentem funkce `= NORM.S.DIST(z;NEPRAVDA)`. Například na straně 161 jsme určili maximální hodnotu normovaného normálního rozdělení, kterou hustota nabývá v čísle 0: zde `= PHI(0)`, tedy $\varphi(0) \doteq 0,398\,942$.

Příklad 8.17. Necht náhodná veličina Z má normované normální rozdělení. Vypočítejte pravděpodobnost, že náhodná veličina Z bude

- menší než -1 ,
- větší než 2 ,
- ležet mezi -1 a 1 .

Řešení.

$$\text{a) } \mathbb{P}(Z < -1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-1} e^{-\frac{t^2}{2}} dt = \Phi(-1) \doteq 0,159,$$

$$= \text{NORM.S.DIST}(-1; \text{PRAVDA})$$

$$\text{b) } \mathbb{P}(X > 2) = \frac{1}{\sqrt{2\pi}} \int_2^{\infty} e^{-\frac{t^2}{2}} dt = 1 - \Phi(2) \doteq 0,023,$$

$$= 1 - \text{NORM.S.DIST}(2; \text{PRAVDA})$$

$$\text{c) } \mathbb{P}(-1 < X < 1) = \frac{1}{\sqrt{2\pi}} \int_{-1}^1 e^{-\frac{t^2}{2}} dt = \Phi(1) - \Phi(-1) \doteq 0,683.$$

$$= \text{NORM.S.DIST}(1; \text{PRAVDA}) - \text{NORM.S.DIST}(-1; \text{PRAVDA})$$

¹²Viz definici 6.5 na straně 105.



Pravděpodobností normovaného normálního rozdělení se též užívá pro interpretaci směrodatné odchylky. Pro normované normální rozdělení totiž platí

- v intervalu $(-1; 1)$ leží 68,3 % populace¹³,
- v intervalu $(-1,96; 1,96)$ leží 95 % populace,
- v intervalu $(-2; 2)$ leží 95,5 % populace,
- v intervalu $(-3; 3)$ leží 99,7 % populace,
- v intervalu $(-3,29; 3,29)$ leží 99,9 % populace.

Řečeno v termínech kvantilové funkce,

$$\Phi^{-1}\left(\frac{1}{2} + \frac{0,683}{2}\right) = \Phi^{-1}(0,841) \doteq 1$$

a podobně například $\Phi^{-1}(0,975) \doteq 1,96$ či $\Phi^{-1}(0,9987) \doteq 3$. Vezmeme-li nyní větu 8.2, dostaneme analogické tvrzení pro obecnou normálně rozdělenou náhodnou veličinu:

- 68,3 % populace leží v intervalu $(\mu - \sigma; \mu + \sigma)$,
- 95 % populace leží v intervalu $(\mu - 1,96\sigma; \mu + 1,96\sigma)$,
- 95,5 % populace leží v intervalu $(\mu - 2\sigma; \mu + 2\sigma)$,
- 99,7 % populace leží v intervalu $(\mu - 3\sigma; \mu + 3\sigma)$,
- 99,9 % populace leží v intervalu $(\mu - 3,29\sigma; \mu + 3,29\sigma)$.

8.6 Logaritmicko-normální rozdělení

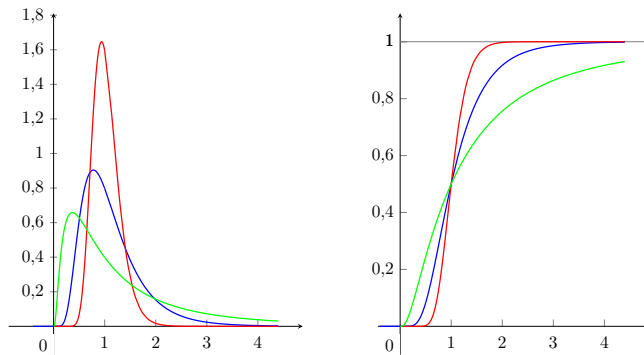
Náhodná veličina X má *logaritmicko-normální* (neformálně též lognormální) *rozdělení* $\text{LN}(\mu; \sigma^2)$ s parametry μ a σ^2 , pokud náhodná veličina $\ln X$ má normální rozdělení $\text{N}(\mu; \sigma^2)$. Hustota pravděpodobnosti a distribuční funkce logaritmicko-normálního rozdělení jsou ve tvaru

$$f(x) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(\ln x - \mu)^2}{\sigma^2}} & x \in (0, \infty), \\ 0 & \text{jinak,} \end{cases} \quad F(x) = \begin{cases} 0 & \text{pro } x \leq 0, \\ \Phi\left(\frac{\ln x - \mu}{\sigma}\right) & \text{pro } x > 0. \end{cases}$$

Hustota logaritmicko-normálního rozdělení je asymetrická, doleva zešikmená křivka. Střední hodnota a rozptyl $X \sim \text{LN}(\mu, \sigma^2)$ jsou rovny

$$\mathbb{E}X = e^{\mu + \frac{\sigma^2}{2}}, \quad \text{var } X = e^{2\mu + \sigma^2} (e^{\sigma^2} - 1).$$

¹³Myslíme tím, že pro $Z \sim \text{N}(0; 1)$ platí $\mathbb{P}(-1 \leq Z \leq 1) = 0,683$.



Obrázek 8.7: Grafy hustot a distribučních funkcí náhodné veličiny s logaritmicko-normálním rozdělením $\text{LN}(0; 0,25)$ (modrá čára), $\text{LN}(0; 0,0625)$ (červená čára), $\text{LN}(0; 1)$ (zelená čára)

Příklad 8.18. Vytvořte graf hustoty pravděpodobnosti a distribuční funkce náhodné veličiny s logaritmicko-normálním rozdělením

- $\text{LN}(0; 0,25)$,
- $\text{LN}(0; 0,0625)$,
- $\text{LN}(0; 1)$.

Vypočítejte hodnoty 17% kvantilu pro všechna zmíněná rozdělení.

Řešení. Grafy hustot pravděpodobnosti a distribučních funkcí ukazuje obrázek 8.7. Hodnoty kvantilu spočteme takto:

a) $F^{-1}(0,17) \doteq 0,6205$,

$$= \text{LOGNORM.INV}(0,17; 0; 0,5)$$

b) $F^{-1}(0,17) \doteq 0,7878$,

$$= \text{LOGNORM.INV}(0,17; 0; 0,25)$$

c) $F^{-1}(0,17) \doteq 0,3851$.

$$= \text{LOGNORM.INV}(0,17; 0; 1)$$

△

Logaritmicko-normální rozdělení má využití u náhodných veličin, které nabývají pouze kladných hodnot, kde je typické velké množství nízkých hodnot s občasným výskytem hodnoty vyšší. Takové náhodné veličiny se objevují například v ekonomii, pojišťovnictví či v teorii spolehlivosti. Uvedeme příklad využití logaritmicko-normálního rozdělení v pojišťovnictví pro výpočet pravděpodobnosti, že bude tzv. uplatněna pojistná událost: jestliže

je škoda při pojistné události větší než hodnota r , která představuje navýšení pojistného v následujícím pojistném období za předpokladu uplatnění pojistné události, má klient motivaci škodu uplatnit. Klient je na počátku po uzavření smlouvy s pojišťovnou zařazený do nulté bonusové třídy, kde platí pojistné ve výši 100 % ceníkové ceny. Zavedme následující náhodné veličiny: necht Y značí počet pojistných událostí, které se řídí Poissonovým rozdělením s parametrem λ : $Y \sim \text{Po}(\lambda)$. Dále necht Z značí výši škody, o níž předpokládáme, že se řídí logaritmicko-normálním rozdělením s parametry μ a σ^2 , tedy $Z \sim \text{LN}(\mu; \sigma^2)^{14}$. Konečně necht A označuje jev „klient uplatní pojistnou událost“. Pak pravděpodobnost uplatnění pojistné události je dána vztahem

$$\mathbb{P}(A) = \mathbb{P}(Y > 0) \cdot \mathbb{P}(Z > r) = (1 - e^{-\lambda}) \left[1 - \Phi\left(\frac{\ln r - \mu}{\sigma}\right) \right]$$

za předpokladu, že jevy výskyt škodní události ($Y > 0$) a překročení výše škody ($Z > r$) jsou vzájemně nezávislé.

Příklad 8.19. Klient pojišťovny je v první bonusové třídě, kde platí 60 % pojistného na sjednané autopojištění, které činí 2 500 Kč. Parametr λ Poissonova rozdělení počtu pojistných událostí je roven 0,15 a výše škody se řídí logaritmicko-normálním rozdělením s parametry $\mu = 6$ a $\sigma^2 = 2$. Určete pravděpodobnost, s jakou klient uplatní pojistnou událost.

Řešení. Klient uplatní pojistnou událost, jestliže bude škoda větší než navýšení pojistného r , zde $r = 2\,500 - 0,6 \cdot 2\,500 = 1\,000$ Kč. Pak

$$\mathbb{P}(A) = (1 - e^{-0,15}) \left[1 - \Phi\left(\frac{\ln 1\,000 - 6}{\sqrt{2}}\right) \right] \doteq 0,036.$$

```
= (1-POISSON.DIST(0;0,15;PRAVDA))*
```

```
= *(1-LOGNORM.DIST(1000;6;ODMOCNINA(2);PRAVDA))
```

Klient uplatní pojistnou událost s pravděpodobností asi 3,6 %. $\triangle$

8.7 Pearsonovo rozdělení χ^2

Z normovaného normálního rozdělení, které jsme představili v odstavci 8.5, se odvozuje trojice rozdělení (Pearsonovo, Studentovo a Fisherovo–Snedecorovo), která nalézají přímé uplatnění ve statistice, jak uvidíme v druhé části této knihy.

¹⁴Pro tuto oblast je právě typické velké množství událostí s relativně nízkou škodou, ale občas nastane i událost se škodou značné velikosti.

Nechť náhodné veličiny $Z_1, Z_2, \dots, Z_k$ jsou nezávislé¹⁵ a mají normované normální rozdělení $N(0; 1)$. Pak veličina

$$Q_k = \sum_{i=1}^k Z_i^2 \quad (8.16)$$

má tzv. *Pearsonovo*¹⁶ *rozdělení* χ^2 (čtete chí kvadrát) s k stupni volnosti. Počet stupňů volnosti je parametr tohoto rozdělení a je shodný s počtem sčítanců v rovnici (8.16). Rozdělení můžeme značit $\chi^2(k)$, ale běžnější je značení pomocí dolního indexu, tedy χ_k^2 , kterého se přidržíme i v této publikaci.

Lze ukázat, že Pearsonovo rozdělení χ_k^2 je speciálním případem rozdělení gama (viz odstavec 8.3 na straně 151) s parametry $\alpha = \frac{k}{2}$ a $\lambda = \frac{1}{2}$. Odtud dostáváme předpis pro hustotu pravděpodobnosti

$$f(z) = \frac{1}{\Gamma(\frac{k}{2}) \cdot 2^{\frac{k}{2}}} \cdot z^{\frac{k}{2}-1} \cdot e^{-\frac{z}{2}}, \quad z > 0,$$

pro distribuční funkci

$$F(x) = \frac{1}{\Gamma(\frac{k}{2})} \int_0^x t^{\frac{k}{2}-1} e^{-t} dt$$

a pro střední hodnotu a rozptyl

$$\mathbb{E}\chi_k^2 = k, \quad \text{var } \chi_k^2 = 2k.$$

Příklad 8.20. Nechť náhodná veličina X má rozdělení χ_5^2 . Vypočítejte pravděpodobnost, že náhodná veličina X :

- bude menší nebo rovna 5,
- nebude menší nebo maximálně rovna 10,
- bude svou hodnotou mezi 5 a 10.

Řešení.

$$\text{a) } \mathbb{P}(X \leq 5) = F(5) \doteq 0,584,$$

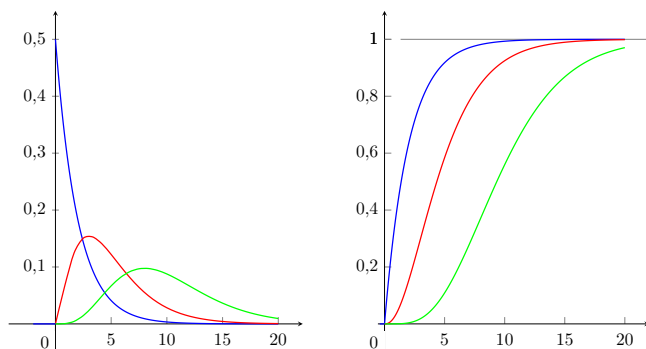
$$= \text{CHISQ.DIST}(5; 5; \text{PRAVDA})$$

$$\text{b) } \mathbb{P}(X > 10) = 1 - \mathbb{P}(X \leq 10) = 1 - F(10) \doteq 0,07,$$

$$= 1 - \text{CHISQ.DIST}(10; 5; \text{PRAVDA})$$

¹⁵Blíže k nezávislosti náhodných veličin viz odstavec 9.4 na straně 191.

¹⁶Karl Pearson (1857–1936), britský matematik.



Obrázek 8.8: Grafy hustot pravděpodobnosti a distribučních funkcí náhodné veličiny s Pearsonovým rozdělením: χ_2^2 (modrá čára), χ_5^2 (červená čára), χ_{10}^2 (zelená čára)

c) $\mathbb{P}(5 < X < 10) = F(10) - F(5) \doteq 0,341$.

= CHISQ.DIST(10;5;PRAVDA)-CHISQ.DIST(5;5;PRAVDA)

△

Příklad 8.21. Vytvořte grafy hustoty pravděpodobnosti a distribuční funkce Pearsonova rozdělení χ_k^2 pro hodnoty parametru $k = 2$, $k = 5$ a $k = 10$.

Řešení. Obrázek 8.8 ukazuje hustoty pravděpodobnosti a distribuční funkce Pearsonova rozdělení χ_k^2 pro $k = 2$, $k = 5$ a $k = 10$. Všimněte si, že pro nízké hodnoty k je hustota pravděpodobnosti výrazně asymetrická. Pro vyšší hodnoty k se hustota začne přibližovat hustotě normálního rozdělení $N(k; 2k)$. △

Pro výpočty kvantilů, které se tradičně označují $\chi_k^2(p)$, se dříve používaly tabulky (většinou pro $k = 1, \dots, 100$). Dnes se využívá software, například v MS Excel je to funkce = CHISQ.INV. Pearsonovo rozdělení χ^2 nachází uplatnění především ve statistice, například v intervalových odhadech rozptylu (viz odstavec 14.3.3), v testech o rozptylu (odstavec 15.3) či v testech dobré shody (kapitola 20).

8.8 Studentovo¹⁷ t -rozdělení

Mějme dvě nezávislé¹⁸ náhodné veličiny, a to náhodnou veličinu Z s normovaným normálním rozdělením $N(0; 1)$ a náhodnou veličinu Q s rozdělením χ^2 s k stupni volnosti. Lze dokázat, že náhodná veličina

$$T = \frac{Z}{\sqrt{Q}} \cdot \sqrt{k} = \frac{Z}{\sqrt{Q/k}}$$

má rozdělení, jehož hustotu lze zapsat ve tvaru

$$f(t) = \frac{1}{B(\frac{1}{2}, \frac{k}{2}) \cdot \sqrt{k}} \cdot \left(1 + \frac{t^2}{k}\right)^{-\frac{(k+1)}{2}}, \quad t \in \mathbb{R}, \quad (8.17)$$

kde B je tzv. *funkce beta*, nazývaná též *Eulerova funkce* nebo *Eulerův integrál prvního druhu*,

$$B(\alpha, \beta) = \frac{\Gamma(\alpha) \cdot \Gamma(\beta)}{\Gamma(\alpha + \beta)} = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx = \int_0^\infty \frac{x^{\alpha-1}}{(1+x)^{\alpha+\beta}}, \quad (8.18)$$

kde $t > 0$, $k > 0$ a Γ je funkce gama zavedená v odstavci 8.3 výrazem (8.7). Rozdělení s hustotou (8.17) se nazývá *Studentovo rozdělení* s k stupni volnosti a značíme jej t_k ¹⁹. Náhodná veličina $T \sim t_k$ má pro $k > 1$ střední hodnotu a pro $k > 2$ i rozptyl ve tvaru

$$\mathbb{E}T = 0, \quad \text{var } T = \frac{k}{k-2}.$$

Hustota Studentova rozdělení je symetrická podle svislé souřadné osy, tudíž pro kvantily, jež se typicky značí $t_k(p)$, platí

$$t_k(p) = -t_k(1-p).$$

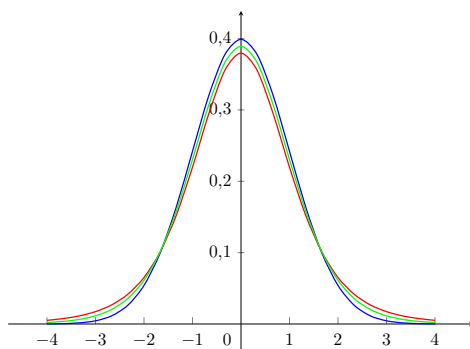
Tohoto vztahu se užívá především pro zjednodušení vzorců ve statistice, jak uvidíme například v odstavci 14.3.1.

Studentovo rozdělení s nízkým počtem stupňů volnosti k má uplatnění ve statistice (viz např. odstavec 15.2) jako náhrada normálního rozdělení v případech, kdy je pravděpodobnost výskytu krajních jevů vyšší (patří mezi

¹⁷William Sealy Gosset (1876–1937), anglický chemik a průkopník statistiky, který vyvinul t -rozdělení a publikoval pod pseudonymem *Student*.

¹⁸Blíže k nezávislosti náhodných veličin viz odstavec 9.4 na straně 191.

¹⁹V literatuře též $t(k)$.



Obrázek 8.9: Aproximace hustoty normovaného normálního rozdělení pomocí Studentova rozdělení: $N(0; 1)$ (modrá čára), t_5 (červená čára), t_{10} (zelená čára)

rozdělení s tzv. těžkými chvosty). Pro větší hodnoty k se blíží normovanému normálnímu rozdělení. Jestliže $T_k \sim t_k$, lze dokonce dokázat, že

$$F_{T_k} \xrightarrow[k \rightarrow \infty]{} \Phi,$$

a podobný vztah platí i pro hustoty rozdělení. Budeme to ilustrovat na příkladu 8.22.

Příklad 8.22. Na grafech hustot pravděpodobností ukažte blízkost normovaného normálního a Studentova rozdělení. Volte t_5 a t_{10} . Určete pravděpodobnost, pro kterou nabude náhodná veličina $T \sim t_5$ hodnoty

- menší než -1 ,
- větší nebo rovno 2 ,
- mezi -1 a 1 .

Řešení. Grafy hustot náhodných veličin s normovaným normálním a Studentovým rozdělením o pěti a deseti stupních volnosti ukazuje obrázek 8.9.

a) $\mathbb{P}(T < -1) = F(-1) \doteq 0,1816,$

$$= \text{T.DIST}(-1; 5; \text{PRAVDA})$$

b) $\mathbb{P}(T \geq 2) = 1 - F(2) \doteq 0,051,$

$$= 1 - \text{T.DIST}(2; 5; \text{PRAVDA})$$

c) $\mathbb{P}(-1 < T < 1) = F(1) - F(-1) \doteq 0,6368.$

$$= \text{T.DIST}(1; 5; \text{PRAVDA}) - \text{T.DIST}(-1; 5; \text{PRAVDA})$$

△

Příklad 8.23. Náhodná veličina T má Studentovo rozdělení s deseti stupni volnosti. Vypočítejte:

- a) 1% kvantil,
- b) 5% kvantil,
- c) 95% kvantil.

Řešení. a) $t_{10}(0,01) = F_T^{-1}(0,01) \doteq -2,76$,

$$= \text{T.INV}(0,01;10)$$

b) $t_{10}(0,05) = F_T^{-1}(0,05) \doteq -1,81$,

$$= \text{T.INV}(0,05;10)$$

c) $t_{10}(0,95) = F_T^{-1}(0,95) \doteq 1,81$.

$$= \text{T.INV}(0,95;10)$$

△

Speciálním případem Studentova rozdělení je Studentovo rozdělení o jednom stupni volnosti, které se nazývá Cauchyho rozdělení a používá se například ve financích pro modelování extrémních investičních rizik. Jeho distribuční funkci a hustotu pravděpodobnosti jsme studovali v úloze 5.5 a v příkladu 6.3. Jak jsme viděli v příkladu 6.3, Cauchyho rozdělení nemá střední hodnotu ani rozptyl (příslušné integrály divergují).

8.9 Fisherovo-Snedecorovo²⁰ rozdělení

Nechť dvě nezávislé²¹ náhodné veličiny mají rozdělení χ^2 , a to Q_1 s ν_1 stupni volnosti a Q_2 s ν_2 stupni volnosti. Lze dokázat, že náhodná veličina

$$F = \frac{Q_1/\nu_1}{Q_2/\nu_2} \quad (8.19)$$

má rozdělení, jehož hustotu lze zapsat ve tvaru

$$f(x) = \frac{1}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \cdot \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} \cdot \frac{x^{\frac{\nu_1-2}{2}}}{\left(1 + x \cdot \frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1+\nu_2}{2}}}, \quad x > 0, \quad (8.20)$$

kde $B(\cdot, \cdot)$ je funkce beta zavedená předpisem (8.18) na straně 169. Rozdělení s hustotou (8.20) se nazývá *Fisherovo–Snedecorovo rozdělení*, stručně též

²⁰Ronald Aylmer Fisher (1890–1962), britský matematik, statistik, biolog a genetik.
George Waddel Snedecor (1881–1974), americký matematik a statistik.

²¹Blíže k nezávislosti náhodných veličin viz odstavec 9.4 na straně 191.

rozdělení F , označené $F(\nu_1; \nu_2)$. ν_1 a ν_2 jsou parametry rozdělení, které nazýváme stupně volnosti (v uvedeném pořadí). Střední hodnota (pro $n > 2$) a rozptyl (pro $n > 4$) mají tvar

$$\mathbb{E}F = \frac{\nu_2}{\nu_2 - 2}, \quad \text{var } F = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}.$$

Kvantily rozdělení značíme tradičně $F_F^{-1}(p) = F_{\nu_1, \nu_2}(p)$. Fisherovo–Snedecorovo rozdělení nachází uplatnění v monoha testech statistických hypotéz, například v analýze rozptylu (viz odstavec 17.1) či v regresní analýze (viz odstavec 19.2).

Příklad 8.24. Necht náhodná veličina F má Fisherovo–Snedecorovo rozdělení s parametry $\nu_1 = 5$ a $\nu_2 = 10$. Vypočítejte pravděpodobnost, že náhodná veličina F bude mít hodnotu

- a) menší nebo rovnu 1,
- b) větší než 3,
- c) mezi 1 a 2.

Řešení.

a) $\mathbb{P}(F \leq 1) = F(1) \doteq 0,535,$

= F.DIST(1;5;10;PRAVDA)

b) $\mathbb{P}(F > 3) = 1 - F(3) \doteq 0,066,$

= 1-F.DIST(3;5;10;PRAVDA)

c) $\mathbb{P}(1 < F < 2) = F(2) - F(1) \doteq 0,301.$

= F.DIST(2;5;10;PRAVDA) - F.DIST(1;5;10;PRAVDA)

△

Příklad 8.25. Náhodná veličina F má rozdělení F se 7 a 9 stupni volnosti. Vypočítejte hodnotu

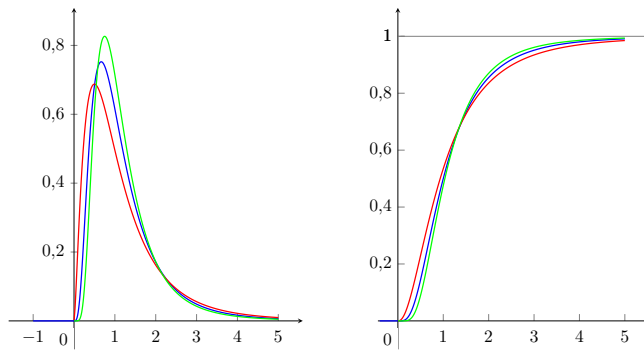
- a) 0,1% kvantilu,
- b) 1% kvantilu,
- c) 5% kvantilu,
- d) 99% kvantilu.

Řešení. a) $F_F^{-1}(0,001) = F_{7,9}(0,001) \doteq 0,0698,$

= F.INV(0,001;7;9)

b) $F_F^{-1}(0,01) = F_{7,9}(0,01) \doteq 0,1488,$

= F.INV(0,01;7;9)



Obrázek 8.10: Grafy hustot pravděpodobnosti a distribučních funkcí náhodné veličiny s Fisherovým–Snedecorovým rozdělením: $F(10; 10)$ (modrá čára), $F(5; 10)$ (červená čára), $F(20; 10)$ (zelená čára)

c) $F_F^{-1}(0,05) = F_{7,9}(0,05) \doteq 0,272\,0,$

= F.INV(0,05;7;9)

d) $F_F^{-1}(0,99) = F_{7,9}(0,99) \doteq 5,612\,9.$

= F.INV(0,99;7;9)

△

Příklad 8.26. Vytvořte grafy hustot pravděpodobnosti a grafy distribučních funkcí náhodných veličin s rozdělením

- a) $F(5; 10),$
- b) $F(10; 10),$
- c) $F(20; 10).$

Řešení. Grafy jsou na obrázku 8.10.

△

8.10 Úlohy

8.1 Ověřte obecnou platnost vzorce pro výpočet střední hodnoty a rozptylu náhodné veličiny X , která má rovnoměrné spojitě rozdělení.

8.2 Dokažte správnost vzorce distribuční funkce rovnoměrného spojitěho rozdělení.

8.3 Červené světlo svítí na semaforu 90 sekund. Cyklista, který přijede k semaforu a neví, jak dlouho již červená svítí, chce určit pravděpodobnost, že bude čekat

- a) méně než deset sekund,
- b) více jak šedesát sekund,
- c) mezi 20 a 40 sekundami.

8.4 Karel jezdí tramvají do práce, která přijíždí na zastávku každých 8 minut. Jaká je pravděpodobnost, že když přijde Karel v libovolný okamžik na zastávku tramvaje,

- a) nebude čekat více jak 3 minuty,
- b) bude čekat více než 8 minut,
- c) nebude čekat a přímo nastoupí do tramvaje v okamžiku svého příchodu,
- d) bude čekat 4–7 minut?



Načrtněte graf hustoty pravděpodobnosti a distribuční funkce v úloze uvažované náhodné veličiny.

8.5 Životnost určitého výrobku se řídí exponenciálním rozdělením se střední hodnotou 3 roky. Jak dlouhou záruční dobu může poskytnout výrobce zákazníkům, jestliže žádá, aby relativní četnost výrobků, které během záruční doby přestanou plnit svou funkci, byla asi 10 %?

8.6 Prodejna očekává dodávku zboží v určitý den v době od 12 do 16 hodin. Podle sdělení dodavatele je uskutečnění dodávky stejně možné kdykoliv během tohoto časového intervalu. Jaká je pravděpodobnost, že zboží bude dodáno v době od jedné hodiny do půl druhé?

8.7 Vytvořte graf hustoty pravděpodobnosti a distribuční funkce náhodné veličiny T s exponenciálním rozdělením $\text{Exp}(\frac{1}{3})$. Určete a na grafech vyznačte:

- a) $\mathbb{P}(T \leq 1)$,
- b) $\mathbb{P}(T = 1,4)$,
- c) $\mathbb{P}(T > 5)$,
- d) $\mathbb{P}(2 \leq T \leq 4)$.

8.8 Stanovte střední dobu obsluhy v prodejně, víte-li, že pravděpodobnost obslužení v době kratší než 4 minuty je 0,25. Předpokládejte, že doba obsluhy má exponenciální rozdělení.

8.9 Ukradený automobil se za dobu t najde s pravděpodobností

$$p(t) = 1 - e^{-\gamma t}, \quad \gamma > 0.$$

Určete střední hodnotu doby hledání T , potřebné k nalezení automobilu.

8.10 Jaká je pravděpodobnost, že náhodná veličina X s normálním rozdělením $N(3; 1)$ nabude hodnoty

- a) menší než 2,
- b) menší nebo rovné 2,
- c) větší než 4,5,
- d) rovné 1,
- e) v mezích od 2,5 do 3,5?

8.11 Vytvořte graf hustoty pravděpodobnosti a distribuční funkce spojité náhodné veličiny X se standardizovaným normálním rozdělením.

8.12 Jaké jsou hodnoty μ a σ^2 normálně rozdělené náhodné veličiny X , jestliže $\mathbb{P}(X < 85) = 0,9$ a $\mathbb{P}(X < 95) = 0,95$?

8.13 Jaká je pravděpodobnost, že náhodná veličina s normovaným normálním rozdělením nabude hodnoty

- a) menší než 1,64,
- b) větší než $-1,64$,
- c) v mezích od $-1,96$ do $1,96$,
- d) větší než 2,33,
- e) menší než $-2,33$,
- f) přesně 0,5?

8.14 Náhodná veličina X , která představuje chybu měření, má rozdělení $N(0,2; 0,64)$. Určete

- pravděpodobnost, že absolutní hodnota veličiny X bude menší než 1,
- hodnotu chyby měření, která může být překročena pouze s pravděpodobností 0,05.

8.15 Dokažte, že pro každé reálné číslo x platí

$$\Phi(x) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \left(\frac{x}{1} + \frac{x^3}{1 \cdot 3} + \frac{x^5}{1 \cdot 3 \cdot 5} + \dots \right).$$



8.16 Při kontrole se přijímají všechny výrobky, jejichž délka přesahuje 77 cm. Bylo zjištěno, že střední hodnota délky výrobku (náhodné veličiny X) μ je 75 cm a směrodatná odchylka (odmocnina z rozptylu) σ je 5 cm. Za předpokladu, že sledovaná náhodná veličina má přibližně normální rozdělení, určete

- pravděpodobnost, že výrobek, který prošel kontrolou, je delší než 80 cm,
- kolik výrobků delších než 80 cm můžeme očekávat, jestliže kontrolou je přijato 2 261 kusů?

8.17 Vytvořte graf hustoty pravděpodobnosti a distribuční funkce náhodné veličiny X s logaritmicko-normálním rozdělením, tedy $X \sim \text{LN}(0; 1)$, na intervalu $(0; 10]$. Vypočítejte 23% kvantil pro rozdělení $\text{LN}(0; 1)$.

8.18 Uvažujte čas, který student potřebuje na vyřešení problému, jako o náhodné veličině s gama rozdělením, u níž je známa střední hodnota 1,5 a rozptyl 0,75. Jaká je pravděpodobnost, že doba pro vyřešení problému přesáhne 2 hodiny?

Řešení úloh

8.1

$$\begin{aligned}
\mathbb{E}X &= \int_{-\infty}^{\infty} xf(x) \, dx = \int_a^b x \frac{1}{b-a} \, dx = \frac{1}{b-a} \left[\frac{x^2}{2} \right]_a^b = \frac{a+b}{2}, \\
\text{var } X &= \mathbb{E}X^2 - (\mathbb{E}X)^2 = \int_a^b x^2 \frac{1}{b-a} \, dx - \left(\frac{a+b}{2} \right)^2 = \\
&= \frac{1}{b-a} \left[\frac{x^3}{3} \right]_a^b - \frac{(a+b)^2}{4} = \frac{a^2 + ab + b^2}{3} - \frac{a^2 + 2ab + b^2}{4} = \\
&= \frac{(a-b)^2}{12}.
\end{aligned}$$

8.2 Pro $x \in [a; b]$ je

$$F(x) = \int_a^x \frac{1}{b-a} \, dt = \frac{1}{b-a} [t]_a^b = \frac{x-a}{b-a}.$$

8.3 $X \sim U(0; 90)$; a) $1/9$, b) $1/3$, c) $2/9$.

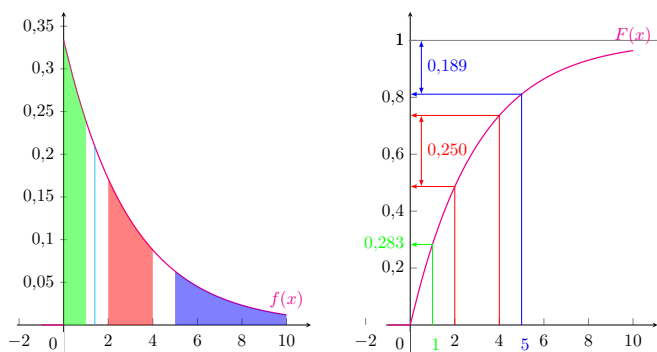
8.4 $X \sim U(0; 8)$

- a) $\mathbb{P}(X \leq 3) = F(3) = 0,375$,
- b) $\mathbb{P}(X \geq 7) = 1 - \mathbb{P}(X \leq 7) = 1 - 0,875 = 0,125$,
- c) $\mathbb{P}(X = 0) = 0$,
- d) $\mathbb{P}(4 \leq X \leq 7) = F(7) - F(3) = 0,375$.

8.5 0,32 roku, tedy necelé 4 měsíce

8.6 $X \sim U(12; 16)$, 0,125.

8.7 a) 0,283 (zelená plocha); b) 0 (azurová „plocha“); c) 0,189 (modrá plocha); d) 0,250 (červená plocha).



Obrázek 8.11: Graf hustoty a distribuční funkce z úlohy 8.7

8.8 13 minut a 54,25 vteřiny

8.9 $\mathbb{E}T = \frac{1}{\gamma}$

8.10 a) $\mathbb{P}(X < 2) \doteq 0,159$; b) $\mathbb{P}(X \leq 2) \doteq 0,159$; c) $\mathbb{P}(X > 4,5) \doteq 0,068$;
d) $\mathbb{P}(X = 1) = 0$; e) $\mathbb{P}(2,5 \leq X \leq 3,5) \doteq 0,383$.

8.11 Řešení je na obrázku 8.6.

8.12 $\mu \doteq 49,7$; $\sigma^2 \doteq 757,64$.

8.13 a) 0,949 5; b) 0,949 5; c) 0,950 0; d) 0,009 9; e) 0,009 9; f) 0.

8.14 a) 0,774 5; b) 1,516.

8.15

8.16

- a) $X \sim N(75; 25)$, $\mathbb{P}(X > 80) = 1 - \mathbb{P}(X \leq 80) \doteq 0,158 7$;
b) $6\,562 \cdot 0,158 7 \doteq 1041$.

8.17 Graf hustoty a distribuční funkce ukazuje obrázek 8.7 na straně 165.
 $F^{-1}(0,23) \doteq 0,478$.

8.18 $X \sim \Gamma(3; 2)$, $\mathbb{P}(X > 2) = 1 - \mathbb{P}(X \leq 2) \doteq 0,238 1$.

Kapitola 9

Náhodný vektor, nezávislost náhodných veličin

9.1 Náhodný vektor

Potřebujeme-li pracovat s několika náhodnými veličinami současně, je vhodné k tomu zavést pojem náhodného vektoru jako kolekce náhodných veličin a na něm zavést potřebné pravděpodobnostní pojmy. To v této kapitole učiníme a nadto ještě zavedeme pojem korelace a rozšíříme pojem nezávislosti i na náhodné veličiny.

Definice 9.1 (Náhodný vektor)

Nechť $(\Omega, \mathcal{A}, \mathbb{P})$ je pravděpodobnostní prostor a na něm jsou definovány náhodné veličiny $X_1, X_2, \dots, X_n$. Pak uspořádanou n -tici $\mathbf{X} = (X_1, \dots, X_n)^T$ nazýváme n -rozměrný náhodný vektor¹.



Náhodný vektor je tedy zobrazení z Ω do $\mathbb{R}^n$. Hodnoty náhodného vektoru je možno geometricky interpretovat jako bod v n -rozměrném prostoru. Speciálně pro $n = 2$ píšeme $\mathbf{X} = (X_1, X_2)^T$ nebo též pouze $\mathbf{X} = (X, Y)^T$, mluvíme o dvourozměrném náhodném vektoru a jeho hodnoty geometricky interpretujeme jako body v rovině o souřadnicích $[x; y]$.

Podobně jako u náhodné veličiny popisuje chování náhodného vektoru $\mathbf{X} = (X_1, \dots, X_n)^T$ distribuční funkce, v tomto případě n -rozměrná.

Definice 9.2 (Distribuční funkce náhodného vektoru)

Nechť $\mathbf{X} = (X_1, \dots, X_n)^T$ je náhodný vektor definovaný na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, \mathbb{P})$. Sdruženou distribuční funkcí $F_{\mathbf{X}}$ náhodného

¹Také někdy n -rozměrná náhodná veličina.

vektoru $\mathbf{X}$ nazveme reálnou funkcí n proměnných definovanou na $\mathbb{R}^n$ vztahem

$$F_{\mathbf{X}}(x_1, \dots, x_n) = \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n) = \\ = \mathbb{P}\left(\bigcap_{i=1}^n \{\omega : X_i(\omega) \leq x_i\}\right), \quad -\infty < x_i < \infty, \quad i = 1, \dots, n.$$

Dvourozměrný náhodný vektor $\mathbf{X} = (X, Y)^T$ tedy charakterizuje distribuční funkce $F_{\mathbf{X}}(x, y) = \mathbb{P}(X \leq x, Y \leq y)$ dvou reálných proměnných x, y .

Pro účely následující věty zavedme pojem n -rozměrného intervalu $(\mathbf{a}; \mathbf{b}]$, kde $\mathbf{a} = (a_1, \dots, a_n)$, $\mathbf{b} = (b_1, \dots, b_n)$, jako

$$(\mathbf{a}, \mathbf{b}] = \bigtimes_{i=1}^n (a_i, b_i].$$

Ve dvourozměrném případě je tedy tímto útvarem obdélník, v třírozměrném kvádr. Označíme dále $\Delta_{k,n}$ množinu $\binom{n}{k}$ všech n -tic $(z_1, \dots, z_n)$ takových, že každé z_i je rovno a_i nebo b_i , přičemž první možnost nastává právě k -krát (a druhá tedy $(n - k)$ -krát). Sjednocení

$$\Delta = \bigcup_{k=0}^n \Delta_{k,n}$$

tedy představuje množinu všech 2^n vrcholů intervalu $(\mathbf{a}; \mathbf{b}]$.

Věta 9.1 (o vlastnostech distribuční funkce). *Nechť je $F_{\mathbf{X}}$ distribuční funkce náhodného vektoru $\mathbf{X} = (X_1, \dots, X_n)^T$. Pak $F_{\mathbf{X}}$ má tyto vlastnosti:*

- $F_{\mathbf{X}}(x_1, \dots, x_n)$ je neklesající funkce v každé ze svých proměnných při pevných hodnotách ostatních proměnných.
- $F_{\mathbf{X}}(x_1, \dots, x_n)$ je zprava spojitá v každé proměnné.
- $\lim_{x_i \rightarrow -\infty} F_{\mathbf{X}}(x_1, \dots, x_n) = 0$, $i = 1, \dots, n$, kde hodnoty x_j jsou pevné pro všechna $j \neq i$, $j = 1, \dots, n$.
- $\lim_{\substack{x_1 \rightarrow \infty \\ x_2 \rightarrow \infty \\ \vdots \\ x_n \rightarrow \infty}} F_{\mathbf{X}}(x_1, \dots, x_n) = 1$.
- Pro $(\mathbf{a}, \mathbf{b}]$ je

$$\sum_{k=0}^n (-1)^k \sum_{\boldsymbol{\delta} \in \Delta_{k,n}} F_{\mathbf{X}}(\boldsymbol{\delta}) \geq 0.$$

Obráceně, každá reálná funkce n proměnných definovaná na $\mathbb{R}^n$ s vlastnostmi a)–e) je distribuční funkcí nějakého náhodného vektoru.

Důkaz. Vlastnosti a), b), c), d) se dokazují stejně jako ve větě 5.1, vlastnost e) je dokázána například v [8]. $\square$

Náhodné veličiny, jež jsou prvky náhodného vektoru, mohou být diskrétního i spojitého charakteru. O něco jednodušším případem je následující situace, kdy jsou tyto prvky stejného typu.

Definice 9.3 (Diskrétní rozdělení náhodného vektoru)

Náhodný vektor $\mathbf{X}$ má diskrétní rozdělení, jestliže existuje posloupnost vektorů hodnot $\{\mathbf{x}_k\}_{k=1}^{\infty}$, $\mathbf{x}_k \in \mathbb{R}^n$ a odpovídající posloupnost kladných čísel $\{p_k\}_{k=1}^{\infty}$ taková, že

$$\sum_{k=1}^{\infty} p_k = 1,$$

kde

$$p_k = \mathbb{P}(\mathbf{X} = \mathbf{x}_k) = \mathbb{P}(\{\omega \in \Omega : \mathbf{X}(\omega) = \mathbf{x}_k\}).$$

Distribuční funkce náhodného vektoru $\mathbf{X}$ diskrétního typu je pro libovolné $\mathbf{x} \in \mathbb{R}^n$ ve tvaru

$$F_{\mathbf{X}}(\mathbf{x}) = \sum_{\{k : \mathbf{x}_k \leq \mathbf{x}\}} p_k,$$

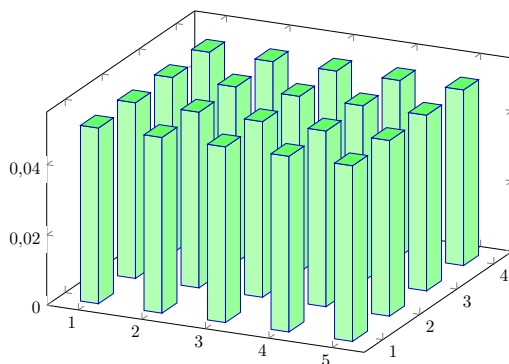
kde $\mathbf{x}_k \leq \mathbf{x}$ znamená $\mathbf{x}_k \in (-\infty, \mathbf{x}]$.

Příklad 9.1. Uvažujme dvě osudí s koulemi: v prvním osudí jsou čtyři koule očíslované 1 až 4, v druhém osudí je pět koulí očíslovaných 1 až 5. Z obou osudí vylosujeme vždy právě jednu kouli. Nechť náhodná veličina X značí číslo koule vytažené z prvního osudí a nechť Y značí číslo koule vytažené z druhého osudí. Určete sdruženou pravděpodobnostní funkci a vytvořte její graf.

Řešení. Náhodné veličiny X a Y tvoří diskrétní náhodný vektor $\mathbf{X} = (X, Y)^T$, který je plně charakterizován svou sdruženou pravděpodobnostní funkcí $p(x, y)$ s definičním oborem $\{(x, y) : x = 1, 2, 3, 4, y = 1, 2, 3, 4, 5\}$. Ke každé dvojici (x, y) přísluší v tomto případě identická pravděpodobnost

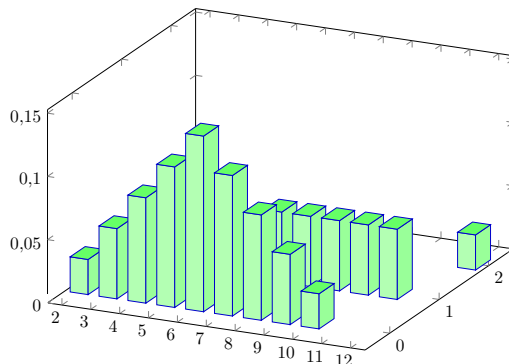
$$p(x, y) = P(X = x, Y = y) = \frac{1}{20}.$$

Grafem sdružené pravděpodobnostní funkce je dvacet bodů $(x, y, p(x, y))$ v trojrozměrné souřadné soustavě, místo nichž se však často pro lepší názornost zobrazují sloupce, viz obrázek 9.1. $\triangle$

Obrázek 9.1: Graf sdružené pravděpodobnostní funkce $p(x, y)$ v příkladu 9.1

Tabulka 9.1: Sdružené a marginální pravděpodobnosti pro příklad 9.2

(x, y)	2	3	4	5	6	7	8	9	10	11	12	$\mathbb{P}(X = x)$
0	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	0	0	$\frac{25}{36}$
1	0	0	0	0	0	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	0	$\frac{10}{36}$
2	0	0	0	0	0	0	0	0	0	0	$\frac{1}{36}$	$\frac{1}{36}$
$\mathbb{P}(Y = y)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	1

Obrázek 9.2: Graf sdružené pravděpodobnostní funkce $p(x, y)$ v příkladu 9.2

Příklad 9.2. Sestrojte náhodný vektor $\mathbf{X} = (X, Y)^T$, kde náhodná veličina X udává počet hozených šestek při hodu dvěma hracími kostkami a náhodná veličina Y udává součet ok padlých na svrchních stěnách obou kostek. Najděte sdruženou pravděpodobnostní funkci a vytvořte její graf.

Řešení. Při hodu dvěma hracími kostkami je množinou všech možných výsledků

$$\Omega = \{(\square, \square), (\square, \blacksquare), \dots, (\blacksquare, \blacksquare), (\blacksquare, \blacksquare)\}.$$

Počet prvků množiny Ω je 36. Podle počtu hozených šestek a součtu ok na obou kostkách přísluší ke každému výsledku dvojice hodnot náhodného vektoru $\mathbf{X}$ z množiny

$$\{(x, y) : x = 0, 1, 2, y = 2, 3, \dots, 12\}.$$



Například výsledek $(\begin{smallmatrix} 1 & 8 \end{smallmatrix})$ odpovídá situaci $X = 1, Y = 8$, tedy dvojici $(1, 8)$. Ke každé dvojici (x, y) ze zmíněné množiny dopočítáme příslušející pravděpodobnost $p(x, y) = \mathbb{P}(X = x, Y = y)$ (viz tabulku 9.1).

Protože $\sum_{(x,y)} p(x, y) = 1$, je výpis pravděpodobností rozdělením a $p(x, y)$ je jeho sdružená pravděpodobnostní funkce. Grafem sdružené pravděpodobnostní funkce je 16 bodů v soustavě souřadné $(x, y, p(x, y))$, místo nichž obvykle znázorňujeme sloupce jako na obrázku 9.2. Grafy podobné obrázkům a 9.2 lze nakreslit i v Excelu nástrojem Prostorový sloupcový graf na kartě Vložení nad daty například tabulky 9.1. $\triangle$

Definice 9.4 (Absolutně spojitě rozdělení náhodného vektoru)

Náhodný vektor $\mathbf{X} = (X_1, \dots, X_n)^T$ má absolutně spojitě rozdělení, jestliže existuje nezáporná funkce $f_{\mathbf{X}}$ n reálných proměnných taková, že

$$F_{\mathbf{X}}(x_1, \dots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{\mathbf{X}}(t_1, \dots, t_n) dt_1 \dots dt_n,$$

kde funkci $f_{\mathbf{X}}$ nazýváme hustotou rozdělení pravděpodobnosti náhodného vektoru $\mathbf{X}$, nebo též sdruženou hustotou náhodných veličin $X_1, \dots, X_n$.

Pro hustotu $f_{\mathbf{X}}$ platí, že

$$f_{\mathbf{X}}(x_1, \dots, x_n) = \frac{\partial^n F_{\mathbf{X}}(X_1, \dots, X_n)}{\partial x_1 \partial x_2 \dots \partial x_n}$$

ve všech bodech $(x_1, \dots, x_n)$, ve kterých derivace existuje, a

$$\int_{\mathbb{R}^n} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(x_1, \dots, x_n) dx_1 \dots dx_n = 1.$$

Příklad 9.3. Uvažujme funkci f spojitých náhodných veličin X a Y , která je pro $c \in \mathbb{R}$ ve tvaru

$$f(x, y) = \begin{cases} c \cdot y \cdot \sin x & \text{pro } 0 \leq x \leq \pi, 0 \leq y \leq \pi, \\ 0 & \text{jinak.} \end{cases}$$

Určete c tak, aby $f(x, y)$ byla hustota pravděpodobnosti náhodného vektoru $\mathbf{X} = (X, Y)^T$. Určete sdruženou distribuční funkci.

Řešení.

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \, dx \, dy = \int_0^{\pi} \int_0^{\pi} cy \sin x \, dx \, dy = 2c \int_0^{\pi} y \, dy = c\pi^2 = 1,$$

odkud již $c = \frac{1}{\pi^2}$. Sdružená hustota pravděpodobnosti má tedy tvar

$$f(x, y) = \begin{cases} \frac{1}{\pi^2} \cdot y \cdot \sin x & \text{pro } 0 \leq x \leq \pi, 0 \leq y \leq \pi, \\ 0 & \text{jinak,} \end{cases}$$

a sdružená distribuční funkce má tvar

$$F(x, y) = \begin{cases} 0 & \text{pro } x < 0 \text{ nebo } y < 0, \\ \int_0^y \int_0^x \frac{1}{\pi^2} t \sin s \, ds \, dt = \frac{1}{\pi^2} \int_0^y t(-\cos x + 1) \, dt = \\ = \frac{y^2}{2\pi^2} (1 - \cos x) & \text{pro } 0 \leq x \leq \pi, 0 \leq y \leq \pi, \\ \frac{1}{2} (1 - \cos x) & \text{pro } 0 \leq x \leq \pi, y > \pi, \\ \frac{y^2}{\pi^2} & \text{pro } x > \pi, 0 \leq y \leq \pi, \\ 1 & \text{pro } x > \pi, y > \pi. \end{cases}$$

△

Distribuční funkci, resp. hustotu části náhodného vektoru $\mathbf{X}$ nazveme *marginální* distribuční funkcí, resp. hustotou. Uvedme si jako příklad marginální distribuční funkci vektoru $(X_1, \dots, X_k)$, $k = 1, \dots, n-1$:

$$F_{(X_1, \dots, X_k)}(x_1, \dots, x_k) = \lim_{\substack{x_{k+1} \rightarrow \infty \\ x_{k+2} \rightarrow \infty \\ \vdots \\ x_n \rightarrow \infty}} F_{\mathbf{X}}(x_1, \dots, x_k, x_{k+1}, \dots, x_n).$$

Marginální hustotu vektoru $(X_1, \dots, X_k)^T$, který je částí náhodného vektoru $\mathbf{X} = (X_1, \dots, X_n)^T$ s absolutně spojitým rozdělením, dostaneme vyintegrováním nadbytečných proměnných:

$$f_{\mathbf{X}}(x_1, \dots, x_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(x_1, \dots, x_n) \, dx_{k+1} \, dx_{k+2} \cdots dx_n.$$

Marginální distribuční funkce vektoru $(X_1, \dots, X_k)$ je pak

$$\begin{aligned} F_{(X_1, X_2, \dots, X_k)}(x_1, x_2, \dots, x_k) &= \\ &= \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_k} f_{(X_1, \dots, X_k)}(t_1, \dots, t_k) \, dt_k \, dt_{k-1} \cdots dt_1, \end{aligned}$$

Analogické vzorce platí pro všechny podvektory vektoru $(X_1, \dots, X_k)^T$. Existuje-li dále sdružená distribuční funkce, resp. sdružená hustota, potom existují všechny marginální distribuční funkce, resp. marginální hustoty. Obrácené tvrzení obecně neplatí.

Příklad 9.4. V příkladu 9.3 určete předpis marginálních hustot veličin X a Y .

Řešení. Marginální hustoty pravděpodobnosti náhodných veličin X a Y jsou

$$f_X(x) = \int_0^\pi \frac{1}{\pi^2} \cdot y \cdot \sin x \, dy = \frac{\sin x}{2} \quad \text{pro } 0 \leq x \leq \pi,$$

$$f_Y(y) = \int_0^\pi \frac{1}{\pi^2} \cdot y \cdot \sin x \, dx = \frac{2y}{\pi^2} \quad \text{pro } 0 \leq y \leq \pi.$$

Mimo uvedené intervaly je hustota nulová, neboť je nulová i sdružená hustota a integrace přes druhou proměnnou na tom nic nezmění. $\triangle$

9.2 Charakteristiky náhodného vektoru

Nejdůležitější charakteristikou náhodného vektoru je, podobně jako v jednorozměrném případě, jeho střední hodnota.

Definice 9.5

Střední hodnotou náhodného vektoru $\mathbf{X} = (X_1, \dots, X_n)^T$ je vektor středních hodnot marginálních rozdělení

$$\mathbb{E}\mathbf{X} = (\mathbb{E}X_1, \dots, \mathbb{E}X_n)^T.$$

Pro charakterizaci vztahu mezi složkami náhodného vektoru zavedeme pojem kovariance a korelace dvou náhodných veličin.

Definice 9.6 (Kovariance náhodných veličin, koeficient korelace)

Nechť X a Y jsou náhodné veličiny a $\mathbb{E}X^2 < \infty$, $\mathbb{E}Y^2 < \infty$. Kovariance $\text{cov}(X, Y)$ náhodných veličin X a Y je definována vztahem

$$\text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]. \quad (9.1)$$

Koeficient korelace $\text{corr}(X, Y)$ je definován vztahem

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var } X} \sqrt{\text{var } Y}} \quad (9.2)$$

pro $\text{var } X > 0$ a $\text{var } Y > 0$.

Kovariance dvou náhodných veličin je vzájemný vztah odchylek od středních hodnot těchto veličin. Jestliže například nízké (podprůměrné) hodnoty X se obvykle pojí s nízkými (podprůměrnými) hodnotami Y a současně vysoké (nadprůměrné) hodnoty X se obvykle pojí s vysokými hodnotami Y , potom součin $Z = (X - EX)(Y - EY)$ bude obvykle vykazovat kladné hodnoty a tedy kovariance (což je střední hodnota veličiny Z) a v důsledku toho i korelace bude kladná (mluvíme o kladně korelovaných veličinách). Analogicky bychom mohli interpretovat i zápornou kovarianci a korelaci.

Kovariance náhodných veličin závisí na jejich měřítku, interpretace numerických hodnot kovariance je tedy velmi obtížná. Cauchyho–Schwarzova–Buňakovského² nerovnost (viz např. [5]) omezuje absolutní hodnotu kovariance shora součinem směrodatných odchylek. Z toho vyplývá, že korelační koeficient může nabývat hodnot pouze mezi -1 a 1 . Korelační koeficient je roven 1 , pokud jsou náhodné veličiny na sobě lineárně závislé a koeficient úměry je kladný. Korelační koeficient je roven -1 , pokud jsou náhodné veličiny lineárně závislé a koeficient úměry je záporný. V těchto případech hovoříme o dokonalé kladné či záporné lineární závislosti.

Hodnoty korelace blízké hodnotám -1 a 1 reprezentují silnou vazbu mezi náhodnými veličinami. Obráceně, jsou-li kovariance a korelace nulové, nejsou náhodné veličiny mezi sebou lineárně svázány vůbec. V těchto případech hovoříme o *nekorelovaných* náhodných veličinách.

Pro samotný výpočet kovariance lze namísto definičního vzorce (9.1) použít tzv. výpočetní tvar

$$\text{cov}(X, Y) = \mathbb{E}(XY) - EXEY. \quad (9.3)$$

Jeho odvození ponecháváme na čtenáři jako cvičení.

Otázka 9.1. Dokažte vztah (9.3)

Příklad 9.5. Vypočítejte korelační koeficient náhodných veličin X a Y z příkladu 9.2 na straně 182.

Řešení. Řádkovými součty tabulky 9.1 na straně 182 dostaneme marginální rozdělení veličiny X a sloupcové součty tabulky dávají marginální rozdělení veličiny Y . Ty následně využijeme pro výpočet středních hodnot:

$$\begin{aligned} EX &= 0 \cdot \frac{25}{36} + 1 \cdot \frac{10}{36} + 2 \cdot \frac{1}{36} = \frac{12}{36} = \frac{1}{3}, \\ EY &= 2 \cdot \frac{1}{36} + \dots + 12 \cdot \frac{1}{36} = \frac{252}{36} = 7. \end{aligned}$$

²Hermann Amandus Schwarz (1843–1921), německý matematik. Viktor Jakovlevič Buňakovský (1804–1889), ruský matematik a fyzik.

Sdružené pravděpodobnosti z téže tabulky využijeme pro výpočet

$$\mathbb{E}(XY) = 0 \cdot 2 \cdot \mathbb{P}(X = 0, Y = 2) + \cdots + 2 \cdot 12 \cdot \mathbb{P}(X = 2, Y = 12) = \frac{114}{36}$$

a tu pak pro výpočet kovariance podle vzorce (9.3):

$$\text{cov}(X, Y) = \frac{114}{36} - \frac{1}{3} \cdot 7 = \frac{30}{36} = \frac{5}{6} \doteq 0,83.$$

Dále vypočítáme rozptyly náhodných veličin X a Y :

$$\text{var } X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = 0 \cdot \frac{25}{36} + 1^2 \cdot \frac{10}{36} + 2^2 \cdot \frac{1}{36} - \left(\frac{1}{3}\right)^2 = \frac{10}{36},$$

$$\text{var } Y = \mathbb{E}Y^2 - (\mathbb{E}Y)^2 = 2^2 \cdot \frac{1}{36} + \cdots + 12^2 \cdot \frac{1}{36} - 7^2 = \frac{210}{36}.$$

Na závěr vypočteme korelaci podle vzorce (9.2):

$$\text{corr}(X, Y) = \frac{\frac{5}{6}}{\sqrt{\frac{10}{36}} \cdot \sqrt{\frac{210}{36}}} \doteq 0,65.$$

Hodnota korelačního koeficientu 0,65 indikuje středně silnou pozitivní vazbu mezi oběma náhodnými veličinami. $\triangle$

Definice 9.7 (Varianční a korelační matice náhodného vektoru)

Nechť $\mathbf{X} = (X_1, \dots, X_n)^T$ je náhodný vektor, jehož složky mají konečný druhý moment. Varianční matice $\text{var } \mathbf{X}$ tohoto náhodného vektoru je definována jako matice typu $n \times n$ s prvky

$$\text{cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)], \quad 1 \leq i, j \leq n.$$

Korelační matice $\text{corr } \mathbf{X}$ je matice s prvky

$$\text{corr}(X_i, X_j) = \frac{\text{cov}(X_i, X_j)}{\sqrt{\text{var } X_i} \sqrt{\text{var } X_j}}, \quad 1 \leq i, j \leq n.$$

Je zřejmé, že pro $i = 1, \dots, n$ platí

$$\text{cov}(X_i, X_i) = \text{var } X_i \quad \text{a} \quad \text{corr}(X_i, X_i) = 1, \quad 1 \leq i \leq n.$$

Například varianční a korelační matice dvourozměrného náhodného vektoru $\mathbf{X} = (X, Y)^T$ mají podobu

$$\text{var } \mathbf{X} = \begin{pmatrix} \text{var } X & \text{cov}(X, Y) \\ \text{cov}(X, Y) & \text{var } Y \end{pmatrix}, \quad \text{corr } \mathbf{X} = \begin{pmatrix} 1 & \text{corr}(X, Y) \\ \text{corr}(X, Y) & 1 \end{pmatrix}.$$

Ve větě 6.2 jsme uvedli tvrzení d) o střední hodnotě součtu náhodných veličin. Obdobné pravidlo platí i pro rozptyl součtu náhodných veličin, jak postulují následující věta, kterou uvádíme bez důkazu.

Věta 9.2. *Jestliže existuje rozptyl součtu náhodných veličin X_i , pak je roven*

$$\text{var} \left(\sum_{i=1}^n X_i \right) = \sum_{i=1}^n \text{var} X_i + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{cov}(X_i, X_j).$$

Jsou-li náhodné veličiny X_i navíc nekorelované, pak platí

$$\text{var} \left(\sum_{i=1}^n X_i \right) = \sum_{i=1}^n \text{var} X_i. \quad (9.4)$$

Příklad 9.6. Diverzifikace portfolia spočívá ve výběru aktiv, jejichž výnosy jsou co nejméně korelované (nebo dokonce negativně korelované). Tím se snižuje celkový rozptyl (riziko) portfolia, protože negativně korelované složky se vzájemně vyvažují. Při negativní korelaci mezi složkami může investor výrazně snížit riziko bez nutnosti snižovat očekávaný výnos.

Uvažujme dva náhodné vektory X a Y , které reprezentují dvě složky určitého (například finančního) portfolia. Necht' mají obě složky stejnou střední hodnotu μ a stejný rozptyl σ^2 . Vypočítejte rozptyl součtu těchto složek portfolia pro různé hodnoty korelačního koeficientu:

- a) $\text{corr}(X, Y) = 0,5$,
- b) $\text{corr}(X, Y) = 0$,
- c) $\text{corr}(X, Y) = -0,5$.

Řešení.

$$\text{var}(X + Y) = \text{var} X + \text{var} Y + 2 \cdot \text{cov}(X, Y),$$

kde $\text{var} X = \text{var} Y = \sigma^2$.

$$\text{cov}(X, Y) = \text{corr}(X, Y) \cdot \sigma_X \cdot \sigma_Y = \text{corr}(X, Y) \cdot \sigma^2.$$

Po dosazení dostaneme

$$\text{var}(X + Y) = \sigma^2 + \sigma^2 + 2 \cdot \text{corr}(X, Y) \cdot \sigma^2 = 2\sigma^2 \cdot (1 + \text{corr}(X, Y)).$$

Pro jednotlivé hodnoty korelačního koeficientu dostaneme:

- a) $\text{var}(X + Y) = 3\sigma^2$,
- b) $\text{var}(X + Y) = 2\sigma^2$,
- c) $\text{var}(X + Y) = \sigma^2$.

Pro kladnou korelaci $\text{corr}(X, Y) = 0,5$ je rozptyl součtu složek $3\sigma^2$, což znamená, že složky se vzájemně posilují a diverzifikace není efektivní. Vysoká korelace mezi složkami vede k větší celkové variabilitě portfolia a tedy ke zvýšení rizika. Pro nulovou korelaci $\text{corr}(X, Y) = 0$ je rozptyl součtu složek

$2\sigma^2$, což je nižší než v případě kladné korelace a vede ke slabší formě diverzifikace, ale ne k maximální diverzifikaci. Teprve pro $\text{corr}(X, Y) = -0,5$ je rozptyl součtu složek σ^2 , což je nejnižší hodnota mezi těmito třemi případy. Záporná korelace mezi složkami znamená, že když jedna složka roste, druhá klesá, což má za následek silnou diverzifikaci portfolia a vede to k výraznému snížení jeho celkového rizika. $\triangle$

9.3 Některá mnohorozměrná rozdělení

9.3.1 Multinomické rozdělení

Multinomické rozdělení je zobecněním binomického rozdělení a je patrně nejdůležitějším diskrétním mnohorozměrným rozdělením. Mějme urnu a v ní kuličky k různých barev. Necht' pravděpodobnost vytažení kuličky i -té barvy je rovna p_i , $i = 1, 2, \dots, k$, přičemž

$$0 < p_i < 1, \quad p_1 + p_2 + \dots + p_k = 1.$$

Z této urny vybereme n -krát nezávisle po jedné kuličce (po vylosování kuličky vracíme zpět do osudí). Označme X_i počet kuliček i -té barvy, které byly vybrány. Z kombinatorických pravidel vyplývá, že náhodný vektor $\mathbf{X} = (X_1, X_2, \dots, X_k)^T$ má sdruženou pravděpodobnostní funkci

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) = \frac{n!}{x_1! x_2! \dots x_k!} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}$$

pro $x_i = 0, 1, \dots, n$, $i = 1, 2, \dots, k$ a $x_1 + x_2 + \dots + x_k = n$.

Parametry multinomického rozdělení jsou $(n, p_1, p_2, \dots, p_k)$. Marginální rozdělení X_i je binomické rozdělení s parametry (n, p_i) pro $i = 1, 2, \dots, k$. Střední hodnota je $\mathbb{E}X_i = np_i$ pro $i = 1, 2, \dots, k$. Prvky varianční matice jsou

$$\begin{aligned} \text{var } X_i &= np_i(1 - p_i) \quad \text{pro } i = 1, 2, \dots, k, \\ \text{cov}(X_i, X_j) &= -np_i p_j \quad \text{pro } i \neq j. \end{aligned}$$

Příklad 9.7. Z dlouhodobých statistik ve firmě je známo, že projekt přijmou s pravděpodobností 0,35, odmítnou ho s pravděpodobností 0,5 a odloží k následnému posouzení s pravděpodobností 0,15. Určete pravděpodobnost, že z 15 projektů firma pět přijme, sedm odmítne a tři odloží na následné posouzení.



Řešení. Náhodná veličina X_1 označuje počet přijatých projektů, náhodná veličina X_2 značí počet odmítnutých projektů a náhodná veličina X_3 značí počet projektů, které budou následně posouzeny, $p_1 = 0,35$, $p_2 = 0,5$, $p_3 = 0,15$.

$$\mathbb{P}(X_1 = 5, X_2 = 7, X_3 = 3) = \frac{15!}{5! \cdot 7! \cdot 3!} \cdot 0,25^4 \cdot 0,5^7 \cdot 0,15^3 \doteq 0,037.$$

První činitel lze vypočítat jako $\text{MULTINOMIAL}(5;7;3)$.

△

9.3.2 Obecné dvourozměrné normální rozdělení

Normální rozdělení může být i vícerozměrné, pro jednoduchost uvedeme pouze dvourozměrný případ, tedy dvourozměrné normální rozdělení.

$$N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$$

je rozdělení náhodného vektoru $(X, Y)^T$ s hustotou

$$f(x, y) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}} \cdot \exp\left\{-\frac{1}{2(1-\rho^2)}\left\{\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{x-\mu_1}{\sigma_1}\frac{y-\mu_2}{\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right\}\right\}$$

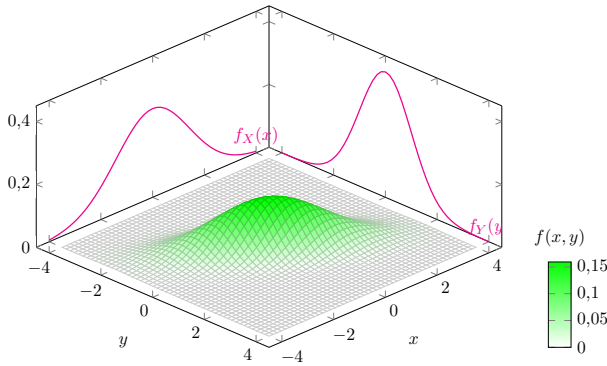
pro $(x, y) \in \mathbb{R}^2$, kde $\mu_i \in \mathbb{R}$, $\sigma_i^2 > 0$, $i = 1, 2$ a $\rho \in (-1, 1)$ jsou parametry. Výraz v exponentu lze také psát jako

$$-\frac{1}{2}\begin{pmatrix} x - \mu_1 \\ y - \mu_2 \end{pmatrix}^T \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}^{-1} \begin{pmatrix} x - \mu_1 \\ y - \mu_2 \end{pmatrix}.$$

Marginální rozdělení X a Y jsou $N(\mu_1; \sigma_1^2)$ a $N(\mu_2; \sigma_2^2)$ a

$$\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

je varianční matice vektoru $(X, Y)^T$; ρ je koeficient korelace $\text{corr}(X, Y)$. Normujeme-li X a Y , pak dvojice veličin $W = \frac{(X - \mu_1)}{\sigma_1}$ a $Z = \frac{(Y - \mu_2)}{\sigma_2}$ má normované dvourozměrné normální rozdělení $N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$.



Obrázek 9.3: Graf hustoty pravděpodobnosti dvourozměrného normálního rozdělení $N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}\right)$

9.4 Nezávislé náhodné veličiny

V odstavci 3.2 jsme definovali nezávislost náhodných jevů. Pojem nyní rozšíříme na nezávislost náhodných veličin $X_1, \dots, X_n$.

Definice 9.8 (Vzájemně nezávislé náhodné veličiny)

Náhodné veličiny $X_1, X_2, \dots, X_n$ jsou vzájemně nezávislé, jestliže

$$\mathbb{P}\left(\bigcap_{j=1}^r \{\omega : X_{i_j} < x_{i_j}\}\right) = \prod_{j=1}^r \mathbb{P}(\{\omega : X_{i_j} < x_{i_j}\}) \quad (9.5)$$

pro všechny r -tice $\{i_1, i_2, \dots, i_r\} \subset \{1, 2, \dots, n\}$, $1 \leq r \leq n$, a všechna $(x_1, \dots, x_n) \in \mathbb{R}^n$.

Podobně jako u náhodných jevů můžeme i zde definovat nezávislost náhodných veličin $X_1, X_2, \dots, X_n$ po dvou. Tuto slabší definici nezávislosti bychom dostali z definice 9.8, pokud bychom omezili hodnotu r pouze na $r = 2$.

Ověřovat nezávislost náhodných veličin podle definice 9.8 je dosti náročné, proto si uvedeme kritéria, podle nichž se ověřuje nezávislost náhodných veličin v praxi.

Věta 9.3. *Nechť náhodný vektor $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$ má sdruženou distribuční funkci $F_{\mathbf{X}}(x_1, x_2, \dots, x_n)$. Nechť $F_{X_i}(x_i)$ je marginální distribuční funkce náhodné veličiny X_i , $i = 1, 2, \dots, n$. Pak náhodné veličiny $X_1, X_2, \dots, X_n$ jsou nezávislé právě tehdy, když platí*

$$F_{\mathbf{X}}(x_1, x_2, \dots, x_n) = F_{X_1}(x_1) \cdot F_{X_2}(x_2) \cdots F_{X_n}(x_n), \quad (9.6)$$

pro všechna $(x_1, x_2, \dots, x_n) \in \mathbb{R}^n$.

Důkaz. Jsou-li $X_1, X_2, \dots, X_n$ nezávislé, pak podle definice 9.8 platí vztah (9.5) pro všechny podmnožiny $\{i_1, i_2, \dots, i_r\} \subset \{1, 2, \dots, n\}$, $1 \leq r \leq n$ a pro libovolné $x \in \mathbb{R}^n$, tudíž podle definice distribuční funkce platí i (9.6).

Předpokládejme nyní, že platí (9.6). Vezměme si libovolnou podmnožinu $\{i_1, i_2, \dots, i_r\}$ množiny $\{1, 2, \dots, n\}$. Pak tvrzení věty plyne z definice marginální distribuční funkce. $\square$

Budeme-li uvažovat náhodný vektor $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$ absolutně spojitého typu, pak nezávislost jeho složek budeme ověřovat podle následujícího kritéria.

Věta 9.4. *Nechť $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$ je náhodný vektor absolutně spojitého typu. Náhodné veličiny $X_1, X_2, \dots, X_n$ jsou vzájemně nezávislé právě tehdy, platí-li*

$$f_{\mathbf{X}}(x_1, x_2, \dots, x_n) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) \cdots f_{X_n}(x_n),$$

pro všechna $(x_1, x_2, \dots, x_n) \in \mathbb{R}^n$.

Důkaz. Plyne z věty 9.3. $\square$

Pro ověření nezávislosti složek náhodného vektoru $\mathbf{X}$ diskrétního typu platí toto kritérium:

Věta 9.5. *Nechť $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$ je náhodný vektor diskrétního typu. Náhodné veličiny $X_1, X_2, \dots, X_n$ jsou vzájemně nezávislé právě tehdy, když platí*

$$\mathbb{P}\left(X_1 = x_1^{(i)}, \dots, X_n = x_n^{(i)}\right) = \prod_{j=1}^n \mathbb{P}\left(X_j = x_j^{(i)}\right),$$

kde $\mathbf{x}^{(i)} = (x_1^{(i)}, x_2^{(i)}, \dots, x_n^{(i)})$, $i = 1, 2, \dots$ jsou všechny výsledky náhodného vektoru $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$.

Důkaz. Plyne z věty 9.3. $\square$

Příklad 9.8. Dvojice součástí má dobu životnosti v letech popsánu hustotou

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{2}e^{-x-\frac{y}{2}} & x > 0, y > 0, \\ 0 & \text{jinak.} \end{cases}$$

- Určete předpis distribuční funkce $F_{\mathbf{X}}$ náhodného vektoru $\mathbf{X} = (X, Y)^T$.
- Jaká je pravděpodobnost, že „druhá součástka přežije první“ (jev A)?

- c) Jaká je pravděpodobnost toho, že druhá součástka bude žít alespoň dvakrát déle, než první?
 d) Spočítejte marginální hustoty.
 e) Určete, zda jsou složky X a Y nezávislé.
 f) Určete $\mathbb{E}\mathbf{X}$ a varianční matici $\text{var } \mathbf{X}$.

Řešení.

a) $\int_0^y \int_0^x \frac{1}{2} e^{-s-\frac{t}{2}} ds dt = e^{-x-\frac{y}{2}} + 2e^{-x} - e^{-\frac{y}{2}} + 1$, tedy

$$F_X(x, y) = \begin{cases} e^{-x-\frac{y}{2}} + 2e^{-x} - e^{-\frac{y}{2}} + 1 & x > 0, y > 0, \\ 0 & \text{jinak.} \end{cases}$$

- b) Hledaná pravděpodobnost je dána integrálem

$$\begin{aligned} \mathbb{P}(A) &= \int_0^\infty \int_x^\infty \frac{1}{2} e^{-x-\frac{y}{2}} dy dx = \\ &= \int_0^\infty \frac{1}{2} e^{-x} \left(\int_x^\infty e^{-\frac{y}{2}} dy \right) dx = \int_0^\infty \frac{1}{2} e^{-x} \left[-2 \cdot e^{-\frac{y}{2}} \right]_x^\infty dx = \\ &= \int_0^\infty \frac{1}{2} e^{-x} \cdot 2e^{-\frac{x}{2}} dx = \int_0^\infty e^{-\frac{3}{2}x} dx = \left[-\frac{2}{3} e^{-\frac{3}{2}x} \right]_0^\infty = \frac{2}{3}. \end{aligned}$$

- c) Pravděpodobnost je dána integrálem

$$\int_0^\infty \int_{2x}^\infty \frac{1}{2} e^{-x-\frac{y}{2}} dy dx = \frac{1}{2},$$

který vypočteme stejně jako v předchozím případě.

- d) Marginální hustoty jsou:

$$\begin{aligned} f_X(x) &= \int_0^\infty \frac{1}{2} e^{-x-\frac{y}{2}} dy = \frac{1}{2} e^{-x} \cdot [-2e^{-\frac{y}{2}}]_0^\infty = e^{-x} \text{ pro } x > 0, \\ f_Y(y) &= \int_0^\infty \frac{1}{2} e^{-x-\frac{y}{2}} dx = \frac{1}{2} e^{-\frac{y}{2}} \cdot [-e^{-x}]_0^\infty = \frac{1}{2} e^{-\frac{y}{2}} \text{ pro } y > 0, \end{aligned}$$

nulové v ostatních případech.

- e) Složky jsou nezávislé právě tehdy, když $f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$ pro libovolná x, y . Pro $x > 0$ a $y > 0$ rovnost platí podle d). V ostatních případech je nulová jak sdružená hustota, tak alespoň jedna z hustot marginálních a rovnost tedy platí rovněž.

f) Pro určení $\mathbb{E}(X, Y)$ a varianční matice nejdříve spočítáme:

$$\mathbb{E}X = \int_0^\infty x f_X(x) dx = \int_0^\infty x \cdot e^{-x} dx = 1$$

$$\mathbb{E}Y = \int_0^\infty y f_Y(y) dy = \int_0^\infty y \cdot \frac{1}{2} e^{-\frac{y}{2}} dy = 2$$

$$\text{var} X = \mathbb{E}X^2 - (\mathbb{E}X)^2, \text{var} Y = \mathbb{E}Y^2 - (\mathbb{E}Y)^2$$

$$\mathbb{E}X^2 = \int_0^\infty x^2 \cdot e^{-x} dx = 2$$

$$\mathbb{E}Y^2 = \int_0^\infty y^2 \cdot \frac{1}{2} e^{-\frac{y}{2}} dy = 8$$

$$\text{var} X = 2 - 1^2 = 1$$

$$\text{var} Y = 8 - 2^2 = 4$$

$$\begin{aligned} \mathbb{E}XY &= \int_0^\infty \int_0^\infty x \cdot y \frac{1}{2} e^{-x-\frac{y}{2}} dx dy = \int_0^\infty y \frac{1}{2} e^{-\frac{y}{2}} \cdot \int_0^\infty x \frac{1}{2} e^{-x} dx dy = \\ &= \int_0^\infty \frac{1}{2} e^{-\frac{y}{2}} dy \cdot \int_0^\infty x \frac{1}{2} e^{-x} dx = (\mathbb{E}Y) \cdot (\mathbb{E}X), \end{aligned}$$

odkud již plyne $\text{cov}(X, Y) = \mathbb{E}XY - (\mathbb{E}X) \cdot (\mathbb{E}Y) = 0$. Varianční matice je tedy rovna

$$\text{var}(X, Y) = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}.$$

△

9.5 Úlohy

9.1 Necht je dána funkce

$$p(x, y) = \begin{cases} \frac{1}{11}(y - x) & x = 1, 2, 3, y = x^2 + 1, \\ 0 & \text{jinde.} \end{cases}$$

- Dokažte, že daná funkce $p(x, y)$ je rozdělení nějakého náhodného vektoru $\mathbf{X}$.
- Znázorněte funkci $p(x, y)$.
- Vypočítejte střední hodnotu $\mathbb{E}\mathbf{X}$.
- Určete varianční matici.

9.2 Funkce dvou proměnných x a y je dána předpisem

$$f(x, y) = \begin{cases} c \cdot x \cdot y \cdot e^{-x^2-y^2} & \text{pro } (x, y) \in (0; \infty) \times (0; \infty), \\ 0 & \text{jinde.} \end{cases}$$

- Určete konstantu c tak, aby funkce $f(x, y)$ byla hustotou nějakého náhodného vektoru $(X, Y)^T$.
- Určete distribuční funkci a marginální distribuční funkce.
- Určete marginální hustoty.
- Určete střední hodnoty.
- Vypočítejte varianční matici.
- Ověřte, zda jsou vektory nezávislé.

9.3 Necht sdružená hustota náhodného vektoru $(R, \Phi)^T$ je

$$g(r, \varphi) = \begin{cases} \frac{1}{2\pi\sigma^2} \cdot e^{-\frac{r^2}{2\sigma^2}} \cdot r & r \in (0; \infty), \varphi \in (0; 2\pi), \\ 0 & \text{jinde,} \end{cases}$$

kde r a φ jsou nezávislé polární souřadnice bodu v Gaussově rovině, náhodná veličina R má rovnoměrné rozdělení a veličina Φ má normální rozdělení. (Funkce $g(r, \varphi)$ je hustotou tzv. Rayleighova³ spojitého rozdělení, které se využívá například v meteorologii.)

- Ukažte, že je $g(r, \varphi)$ je hustota vektoru $(R, \Phi)^T$.
- Určete distribuční funkci $G(r, \varphi)$.
- Najděte marginální hustoty náhodných veličin R a Φ .
- Zjistěte, zda náhodné veličiny R a Φ jsou nezávislé.

³John William Strutt, 3. baron Rayleigh (1842–1919), anglický fyzik.

9.4 Je dána funkce

$$f(x, y) = \begin{cases} c \cdot (x^2 + y) \cdot e^x & \text{pro } (x, y) \in [0; 1] \times [0; 1], \\ 0 & \text{jinde.} \end{cases}$$

- Najděte konstantu c tak, aby $f(x, y)$ byla hustota nějakého náhodného vektoru $(X, Y)^T$.
- Vypočítejte distribuční funkci náhodného vektoru $\mathbf{X} = (X, Y)^T$.
- Určete marginální hustoty $f_X(x)$, $f_Y(y)$.
- Určete střední hodnotu $\mathbb{E}\mathbf{X}$.
- Určete varianční matici a zjistěte, zda jsou náhodné veličiny X a Y nezávislé.

9.5 Zkoušený přístroj je složen z pěti prvků. n -tý prvek se porouchá s pravděpodobností

$$p_n = 0,2 + 0,1(n - 1).$$

Poruchy jednotlivých prvků jsou nezávislé. Určete rozptyl počtu porouchaných prvků.

9.6 Házíme desetkrát po sobě hrací kostkou. Nechť náhodná veličina X znamená počet šestek, které padnou při 10 hodech, a náhodná veličina Y znamená počet jedniček.

- Najděte rozdělení pravděpodobnosti (sdruženou pravděpodobnostní funkci) náhodného vektoru $(X, Y)^T$ a sestrojte její graf.
- Spočtěte střední hodnotu a varianční matici.

9.7 Nechť náhodný vektor $(X, Y)^T$ má hustotu

$$f(x, y) = \begin{cases} \frac{1}{(b_1 - a_1)(b_2 - a_2)} & a_1 \leq x \leq b_1, \quad a_2 \leq y \leq b_2, \\ 0 & \text{jinde.} \end{cases}$$

(Náhodný vektor $(X, Y)^T$ s touto hustotou má tzv. dvourozměrné rovnoměrné rozdělení.)

- Ověřte, že $f(x, y)$ je hustotou rozdělení.
- Najděte distribuční funkci náhodného vektoru $(X, Y)^T$.
- Spočtěte střední hodnotu a varianční matici.

9.8 Najděte konstantu c tak, aby funkce

$$f(x, y, z) = \begin{cases} c \cdot z & 0 \leq x \leq 2, \\ & 0 \leq y \leq \sqrt{4 - x^2}, \\ & 0 \leq z \leq \sqrt{4 - x^2 - y^2}, \\ 0 & \text{jinde.} \end{cases}$$

byla hustotou náhodného vektoru $(X, Y, Z)^T$.

9.9 Náhodný vektor $(X, Y, Z)^T$ má sdruženou hustotu

$$f(x, y, z) = \begin{cases} \frac{1}{8}(1 + x \cdot y \cdot z) & -1 \leq x \leq 1, -1 \leq y \leq 1, -1 \leq z \leq 1, \\ 0 & \text{jinde.} \end{cases}$$

- Ověřte, že $f(x, y, z)$ vyjadřuje hustotu.
- Najděte marginální hustoty náhodných veličin X, Y, Z .
- Najděte marginální hustoty náhodných vektorů $(X, Y)^T, (X, Z)^T$ a $(Y, Z)^T$.
- Zjistěte, zda jsou náhodné veličiny X, Y, Z nezávislé po dvou a dále zda jsou vzájemně nezávislé.

9.10 n zaměstnanců jistého podniku obědvá v jedné ze tří restaurací, každý zaměstnanec volí restauraci náhodně, kapacity restaurací jsou n_1, n_2 , a n_3 , kde $n_i \geq n$, $i = 1, 2, 3$. Odvoďte rozdělení $(X_1, X_2, X_3)^T$, kde X_i je počet zaměstnanců, kteří obědvají v i -té restauraci. Odvoďte očekávaný počet neobsazených míst v jednotlivých restauracích.

9.11 Z urny obsahující dvě bílé a dvě černé koule vybíráme za sebou s vracením dvě koule. Definujeme náhodné veličiny X_1, X_2 následovně:

$$X_1 = \begin{cases} 1 & \text{jestliže 1. tažená koule je bílá,} \\ 0 & \text{jinak,} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{jestliže 2. tažená koule je bílá,} \\ 0 & \text{jinak.} \end{cases}$$

Určete distribuční funkci vektoru $(X_1, X_2)^T$ a zjistěte, zda jsou náhodné veličiny X_1 a X_2 nezávislé.

9.12 Nechtě náhodné veličiny U, V mají diskrétní rozdělení určené tabulkou 9.2. Najděte marginální rozdělení každé z obou náhodných veličin, jejich střední hodnoty, rozptyly a kovarianci.

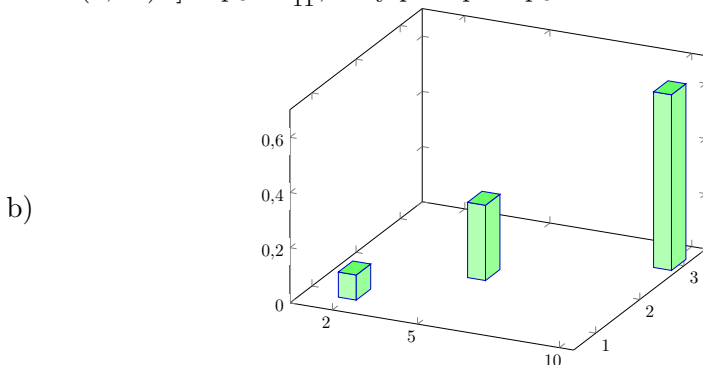
Tabulka 9.2: Rozdělení náhodných veličin pro úlohu 9.12

U / V	1	2	3
1	0,1	0,2	0,3
2	0,2	0,1	0,1

Řešení úloh

9.1

- a) $p(x, y)$ je diskrétní rozdělení nějakého náhodného vektoru $\mathbf{X}$, neboť $\mathbb{P}[\mathbf{X} = (1; 2)^T] = p_1 = \frac{1}{11}$, $\mathbb{P}[\mathbf{X} = (2; 5)^T] = p_2 = \frac{3}{11}$ a $\mathbb{P}[\mathbf{X} = (3; 10)^T] = p_3 = \frac{7}{11}$, tedy $p_1 + p_2 + p_3 = 1$.



Obrázek 9.4: Graf rozdělení pravděpodobnosti $p(x, y)$ z úlohy 9.1

- c) $\mathbb{E}X = \frac{28}{11}$, $\mathbb{E}Y = \frac{87}{11}$, tj. $\mathbb{E}\mathbf{X} = \left(\frac{28}{11}, \frac{87}{11}\right)^T$.
d) $\text{var } X = \frac{52}{121}$, $\text{var } Y = \frac{1000}{121}$, $\text{cov}(X, Y) = \frac{226}{121}$,
tj. $\text{var } \mathbf{X} = \begin{pmatrix} \frac{52}{121} & \frac{226}{121} \\ \frac{226}{121} & \frac{1000}{121} \end{pmatrix}$.

9.2

- a) $\int_0^\infty \int_0^\infty c \cdot x \cdot y \cdot e^{-x^2} \cdot e^{-y^2} dx dy = \frac{c}{4} = 1$, z toho $c = 4$, za využití vzorce $\int x \cdot e^{kx^2} dx = \frac{1}{2k} \cdot e^{kx^2}$, zde pro $k = -1$.
b) $F(X, Y) = (1 - e^{-x^2}) \cdot (1 - e^{-y^2})$, $F_X(x) = 1 - e^{-x^2}$, $x \geq 0$,
 $F_Y(y) = 1 - e^{-y^2}$, $y \geq 0$.
c) $f_X(x) = 2 \cdot x \cdot e^{-x^2}$, $f_Y(y) = 2 \cdot y \cdot e^{-y^2}$.

- d) $\mathbb{E}X = 1, \mathbb{E}Y = 1$.
- e) $\text{var } \mathbf{X} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.
- f) $f_X(x) \cdot f_Y(y) = 4xye^{-x^2}e^{-y^2} = f(x, y)$ pro $x > 0, y > 0$ (jinak 0).
Náhodné veličiny X a Y jsou nezávislé.

9.3

- a) $\int_0^{2\pi} \int_0^\infty \frac{1}{2\pi\rho^2} \cdot e^{-\frac{r^2}{2\rho^2}} \cdot r \, dr \, d\varphi = 1$,
- b) distribuční funkce $G(r, \varphi) = \frac{\varphi}{2\pi} \cdot \left(1 - e^{-\frac{r^2}{2\rho^2}}\right)$,
- c) marginální hustoty $g_R(r) = \frac{r}{\rho^2} \cdot e^{-\frac{r^2}{2\rho^2}}$, $g_\Phi(\varphi) = \frac{1}{2\pi}$,
- d) protože $g(r, \varphi) = g_R(r) \cdot g_\Phi(\varphi)$, jsou R a Φ nezávislé náhodné veličiny.

9.4

- a) $c = \frac{2}{3e-5}$, hustota má tvar $f(x, y) = \frac{2}{3e-5} \cdot (x^2 + y) \cdot e^x$ pro $(x, y) \in [0; 1] \times [0; 1]$,
- b) $F(x, y) = \frac{2e^x y}{3e-5} \left(x^2 + \frac{y}{2} - 2x + 2\right)$ pro $(x, y) \in [0; 1] \times [0; 1]$,
- c) $f_X(x) = c \cdot e^x \cdot \left(x^2 + \frac{1}{2}\right)$ pro $x \in [0; 1]$, $f_Y(y) = c \cdot (2e - ye + y - 3)$ pro $y \in [0; 1]$.
- d) Střední hodnota:

$$\mathbb{E}X = \frac{-4e + 13}{3e - 5}, \mathbb{E}Y = \frac{4e - 7}{3 \cdot (3e - 5)},$$

$$\mathbb{E}\mathbf{X} = \left(\frac{-4e + 13}{3e - 5}; \frac{4e - 7}{3 \cdot (3e - 5)} \right)^T,$$

$$\mathbb{E}X^2 = \frac{-30e}{3e - 5}, \text{var } X = \frac{-106e^2 + 254e - 169}{(3e - 5)^2},$$

$$\mathbb{E}Y^2 = \frac{5e - 9}{6 \cdot (3e - 5)}, \text{var } Y = \frac{13e^2 - 24e + 37}{18 \cdot (3e - 5)},$$

$$\text{cov}(X, Y) = \frac{-56x^2 + 345e - 284}{3 \cdot (3e - 5)^2},$$

$$\text{var } \mathbf{X} = \begin{pmatrix} \frac{-106e^2 + 254e - 169}{(3e - 5)^2} & \frac{-56x^2 + 345e - 284}{3 \cdot (3e - 5)^2} \\ \frac{-56x^2 + 345e - 284}{3 \cdot (3e - 5)^2} & \frac{13e^2 - 24e + 37}{18 \cdot (3e - 5)} \end{pmatrix}.$$
- e) Náhodné veličiny X a Y nejsou nezávislé, mají nenulovou kovarianci.

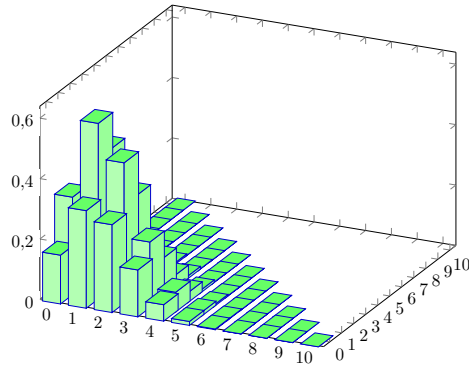
9.5 $\text{var } X = 1,1$.

9.6

- a) X nabývá některé z hodnot $0, 1, \dots, 10$, Y nabývá některé z hodnot $0, 1, \dots, 10$, přitom $x + y \leq 10$.

$$p(x, y) = \binom{10}{x} \binom{10-x}{y} \left(\frac{1}{6}\right)^{x+y} \left(\frac{5}{6}\right)^{10-x-y}$$

pro $0 \leq x \leq 10$, $0 \leq y \leq 10 - x$, v ostatních případech je nulová.



Obrázek 9.5: Graf sdružené pravděpodobnostní funkce k úloze 9.6

- b) $\mathbb{E}X = \frac{5}{3}$, $\mathbb{E}Y = \frac{5}{3}$, $\text{var } X = \frac{25}{18}$, $\text{var } Y = \frac{25}{18}$, $\text{cov}(X, Y) = 0$,
 $\text{var}(X, Y)^T = \begin{pmatrix} \frac{25}{18} & 0 \\ 0 & \frac{25}{18} \end{pmatrix}$.

9.7

a) $\int_{a_2}^{b_2} \int_{a_1}^{b_1} \frac{1}{(b_1 - a_1)(b_2 - a_2)} dx dy = 1.$

b) $F(x, y) = \begin{cases} 1, & x > b_1, y > b_2, \\ \frac{(x-a_1)(y-a_2)}{(b_1-a_1)(b_2-a_2)}, & a_1 \leq x \leq b_1, a_2 \leq y \leq b_2, \\ 0, & \text{jinde.} \end{cases}$

c) $\mathbb{E}X = \frac{a_1+b_1}{2}$, $\mathbb{E}Y = \frac{a_2+b_2}{2}$, $\text{var } X = \frac{(b_1-a_1)^2}{12}$, $\text{var } Y = \frac{(b_2-a_2)^2}{12}$,
 $\text{cov}(X, Y) = 0$, $\text{var}(X, Y)^T = \begin{pmatrix} \frac{(b_1-a_1)^2}{12} & 0 \\ 0 & \frac{(b_2-a_2)^2}{12} \end{pmatrix}$.

9.8 $c = \frac{4}{\pi}$

9.9 a) $\int_{-1}^1 \int_{-1}^1 \int_{-1}^1 \frac{1}{8}(1+x \cdot y \cdot z) \, dx \, dy \, dz = 1$. b) $f_1(x) = f_2(y) = f_3(z) = \frac{1}{2}$.
c) $f_{12}(x, y) = f_{13}(x, z) = f_{23}(y, z) = \frac{1}{4}$. d) X, Y, Z jsou nezávislé po dvou a nejsou vzájemně nezávislé.

9.10 Náhodný vektor $(X_1, X_2, X_3)^T$ má multinomické rozdělení s parametry n a $(p_1; p_2; p_3)$:

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \frac{n!}{x_1! x_2! x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3},$$

kde x_1, x_2, x_3 jsou nezáporná celá čísla splňující $x_1 + x_2 + x_3 = n$. Očekávaný počet neobsazených míst v i -té restauraci je $\mathbb{E}Y_i = n_i - np_i$.

9.11 $(X_1, X_2) \in \{(0; 0), (0; 1), (1; 0), (1; 1)\}$. $\mathbf{X} = (X_1, X_2)^T$ s oborem hodnot $\{(j, k) : j = 0, 1, k = 0, 1\}$, $\mathbb{P}[\mathbf{X} = (1, 1)^T] = \mathbb{P}[\mathbf{X} = (1, 0)^T] = \mathbb{P}[\mathbf{X} = (0, 1)^T] = \mathbb{P}[\mathbf{X} = (0, 0)^T] = \frac{1}{4}$. Distribuční funkce náhodného vektoru

$$F_{X_1, X_2}(x_1, x_2) = \begin{cases} 0 & \text{pro } (x_1, x_2) \in (-\infty; 0) \times (-\infty; 0) \\ 0 & \text{pro } (x_1, x_2) \in [0; 0] \times (-\infty; 0) \\ 0 & \text{pro } (x_1, x_2) \in (-\infty; 0) \times [0; 0] \\ \frac{1}{4} & \text{pro } (x_1, x_2) \in [0; 0] \times [0; 0] \\ \frac{1}{2} & \text{pro } (x_1, x_2) \in [0; 0] \times [0; 1] \\ \frac{1}{2} & \text{pro } (x_1, x_2) \in [0; 1] \times [0; 0] \\ 1 & \text{pro } (x_1, x_2) \in [0; 1] \times [0; 1] \end{cases}.$$

Náhodné veličiny X_1 a X_2 jsou nezávislé.

9.12 $\mathbb{P}(U = 1) = 0,6$, $\mathbb{P}(U = 2) = 0,4$, $\mathbb{P}(V = 1) = 0,3$, $\mathbb{P}(V = 2) = 0,3$, $\mathbb{P}(V = 3) = 0,4$. $\mathbb{E}U = 1,4$, $\mathbb{E}V = 2,1$. $\text{var } U = 0,24$, $\text{var } V = 0,69$. $\text{cov}(U, V) = -0,14$.

Kapitola 10

Funkce náhodných veličin

10.1 Funkce náhodných veličin

Při řešení některých pravděpodobnostních úloh se setkáváme se situací, kdy známe rozdělení náhodné veličiny X a hledáme rozdělení náhodné veličiny Y , která je funkcí veličiny X , tj. $Y = \phi(X)$. Následující věty udávají tvar distribučních funkcí, resp. hustot takových veličin.

Věta 10.1. *Nechť X je náhodná veličina s distribuční funkcí F a nechť $\phi: \mathbb{R} \rightarrow \mathbb{R}$. Označme $Y = \phi(X)$ a G její distribuční funkci. Pak pro $y \in \mathbb{R}$*

$$G(y) = \int_{\{x: \phi(x) \leq y\}} dF(x). \quad (10.1)$$

Speciálně, je-li F diskrétní $\{x_n, p_n\}$, je

$$G(y) = \sum_{\{x_n: \phi(x_n) \leq y\}} p_n,$$

a je-li absolutně spojitá s hustotou f , je

$$G(y) = \int_{\{x: \phi(x) \leq y\}} f(x) dx.$$

Důkaz. Označme $B_y = \{\omega : X(\omega) \leq y\}$; máme tyto rovnosti:

$$\begin{aligned} G(y) &= \mathbb{P}(Y \leq y) = \mathbb{P}(\phi(X) \leq y) = \mathbb{P}(X \in B_y) = \\ &= \int_{B_y} d\mu_F = \int_{\{x: \phi(x) \leq y\}} dF(x). \end{aligned}$$

□

Věta 10.2. *Nechť X je absolutně spojitá náhodná veličina s distribuční funkcí $F(x)$ a hustotou $f(x)$. Nechť t je ryze monotónní reálná funkce, která má derivaci všude. Položme $Y = t(X)$. Označme t^{-1} inverzní funkci k t . Pak Y má pro $y \in \mathbb{R}$ hustotu*

$$g(y) = f(t^{-1}(y)) |t^{-1}(y)'|.$$

Důkaz. Označme $G(y)$ distribuční funkci Y . Předpokládejme, že t je rostoucí. Pak platí

$$G(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(t(X) \leq y) = \mathbb{P}(X \leq t^{-1}(y)) = F(t^{-1}(y)).$$

Vidíme, že G je spojitá a má derivaci všude, až nanejvýš s výjimkou konečně mnoha bodů. Platí tedy

$$G'(y) = f(t^{-1}(y))(t^{-1}(y))' = g(y),$$

kde g je hustota veličiny Y . V případě, že t je klesající, je důkaz analogický. $\square$

Příklad 10.1.

- a) Nechť X je náhodná veličina s distribuční funkcí F , definujme $Y = a + bX$ pro $b \neq 0$, G budiž distribuční funkce Y . Je-li $b > 0$, pak $a + bX < Y$ je ekvivalentní s $X < \frac{Y-a}{b}$ a z (10.1) plyne, že

$$G(y) = \int_{\{x: x < \frac{y-a}{b}\}} dF(x) = F\left(\frac{y-a}{b}\right).$$

Naopak je-li $b < 0$, pak $a + bX < Y$ je ekvivalentní s $X > \frac{Y-a}{b}$ a dostáváme, že

$$G(y) = \int_{\{x: x > \frac{y-a}{b}\}} dF(x) = 1 - F\left(\frac{y-a}{b}\right).$$

Má-li F hustotu f , pak G má hustotu g , kterou najdeme derivováním předchozích rovnic. Dostaneme (pro $b > 0$ i $b < 0$) vztah

$$g(y) = \frac{1}{|b|} f\left(\frac{y-a}{b}\right).$$

- b) Má-li X normované normální rozdělení a $Y = \mu + \sigma X$, pak z a) plyne, že Y má normální rozdělení s parametry μ a σ^2 .

- c) Má-li X normální rozdělení s parametry μ, σ^2 a je-li $Y = a + bX$, pak z a) plyne, že Y má opět normální rozdělení s parametry $a + b\mu$ a $b^2\sigma^2$, neboť

$$g(y) = \frac{1}{\sqrt{2\pi}|b|\sigma} e^{-\frac{\left(\frac{y-a}{b}-\mu\right)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}|b|\sigma} e^{-\frac{(y-(a+b\mu))^2}{2\sigma^2 b^2}}.$$

- d) Je-li hustota f náhodné veličiny X sudá funkce, tj. je-li $f(x) = f(-x)$ pro každé x , pak náhodné veličiny X a $-X$ mají totéž rozdělení. Položme $Y = -X$, tj. $a = 0$ a $b = -1$, z (10.1) plyne, že

$$g(y) = f(-y), \quad \text{tj.} \quad g(y) = f(y),$$

jelikož je f sudá funkce.

- e) Je-li f hustota náhodné veličiny X sudá funkce, pak pro distribuční funkci F platí

$$F(x) = 1 - F(-x) \text{ pro libovolné } x \in \mathbb{R},$$

neboť

$$F(-x) = \int_{-\infty}^{-x} f(t) dt = \int_x^{\infty} f(t) dt = 1 - F(x).$$

- f) Necht X má normované normální rozdělení. Necht $Y = X^2$, pak Y má hustotu

$$g(y) = \begin{cases} 0 & \text{pro } y < 0, \\ \frac{1}{\sqrt{2\pi y}} e^{-\frac{y}{2}} & \text{pro } y \geq 0. \end{cases}$$

Jelikož dle (10.1) platí

$$\begin{aligned} G(y) &= \int_{\{x: \phi(x) < y\}} f(x) dx = \frac{1}{\sqrt{2\pi}} \int_{\{x: x^2 < y\}} e^{-\frac{x^2}{2}} dx \\ &= \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{y}} e^{-\frac{x^2}{2}} dx, \end{aligned}$$

položme substituci $x^2 = t$ a dostáváme

$$G(y) = \frac{1}{\sqrt{2\pi}} \int_0^y t^{-\frac{1}{2}} e^{-\frac{t}{2}} dt.$$

Odtud je zřejmé, že hustota $g(y)$ má výše uvedený tvar.

10.2 Funkce náhodného vektoru

Podobně jako v jednorozměrném případě (viz odstavec 5.6) definujeme Lebesgueovu–Stieltjesovu míru μ_F indukovanou distribuční funkcí F i pro více-rozměrný případ (tedy na prostoru $\mathbb{R}^n$) předpisem

$$\mu_F((\mathbf{a}, \mathbf{b}]) = \sum_{k=0}^n (-1)^k \sum_{\boldsymbol{\delta} \in \Delta_{k,n}} F(\boldsymbol{\delta})$$

pro polouzavřené intervaly $(\mathbf{a}; \mathbf{b}] = \times_{i=1}^n (a_i; b_i]$. Tuto definici lze dále jednoznačně rozšířit na všechny borelovské množiny $\mathcal{B}^n$. Integrál podle Lebesgueovy–Stieltjesovy míry z funkce $\phi(\mathbf{x})$ budeme stejně jako v jednorozměrném případě zapisovat

$$\int_{\mathbb{R}^n} \phi(\mathbf{x}) dF_{\mathbf{X}}(\mathbf{x}).$$

Příklad 10.2 (rozdělení minima a maxima). Necht X_i , $1 \leq i \leq n$, jsou nezávislé stejně rozdělené náhodné veličiny se spojitou distribuční funkcí F . Nalezněte distribuční funkce náhodných veličin

$$U = \max_{1 \leq i \leq n} X_i \quad \text{a} \quad V = \min_{1 \leq i \leq n} X_i.$$

Řešení. Z nezávislosti $X_1, \dots, X_n$ dostaneme podle věty 9.3

$$\begin{aligned} F_U(u) &= \mathbb{P}(U \leq u) = \mathbb{P}(X_1 \leq u, X_2 \leq u, \dots, X_n \leq u) = \\ &= \prod_{i=1}^n \mathbb{P}(X_i \leq u) = [F(u)]^n. \end{aligned}$$

$$\begin{aligned} F_V(v) &= \mathbb{P}(V \leq v) = 1 - \mathbb{P}(V > v) = \\ &= 1 - \mathbb{P}(X_1 > v, X_2 > v, \dots, X_n > v) = \\ &= 1 - \prod_{i=1}^n \mathbb{P}(X_i > v) = 1 - [1 - F(v)]^n. \end{aligned}$$

△

V případě, že k charakterizaci náhodné veličiny $Z = \phi(\mathbf{X})$ postačí spočítat její střední hodnotu, není potřeba odvozovat rozdělení Z , postačí (analogicky k jednorozměrnému případu ve větě 6.3) znát rozdělení náhodného vektoru $\mathbf{X}$.

Věta 10.3. *Nechť $\mathbf{X}$ je náhodný vektor a nechť $\phi: \mathbb{R}^n \rightarrow \mathbb{R}$ je reálná funkce. Pak platí*

$$\mathbb{E}\phi(\mathbf{X}) = \int_{\mathbb{R}^n} \phi(\mathbf{x}) \, dF_{\mathbf{X}}(\mathbf{x}),$$

pokud jeden z integrálů existuje. Speciálně

$$\mathbb{E}\phi(\mathbf{X}) = \sum_{m=1}^{\infty} \phi(\mathbf{x}_m) p_m$$

pro diskrétní rozdělení a

$$\mathbb{E}\phi(\mathbf{X}) = \int_{\mathbb{R}^n} \phi(\mathbf{x}) f(\mathbf{x}) \, d\mathbf{x}$$

pro absolutně spojitě rozdělení.

V případě nezávislosti dokonce postačí znát marginální rozdělení. Následující větu pro jednoduchost zformulujeme pouze pro dvojici náhodných veličin.

Věta 10.4. *Nechť X, Y jsou nezávislé náhodné veličiny s distribučními funkcemi F_X, F_Y a Lebesgueovými–Stieltjesovými měrami μ_{F_X}, μ_{F_Y} . Nechť $\phi: \mathbb{R}^2 \rightarrow \mathbb{R}$ je reálná funkce, pak pro náhodnou veličinu $Z = \phi(X, Y)$ platí*

$$\mathbb{E}Z = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \phi(x, y) \, d\mu_{F_Y}(y) \right) d\mu_{F_X}(x) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \phi(x, y) \, dF_Y(y) \right) dF_X(x),$$

je-li $\mathbb{E}|Z| < \infty$.

Důkaz. Z věty 9.3 plyne, že pro Lebesgueovy–Stieltjesovy míry nezávislých náhodných veličin platí

$$\mu_{F_{X,Y}} = \mu_{F_X} \cdot \mu_{F_Y}.$$

Odtud dostáváme

$$\mathbb{E}Z = \int_{\mathbb{R}^2} \phi(x, y) \, d\mu_{F_{X,Y}}(x, y) = \int_{\mathbb{R}^2} \phi(x, y) \, d\mu_{F_Y}(y) \, d\mu_{F_X}(x)$$

a podle Fubiniovy věty dostaneme tvrzení věty 10.4. □

Důsledek 10.5. *Distribuční funkce náhodné veličiny $Z = \phi(X, Y)$ je*

$$G(z) = \int_{-\infty}^{\infty} \left(\int_{\{y: \phi(x,y) \leq z\}} d\mu_{F_Y}(y) \right) d\mu_{F_X}(x).$$

Důkaz. Nahradíme-li ve větě 10.4 funkci $\phi(x, y)$ tzv. indikátorem $I_{\{(x,y) \in \mathbb{R}^2 : \phi(x,y) \leq z\}}(x, y)$, plyne tvrzení ihned. $\square$

Důsledek 10.6. *Jsou-li X, Y nezávislé náhodné veličiny s konečnými středními hodnotami, pak*

$$a) \mathbb{E}XY = (\mathbb{E}X)(\mathbb{E}Y).$$

Jsou-li navíc $\mathbb{E}X^2 < \infty$ a $\mathbb{E}Y^2 < \infty$, pak

$$b) \text{cov}(X, Y) = 0,$$

$$c) \text{var}(aX + bY) = a^2 \text{var} X + b^2 \text{var} Y \text{ pro libovolná } a, b \in \mathbb{R}.$$

Důkaz.

a) Uvažujeme-li $|x|I_{[|x| \leq n]}|y|I_{[|y| \leq n]}$, dostaneme pomocí Lévyho věty

$$\mathbb{E}|XY| = \mathbb{E}|X| \cdot \mathbb{E}|Y| < \infty.$$

Můžeme tedy použít větu 10.4, ve které dosadíme za $\phi(x, y) = x \cdot y$.

b) $(X - \mathbb{E}X), (Y - \mathbb{E}Y)$ jsou také nezávislé náhodné veličiny. Tudíž podle a) dostáváme

$$\text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] = \mathbb{E}(X - \mathbb{E}X)\mathbb{E}(Y - \mathbb{E}Y) = 0.$$

c) Podle b) dostáváme

$$\begin{aligned} \text{var}(aX + bY) &= \mathbb{E}[a(X - \mathbb{E}X) + b(Y - \mathbb{E}Y)]^2 = \\ &= a^2\mathbb{E}(X - \mathbb{E}X)^2 + b^2\mathbb{E}(Y - \mathbb{E}Y)^2 + 2ab\mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y) = \\ &= a^2 \text{var} X + b^2 \text{var} Y + 0. \end{aligned}$$

$\square$

Tvrzení b) postulují, že nezávislé náhodné veličiny jsou vždy nekorelované. Obrácené tvrzení však neplatí, nekorelované veličiny nemusí být nezávislé¹. Výjimku tvoří normálně rozdělené náhodné veličiny, pro něž platí ekvivalence mezi nekorelovaností a nezávislostí.

10.3 Konvoluce

Mějme dvě nezávislé náhodné veličiny X a Y s distribučními funkcemi $F(x)$ a $G(y)$. Zajímá nás rozdělení součtu $Z = X + Y$. Distribuční funkci náhodné

¹Nekorelovanost je nezávislost pouze „v lineárním pohledu“.

veličiny Z označme $H(z)$. Pak platí

$$\begin{aligned} H(z) &= \int \int_{x+y \leq z} dF(x) dG(y) = \int_{-\infty}^{\infty} F(z-y) dG(y) = \\ &= \int_{-\infty}^{\infty} G(z-x) dF(x). \end{aligned} \quad (10.2)$$

Rozdělení s distribuční funkcí $H(z)$ se nazývá konvoluce rozdělení s distribučními funkcemi $F(x)$ a $G(y)$. H se nazývá konvoluce distribučních funkcí F a G . Operaci konvoluce budeme značit $H = F * G$.

Operace konvoluce je zjevně komutativní a asociativní, neboť pro nezávislé náhodné veličiny X_1, X_2, X_3 platí

$$X_1 + X_2 = X_2 + X_1$$

a

$$X_1 + (X_2 + X_3) = (X_1 + X_2) + X_3.$$

Z toho plynou pro distribuční funkce

$$F_1 * F_2 = F_2 * F_1$$

a

$$F_1 * (F_2 * F_3) = (F_1 * F_2) * F_3.$$

Věta 10.7. *Nechť náhodné veličiny X a Y jsou nezávislé a mají absolutně spojitě distribuční funkce $F(x)$ a $G(y)$ s hustotami $f(x)$ a $g(y)$. Pak také konvoluce $H = F * G$ je absolutně spojitá a pro její hustotu $h(z)$ (tj. pro hustotu náhodné veličiny $Z = X + Y$) platí*

$$h(z) = \int_{-\infty}^{\infty} f(x)g(z-x) dx = \int_{-\infty}^{\infty} f(z-y)g(y) dy. \quad (10.3)$$

Důkaz. Formule (10.2) je za našich předpokladů totožná s

$$H(z) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^z f(x-y) dx \right) dG(y) = \int_{-\infty}^z \left(\int_{-\infty}^{\infty} f(x-y) dG(y) \right) dx.$$

Odtud derivováním dostaneme

$$h(z) = \int_{-\infty}^{\infty} f(z-y) dG(y)$$

a z toho plyne (10.3). □

Z důkazu je mimochodem vidět, že k absolutní spojitosti rozdělení náhodné veličiny $Z = X + Y$ ve skutečnosti stačí, aby bylo absolutně spojitě rozdělení jedné z náhodných veličin X a Y bez jakýchkoliv předpokladů o rozdělení druhé náhodné veličiny. Funkce $h(z)$ definovaná vztahem (10.3) se nazývá *konvoluce hustot* $f(x)$ a $g(y)$ a budeme ji značit $h = f * g$. Je to skutečně hustota, neboť z (10.3) plyne, že $h(z) \geq 0$ a

$$\begin{aligned} \int_{-\infty}^{\infty} h(z) dz &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x-y)g(y) dy dx = \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(x-y) dx \right) g(y) dy = \\ &= \int_{-\infty}^{\infty} 1 \cdot g(y) dy = 1. \end{aligned}$$

Diskrétní analogií vztahu (10.3) je následující věta.

Věta 10.8. *Nechť F, G jsou diskrétní distribuční funkce se skoky v přirozených číslech o velikosti $\{p_n\}, \{q_n\}$, tj.*

$$F(x) = \sum_{0 \leq n < x} p_n, \quad G(y) = \sum_{0 \leq n < y} q_n.$$

*Nechť $H = F * G$. Pak H je diskrétní distribuční funkce se skoky v přirozených číslech a platí*

$$H(z) = \sum_{0 \leq n < z} h_n, \quad \text{kde } h_n = \sum_{k=0}^n p_k q_{n-k}.$$

Důkaz. Věta plyne přímo po aplikaci věty o úplné pravděpodobnosti. □

Nyní si uvedeme některé příklady konvoluce rozdělení.

1) Konvoluce rovnoměrných rozdělení

Nechť

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b, \\ 0 & \text{jinak} \end{cases}$$

a

$$g(y) = \begin{cases} \frac{1}{d-c} & c \leq y \leq d, \\ 0 & \text{jinak.} \end{cases}$$

Předpokládejme, že $d - c \geq b - a$. Pak pro konvoluci $h(z)$ hustot náhodných veličin s hustotami $f(x)$ a $g(y)$ platí

$$h(z) = \begin{cases} 0 & z \leq a + c \text{ nebo } b + d \leq z, \\ \frac{z - (a + c)}{(b - a)(d - c)} & a + c \leq z \leq b + c, \\ \frac{1}{d - c} & b + c \leq z \leq a + d, \\ \frac{(b + d) - z}{(b - a)(d - c)} & a + d \leq z \leq b + d. \end{cases}$$

Grafem je lichoběžník se základnou v ose x . Vidíme, že hustota $h(z)$ je všude spojitá, ačkoliv $f(x)$ a $g(y)$ mají body nespojitosti (konvoluce „vyhlazuje“ nespojitosti). Ve speciálním případě, kdy obě náhodné veličiny X a Y mají stejné rozdělení (tj. $a = c$, $b = d$), má hustota $h(x)$ tvar trojúhelníku; toto rozdělení se nazývá *Simpsonovo rozdělení*.

2) Konvoluce binomických rozdělení

Nechť X_1, X_2 jsou nezávislé náhodné veličiny s rozdělením $\text{Bi}(n_1; p)$ a $\text{Bi}(n_2; p)$. Pak náhodná veličina $Y = X_1 + X_2$ má binomické rozdělení s parametry $(n_1 + n_2; p)$.

3) Konvoluce Poissonových rozdělení

Nechť X_1 má $\text{Po}(\lambda_1)$ a X_2 má $\text{Po}(\lambda_2)$. Předpokládejme navíc, že X_1 a X_2 jsou nezávislé. Pak náhodná veličina $Y = X_1 + X_2$ má Poissonovo rozdělení s parametrem $\lambda_1 + \lambda_2$.

4) Konvoluce normálních rozdělení

Nechť X_1, X_2 jsou nezávislé náhodné veličiny, X_1 má $\text{N}(\mu_1; \sigma_1^2)$ a X_2 má $\text{N}(\mu_2; \sigma_2^2)$. Pak $Y = X_1 + X_2$ má rozdělení $\text{N}(\mu_1 + \mu_2; \sigma_1^2 + \sigma_2^2)$.

5) Konvoluce exponenciálních rozdělení

Jsou-li X_1, X_2 nezávislé náhodné veličiny s týmž exponenciálním rozdělením s parametrem $\lambda > 0$, pak náhodná veličina $Y = X_1 + X_2$ má rozdělení $\text{Erlang}(2, \lambda)$.

Příklad 10.3 (Rozdělení průměru $\bar{X}$). Uvažujme hypotetickou populaci lučních květin, jejichž výška se řídí normálním rozdělením $\text{N}(20; 25)$. Vypočítejte pravděpodobnost, že průměrná výška spočtená z 20 náhodně vybraných květin se bude od 20 lišit o více než 3 cm.

Řešení. Máme $X_1, \dots, X_{20}$ nezávislých stejně rozdělených náhodných veličin s rozdělením $X_i \sim \text{N}(20; 25)$. Konvoluce normálních rozdělení nám dá

$$\sum_{i=1}^{20} X_i \sim \text{N}(20 \cdot 20; 20 \cdot 25).$$

a po jednoduché lineární transformaci $\bar{X} = \frac{1}{20} \cdot \sum_{i=1}^{20} X_i$ dostáváme, že

$$\bar{X} \sim \text{N}\left(20; \frac{25}{20}\right).$$

Hledaná pravděpodobnost je tedy $\mathbb{P}(\bar{X} < 17) + \mathbb{P}(\bar{X} > 23)$. Standardizujeme-li $\bar{X}$ na veličinu Z s normovaným normálním rozdělením, dostaneme

$$\begin{aligned} & \mathbb{P}\left(\frac{\bar{X} - 20}{\sqrt{\frac{5}{4}}} < \frac{17 - 20}{\sqrt{\frac{5}{4}}}\right) + \mathbb{P}\left(\frac{\bar{X} - 20}{\sqrt{\frac{5}{4}}} > \frac{23 - 20}{\sqrt{\frac{5}{4}}}\right) = \\ &= \mathbb{P}\left(Z < \frac{-3}{\sqrt{\frac{5}{4}}}\right) + \mathbb{P}\left(Z > \frac{3}{\sqrt{\frac{5}{4}}}\right) = \Phi\left(\frac{-3}{\sqrt{\frac{5}{4}}}\right) + 1 - \Phi\left(\frac{3}{\sqrt{\frac{5}{4}}}\right) = \\ &= 2 \cdot \left(1 - \Phi\left(\frac{3}{\sqrt{\frac{5}{4}}}\right)\right) \doteq 2 \cdot (1 - 0,996) \doteq 0,008. \end{aligned}$$

△

Příklad 10.4. Samička určitého hmyzu naklade r vajíček s pravděpodobností $p_r = \frac{e^{-\lambda} \lambda^r}{r!}$. Z každého vajíčka se vylíhne živý jedinec s pravděpodobností p , která je stejná pro všechna vajíčka. Osudy jednotlivých vajíček jsou nezávislé. Zajímá nás pravděpodobnost, s níž dá samička život právě k novým jedincům.

Řešení. Necht $N \sim \text{Po}(\lambda)$ je náhodná veličina udávající počet vajíček a X je náhodná veličina udávající počet vylíhnutých jedinců. Víme, že za podmínky, že známe N , má $X \sim \text{Bi}(N; p)$. Tedy $\mathbb{P}(X = k)$ spočteme podle věty o úplné pravděpodobnosti

$$\begin{aligned} \mathbb{P}(X = k) &= \sum_{n=k}^{\infty} \mathbb{P}(X = k | N = n) \mathbb{P}(N = n) = \\ &= \sum_{n=k}^{\infty} \binom{n}{k} p^k (1-p)^{n-k} \cdot e^{-\lambda} \frac{\lambda^n}{n!} = \\ &= e^{-\lambda} p^k \cdot \sum_{n=k}^{\infty} \frac{n!}{k!(n-k)!} \cdot (1-p)^{n-k} \frac{\lambda^n}{n!} = \\ &= \frac{e^{-\lambda} p^k \cdot \lambda^k}{k!} \cdot \sum_{n=k}^{\infty} \frac{(1-p)^{n-k} \lambda^{n-k}}{(n-k)!} = \\ &= \frac{e^{-\lambda} p^k \cdot \lambda^k}{k!} \cdot \sum_{n=0}^{\infty} \frac{((1-p) \cdot \lambda)^n}{n!} = \\ &= \frac{e^{-\lambda} p^k \cdot \lambda^k}{k!} \cdot e^{(1-p) \cdot \lambda} = e^{-\lambda p} \cdot \frac{(\lambda p)^k}{k!}. \end{aligned}$$

Neboli $X \sim \text{Po}(\lambda p)$.

△

Příklad 10.5. Spočítejte kovarianci náhodných veličin X a $Y = X^2$, kde X má rovnoměrné rozdělení na intervalu $[-1; 1]$. Jsou tyto náhodné veličiny nezávislé?

Řešení.

$$\text{cov}(X, Y) = \mathbb{E}XY - \mathbb{E}X \cdot \mathbb{E}Y = \mathbb{E}X^3 - \mathbb{E}X \cdot \mathbb{E}X^2$$

$$X \sim U[-1; 1] \Rightarrow \mathbb{E}X = 0$$

$$\mathbb{E}X^3 = \int_{-1}^1 x^3 \cdot \frac{1}{2} dx = 0.$$

Odtud dostáváme, že $\text{cov}(X, Y) = 0$, ale náhodné veličiny X a $Y = X^2$ zjevně nejsou nezávislé. $\triangle$

Příklad 10.6. Náhodná veličina Y je funkcí absolutně spojitě náhodné veličiny X . Čemu se rovná hustota $g(y)$, jestliže

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad -\infty < x < \infty \quad \text{a} \quad Y = X^2?$$

Řešení.

$$t(x) = x^2, \quad -\infty < x < \infty.$$

V tomto intervalu není $t(x)$ monotónní. Distribuční funkce náhodné veličiny Y je obecně tvaru

$$\begin{aligned} G(y) &= \mathbb{P}(Y \leq y) = \mathbb{P}(X^2 \leq y) = \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) = \\ &= \mathbb{P}(X \leq \sqrt{y}) - \mathbb{P}(X \leq -\sqrt{y}) = F(\sqrt{y}) - F(-\sqrt{y}). \end{aligned}$$

Odtud pro hustotu $g(y)$ dostáváme

$$g(y) = f(\sqrt{y}) \frac{1}{2\sqrt{y}} - f(-\sqrt{y}) \left(-\frac{1}{2\sqrt{y}}\right) = \frac{1}{2\sqrt{y}} [f(\sqrt{y}) + f(-\sqrt{y})].$$

Speciálně pro $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ platí, že

$$g(y) = \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{2\sqrt{y}} \left(e^{-\frac{y}{2}} + e^{-\frac{y}{2}} \right) = \begin{cases} \frac{1}{\sqrt{2\pi y}} e^{-\frac{y}{2}} & y > 0, \\ 0 & \text{jinak.} \end{cases}$$

Dostali jsme hustotu rozdělení χ^2 o jednom stupni volnosti, která se používá v matematické statistice [13]. $\triangle$

10.4 Úlohy

10.1 Náhodná veličina Y je funkcí absolutně spojitě náhodné veličiny X . Čemu se rovná hustota pravděpodobnosti $g(y)$, jestliže

- a) $Y = 8X^3$ a $f(x) = \begin{cases} 2x & 0 < x < 1, \\ 0 & \text{jinak,} \end{cases}$
- b) $Y = -2 \ln X$ a $f(x) = \begin{cases} 1 & 0 < x < 1, \\ 0 & \text{jinak,} \end{cases}$
- c) $Y = X^3$ a $f(x) = \begin{cases} \frac{x^2}{9} & 0 < x < 3, \\ 0 & \text{jinak,} \end{cases}$
- d) $Y = X^2$ a $f(x) = \begin{cases} xe^{-x^2} & x > 0, \\ 0 & \text{jinak,} \end{cases}$
- e) $Y = e^X$ a $f(x) = \begin{cases} 1 & 0 < x < 1, \\ 0 & \text{jinak,} \end{cases}$
- f) $Y = |X|$ a $f(x)$ je libovolné $-\infty < x < \infty$,
- g) $Y = \sin X$ a $f(x) = \begin{cases} \frac{1}{\pi} & -\frac{\pi}{2} < x < \frac{\pi}{2}, \\ 0 & \text{jinak,} \end{cases}$
- h) $Y = |1 - X|$ a $f(x) = \begin{cases} \frac{1}{2} & 0 < x < 2, \\ 0 & \text{jinak.} \end{cases}$

10.2 Najděte rozdělení náhodné veličiny Y , jestliže

- a) $Y = 2X + 1$ a

$$p_n = \begin{cases} \frac{1}{3} & n = 1, 2, 3, \\ 0 & \text{jinak,} \end{cases}$$

- b) $Y = X^3$

$$p_n = \begin{cases} \left(\frac{1}{2}\right)^n & n = 1, 2, 3, \dots, \\ 0 & \text{jinak.} \end{cases}$$

- c) X, Y jsou nezávislé náhodné veličiny, X má exponenciální rozdělení s parametrem $\lambda > 0$, Y má rovnoměrné rozdělení na $(0; \theta)$, $\theta > 0$.
- (a) Určete rozdělení $X + Y$.
 - (b) Určete $\mathbb{E}(X + Y)$.
 - (c) Určete $\text{var}(X + Y)$.

10.3 Necht X_i , $1 \leq i \leq n$ jsou nezávislé náhodné veličiny, nabývající hodnot ± 1 s pravděpodobnostmi $\mathbb{P}(X_i = 1) = p$, $\mathbb{P}(X_i = -1) = 1 - p$. Najděte rozdělení náhodné veličiny $S_n = \sum_{i=1}^n X_i$.

10.4 Házíme třemi korunovými mincemi a čtyřmi pětikorunovými mincemi. Necht X je celková hodnota těch korunových mincí, na nichž padl líc. Podobně necht Y je celková hodnota těch pětikorunových mincí, na nichž padl líc. Zaveďme $W = X + Y$. Spočítejte kovarianci náhodných veličin X a W .



10.5 Necht $X \sim U[1; 2]$. Určete kovarianci náhodných veličin X , $\frac{1}{X}$.

10.6 Necht $(X_1, \dots, X_n)^T$ je posloupnost nezávislých náhodných veličin s rovnoměrným rozdělením na $(0; 1)$. Označme

$$U = \max_{1 \leq i \leq n} X_i, \quad V = \min_{1 \leq i \leq n} X_i.$$

- Stanovte distribuční funkce a hustoty náhodných veličin U a V .
- Určete $\mathbb{E}U$, $\text{var } U$, $\mathbb{E}V$, $\text{var } V$.

10.7 Náhodné veličiny X , Y jsou nezávislé a mají obě stejné exponenciální rozdělení

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{pro } x \geq 0, \\ 0 & \text{pro } x < 0. \end{cases}$$

Najděte distribuční funkci náhodné veličiny $Z = \max(X, Y)$ a její střední hodnotu. Jsou náhodné veličiny $\max(X, Y)$ a $\min(X, Y)$ nezávislé?

10.8 Necht náhodný vektor $Y = (X_1, X_2)^T$ má sdruženou hustotu

$$f_Y(x_1, x_2) = \frac{1}{2\pi\sigma^2} \cdot e^{-\frac{(x_1^2 + x_2^2)}{2\sigma^2}}.$$

Stanovte hustotu náhodné veličiny $Z = X_1 + X_2$.

Řešení úloh

10.1

- a) $g(y) = \begin{cases} \frac{1}{6}y^{-\frac{1}{3}} & 0 < y < 8, \\ 0 & \text{jinak,} \end{cases};$
- b) $g(y) = \begin{cases} \frac{1}{2}e^{-\frac{y}{2}} & y > 0, \\ 0 & \text{jinak,} \end{cases}$
- c) $g(y) = \begin{cases} \frac{1}{27} & 0 < y < 27, \\ 0 & \text{jinak,} \end{cases}$
- d) $g(y) = \begin{cases} e^{-y} & y \geq 0, \\ 0 & \text{jinak,} \end{cases}$
- e) $g(y) = \begin{cases} \frac{1}{y} & 1 < y < e, \\ 0 & \text{jinak,} \end{cases}$
- f) $g(y) = \begin{cases} f(y) + f(-y) & y \geq 0, \\ 0 & y < 0, \end{cases}$
- g) $g(y) = \begin{cases} \frac{1}{\pi\sqrt{1-y^2}} & -1 < y < 1, \\ 0 & \text{jinak,} \end{cases}$
- h) $g(y) = \begin{cases} 1 & 0 \leq y \leq 1, \\ 0 & \text{jinak.} \end{cases}$

10.2

- a) $\mathbb{P}(Y = y) = \begin{cases} \frac{1}{3} & y = 3, 5, 7, \\ 0 & \text{jinak.} \end{cases}$
- b) $\mathbb{P}(Y = y) = \begin{cases} \left(\frac{1}{2}\right)^n & y = n^3, n = 1, 2, 3, \dots, \\ 0 & \text{jinak.} \end{cases}$
- c) $f_Z(z) = \frac{1}{\theta} \cdot e^{-z\lambda} \cdot (e^{\theta\lambda} - 1), z \geq 0,$
 $\mathbb{E}(X + Y) = \frac{1}{\lambda} + \frac{\theta}{2},$
 $\text{var}(X + Y) = \frac{1}{\lambda^2} + \frac{\theta^2}{12}.$

10.3 Náhodná veličina S_n má rozdělení definované pravděpodobností

$$\mathbb{P}(S_n = s) = \binom{n}{\frac{s+n}{2}} p^{\frac{s+n}{2}} (1-p)^{n-\frac{s+n}{2}},$$

kde $s = \{-n, -n+2, -n+4, \dots, n\}$ a $\frac{s+n}{2}$ je celé číslo.

10.4 $\text{cov}(X, W) = 0,75$

10.5 $\text{cov}(X, \frac{1}{X}) = 1 - \frac{3}{2} \ln 2$

10.6

$$F_U(u) = \begin{cases} 0, & u \leq 0, \\ u^n, & u \in (0; 1), \\ 1, & u \geq 1, \end{cases} \quad f_U(u) = \begin{cases} nu^{n-1}, & u \in (0; 1), \\ 0, & u \notin (0; 1), \end{cases}$$

$$F_V(v) = \begin{cases} 0, & v \leq 0, \\ 1 - (1 - v)^n, & v \in (0; 1), \\ 1, & v \geq 1, \end{cases} \quad f_V(v) = \begin{cases} n(1 - v)^{n-1}, & v \in (0; 1), \\ 0, & v \notin (0; 1). \end{cases}$$

$$\mathbb{E}U = \frac{n}{n+1}, \mathbb{E}V = \frac{1}{n+1}, \text{var } U = \frac{n}{n+2} - \left(\frac{n}{n+1}\right)^2, \text{var } V = \frac{1}{(n+1)(n+2)}.$$

10.7 $F_Z(z) = \mathbb{P}(Z \leq z) = \mathbb{P}(\max(X, Y) \leq z) = \begin{cases} 0 & \text{pro } z < 0, \\ (1 - e^{-\lambda z})^2 & \text{pro } z \geq 0. \end{cases}$

$\mathbb{E}Z = \frac{3}{2\lambda}$. Náhodné veličiny $\max(X, Y)$ a $\min(X, Y)$ jsou nezávislé.

10.8 $f(z) = \frac{1}{2\sqrt{\pi\sigma^2}} \cdot e^{-\frac{z^2}{4\sigma^2}}, z \in \mathbb{R}.$

Kapitola 11

Zákon velkých čísel, centrální limitní věta

11.1 Konvergence náhodných veličin

Až dosud jsme se zabývali náhodnou veličinou a jejími charakteristikami za předpokladu, že tyto charakteristiky dokonale známe. To ovšem není příliš obvyklý případ; častější je případ, kdy náhodnou veličinou popisujeme nějakou vlastnost populace, ale její charakteristiky neznáme – hovoříme o teoretických vlastnostech populace, resp. náhodné veličiny (teoretickém rozdělení, teoretické střední hodnotě atd.). Jestliže však máme možnost nezávisle opakovat nějaký náhodný pokus, související s teoretickou náhodnou veličinou, můžeme z pozorovaných hodnot sestavit rozdělení relativních četností a informace o tomto rozdělení shrnout opět do charakteristik. Toto rozdělení a jeho charakteristiky nazveme empirickým rozdělením a empirickými charakteristikami. Například střední hodnotu nějakého teoretického rozdělení odhadujeme tak, že uskutečníme náhodný výběr z tohoto rozdělení – posloupnost nezávislých náhodných pokusů, jejichž výsledkem jsou nezávislé náhodné veličiny se stejným rozdělením, jako má teoretická náhodná veličina – a jako odhad teoretické střední hodnoty použijeme empirickou charakteristiku aritmetický průměr.

Při dodržování jistých podmínek můžeme očekávat, že empirické rozdělení (resp. jeho charakteristika) se bude blížit k teoretickému rozdělení (resp. k teoretické charakteristice), a to tím více, čím větší bude rozsah realizovaných pokusů. To je obecně vyjádřeno tzv. zákonem velkých čísel. Zde je však na místě poznamenat, že přibližování empirických hodnot k teoretickým hodnotám nemá charakter obvyklé matematické konvergence, ale konvergence pravděpodobnostního typu. Pravděpodobnostní konvergenčí rozu-

míme skutečnost, že při vzrůstajícím počtu pokusů se pravděpodobnost velkých odchylek empirických hodnot od teoretických stále zmenšuje. Typů pravděpodobnostní konvergence je ovšem více a mají různou sílu. Ty nejpožívanější typy zavedeme v následující definici. Poznamenejme ještě úvodem, že některá tvrzení v této kapitole uvedeme bez důkazu. Důkazy lze nalézt například v [8].

Definice 11.1 (Konvergence náhodných veličin)

Mějme posloupnost náhodných veličin $X_1, X_2, X_3, \dots$ a náhodnou veličinu X . Necht' jsou všechny tyto veličiny definovány na témže pravděpodobnostním prostoru $(\Omega, \mathcal{A}, \mathbb{P})$.

a) *Ríkáme, že X_n konverguje k X skoro jistě, jestliže*

$$\mathbb{P}\{\omega : X_n(\omega) \xrightarrow[n \rightarrow \infty]{} X(\omega)\} = 1.$$

b) *Jestliže pro každé $\varepsilon > 0$ platí*

$$\mathbb{P}\{\omega : |X_n(\omega) - X(\omega)| > \varepsilon\} \xrightarrow[n \rightarrow \infty]{} 0,$$

pak říkáme, že X_n konverguje k X podle pravděpodobnosti.

c) *Pokud*

$$\mathbb{E}X_n^2 < \infty, \quad n = 1, 2, 3, \dots,$$

a jestliže

$$\mathbb{E}(X_n - X)^2 \xrightarrow[n \rightarrow \infty]{} 0,$$

pak říkáme, že X_n konverguje k X podle středu stupně 2.

Následující věta ukazuje vztah mezi jednotlivými konvergencemi.

Věta 11.1 (o vztahu konvergencí). *Uvažujme konvergenci posloupnosti náhodných veličin skoro jistě a konvergenci podle pravděpodobnosti. Pak platí tvrzení:*

- a) *Z konvergence skoro jistě plyne konvergence podle pravděpodobnosti.*
- b) *Z konvergence podle středu stupně 2 plyne konvergence podle pravděpodobnosti.*

Tvrzení věty 11.1 nelze bez dodatečných předpokladů zesílit. Žádnou implikaci nelze obrátit.

Lemma 11.2. *Pro libovolnou nezápornou veličinu X se střední hodnotou $\mathbb{E}X$ platí:*

$$\mathbb{P}(X \geq \varepsilon) \leq \frac{\mathbb{E}X}{\varepsilon}.$$

Důkaz.

$$\mathbb{P}(X \geq \varepsilon) = \int_{x \geq \varepsilon} dF_X(x) \leq \int_{x \geq \varepsilon} \frac{x}{\varepsilon} dF_X(x) \leq \frac{1}{\varepsilon} \int x dF_X(x) = \frac{1}{\varepsilon} \mathbb{E}X.$$

□

Lemma 11.2 má význam i pro větší ε , nikoli jen pro velmi malé ε , jak jsme zvyklí z matematické analýzy.

Příklad 11.1. Mějme náhodnou veličinu X , která má $\mathbb{E}X = 3$ a pro které platí $\mathbb{E}X^2 = 13$. Odhadněte pravděpodobnost, že veličina X nabude hodnoty z intervalu $(-2; 8)$.

Řešení. Střední hodnota $\mathbb{E}X = 3$. Určíme rozptyl veličiny X :

$$\text{var } X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = 13 - 3^2 = 4.$$

Hledanou pravděpodobnost odhadneme pomocí Čebyševovy nerovnosti:

$$\mathbb{P}(-2 < X < 8) = \mathbb{P}(|X - \mathbb{E}X| < 5) \geq 1 - \frac{4}{25} \doteq 0,84.$$

△

11.2 Zákon velkých čísel

Mezi klasické výsledky teorie pravděpodobnosti patří chování aritmetického průměru, jestliže počet sčítanců v jeho definici roste nade všechny meze. Tyto vlastnosti jsou popisovány zákony velkých čísel.

Věta 11.3 (slabý zákon velkých čísel). *Nechť $X_1, X_2, \dots$ jsou nezávislé náhodné veličiny se stejnými středními hodnotami μ a stejnými rozptyly $\sigma^2 < \infty$. Pak pro $n \rightarrow \infty$ platí*

$$\frac{1}{n}(X_1 + X_2 + \dots + X_n) \rightarrow \mu$$

podle pravděpodobnosti.

Důkaz. Označme

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \dots + X_n).$$

Vypočtěme střední hodnotu a rozptyl veličiny $\bar{X}_n$:

$$\begin{aligned}\mathbb{E}\bar{X}_n &= \mathbb{E}\left[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right] = \frac{1}{n}[\mathbb{E}X_1 + \mathbb{E}X_2 + \cdots + \mathbb{E}X_n] = \\ &= \frac{1}{n}n\mu = \mu, \\ \text{var } \bar{X}_n &= \text{var}\left[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right] = \\ &= \frac{1}{n^2}[\text{var } X_1 + \text{var } X_2 + \cdots + \text{var } X_n] = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.\end{aligned}$$

Nyní využijeme Čebyševovy nerovnosti (věta 6.5) pro veličinu $\bar{X}_n$ a dostáváme, že pro každé $\varepsilon > 0$ platí

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

Odtud pro $n \rightarrow \infty$ máme $\frac{\sigma^2}{n\varepsilon^2} \rightarrow 0$, takže také

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \varepsilon) \rightarrow 0.$$

□

Tato verze zákona velkých čísel patří mezi historicky nejstarší. V principu říká, že jestliže se zajímáme o odhad teoretické střední hodnoty μ nějaké náhodné veličiny, můžeme se k němu přiblížit libovolně blízko pomocí aritmetického průměru dostatečně velkého počtu náhodných veličin se stejnou střední hodnotou a stejným rozptylem. Tvzení věty 11.3 se dá dále zesílit. Jeden z výsledků je uveden v následující větě. Další zobecnění například pro náhodné vektory s různými středními hodnotami můžeme nalézt v [8].

Věta 11.4 (silný zákon velkých čísel). *Nechť $\{X_n\}_{n=1}^\infty$ je posloupnost nezávislých stejně rozdělených náhodných veličin se střední hodnotou $\mathbb{E}X_1 = \mu$ a nechť $\mathbb{E}|X_1| < \infty$. Pak*

$$\lim_{n \rightarrow \infty} \bar{X}_n = \mu$$

v pravděpodobnosti a skoro jistě.

11.3 Centrální limitní věta

Podstatou centrální limitní věty (dále CLV) je tvrzení, že náhodná veličina Y_n , která vznikla jako součet velkého počtu vzájemně nezávislých náhodných

veličin $X_1, X_2, \dots, X_n$, má za velmi obecných podmínek přibližně normální rozdělení. Budeme říkat, že náhodná veličina Y_n , jejímž limitním rozdělením je rozdělení normální, má tzv. asymptoticky normální rozdělení.

Nejstarší a nejjednodušší případ CLV je věta de Moivreova¹-Laplaceova. Náhodnou veličinou Y_n je součet n vzájemně nezávislých náhodných veličin, z nichž každá má alternativní rozdělení s parametrem p . Pak náhodná veličina Y_n má rozdělení $\text{Bi}(n; p)$ se střední hodnotou $\mathbb{E}Y_n = np$ a rozptylem $\text{var } Y_n = np(1-p)$. De Moivreova-Laplaceova věta tvrdí, že pro normovanou náhodnou veličinu

$$Z_n = \frac{Y_n - np}{\sqrt{np(1-p)}}$$

platí limitní vztah

$$\lim_{n \rightarrow \infty} \mathbb{P}(Z_n \leq x) = \Phi(x), \quad -\infty < x < \infty,$$

kde $\Phi(x)$ je distribuční funkce rozdělení $N(0; 1)$. Tedy věta de Moivreova-Laplaceova říká, že při dostatečně velkém počtu nezávislých pokusů lze binomické rozdělení aproximovat rozdělením normálním.

Věta 11.5 (De Moivreova-Laplaceova). *Nechť pro libovolné $n \geq 1$ je Y_n náhodná veličina s rozdělením $\text{Bi}(n; p)$, $0 < p < 1$. Položme*

$$Z_n = \frac{Y_n - np}{\sqrt{np(1-p)}}$$

a označme $F_n(x)$ distribuční funkci náhodné veličiny Z_n . Pak

$$\lim_{n \rightarrow \infty} F_n(x) = \Phi(x), \quad -\infty < x < \infty. \quad (11.1)$$

Příklad 11.2. Jaká je pravděpodobnost, že ze 120 hodů kostkou padne alespoň 14 šestek?

Řešení. Označme $X_i \sim A\left(\frac{1}{6}\right)$ náhodnou veličinu, která představuje to, zda nám v i -tém hodu kostkou padne či nepadne šestka. Pro X_i platí: $\mathbb{E}X_i = \frac{1}{6}$, $\sigma^2 = \frac{5}{36}$. Chceme tudíž vypočítat

$$\mathbb{P}\left(\sum_{i=1}^{120} X_i \geq 14\right).$$

¹Abraham de Moivre (1667–1754), francouzský matematik.

Tento příklad vypočítáme nejprve přesně. Náhodná veličina $Y_{120} = \sum_{i=1}^{120} X_i$ má binomické rozdělení $\text{Bi}(120; \frac{1}{6})$. Pomocí vzorce (7.2) pro pravděpodobnostní funkci binomického rozdělení (anebo pomocí funkce Excelu pro binomické rozdělení) spočteme, že

$$\mathbb{P}(Y_{120} \geq 14) = 1 - \sum_{k=0}^{13} \mathbb{P}(Y_{120} = k) = 1 - \sum_{k=0}^{13} \binom{120}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{120-k} \doteq 0,950.$$

= 1-BINOM.DIST(13;120;1/6;PRAVDA)

Nyní spočteme tuto pravděpodobnost přibližně pomocí aproximace CLV. Použití CLV spočívá v transformaci veličiny Y_{120} na veličinu Z_{120} podle věty 11.5. Pišme tedy

$$\begin{aligned} \mathbb{P}\left(\sum_{i=1}^{120} X_i \geq 14\right) &= \mathbb{P}\left(\frac{\sum_{i=1}^{120} X_i - 120 \cdot \frac{1}{6}}{\sqrt{120 \cdot \frac{1}{6} \cdot \frac{5}{6}}} \geq \frac{14 - 120 \cdot \frac{1}{6}}{\sqrt{120 \cdot \frac{1}{6} \cdot \frac{5}{6}}}\right) \doteq \\ &\doteq \mathbb{P}(Z_{120} \geq -1,470) = 1 - \mathbb{P}(Z_{120} \leq -1,470). \end{aligned}$$

Náhodná veličina Z_{120} má podle CLV asymptoticky normované normální rozdělení, můžeme tedy konečně psát, že

$$\mathbb{P}\left(\sum_{i=1}^{120} X_i \geq 14\right) \doteq 1 - \Phi(-1,470) \doteq 0,929.$$

= 1-NORM.S.DIST(-1,470;PRAVDA)

Hledaná pravděpodobnost je přibližně (po aproximaci CLV a po zaokrouhlení) 0,929. Vidíme, že odchylka od přesně spočítané pravděpodobnosti činí pouze asi dva procentní body. $\triangle$

Příklad 11.3. Kolikrát musíme hodit kostkou, aby pravděpodobnost, že padne alespoň deset šestek, byla větší nebo rovna 0,95?

Řešení. Obdobně jako v příkladu 11.2 označme $X_i \sim A(\frac{1}{6})$ náhodnou veličinu, která představuje to, zda nám v i -tém hodu kostkou padne šestka či nikoliv. Problém můžeme přepsat na nerovnici

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq 10\right) \geq 0,95, \quad (11.2)$$

kde neznámá je n , tj. počet hodů kostkou. Tuto úlohu lze řešit hrubou silou: pro rostoucí hodnotu n napočítávat binomické pravděpodobnosti na levé

straně nerovnice (11.2) do té doby, dokud některá z nich nepřekročí hodnotu 0,95. Elegantnější a rychlejší způsob je hodnotu n odhadnout pomocí CLV. To opět spočívá v úpravě nerovnice do podoby, ve které se nachází náhodná veličina asymptoticky se blížící veličině s normálním rozdělením:

$$\mathbb{P} \left(\frac{\sum_{i=1}^n X_i - \frac{n}{6}}{\sqrt{\frac{5n}{36}}} \geq \frac{10 - \frac{n}{6}}{\sqrt{\frac{5n}{36}}} \right) \geq 0,95. \quad (11.3)$$

Takto je ovšem rovněž určena hodnota $\frac{10 - \frac{n}{6}}{\sqrt{\frac{5n}{36}}}$ jako 5% kvantil² náhodné ve-

ličiny $Z_n = \frac{\sum_{i=1}^n X_i - \frac{n}{6}}{\sqrt{\frac{5n}{36}}}$. Tato veličina má podle CLV asymptoticky standardní normální rozdělení, jehož 5% kvantil umíme spočítat: $\Phi^{-1}(0,05) \doteq -1,64$.

= **NORM.S.INV(0,05)**

Tedy

$$\frac{10 - \frac{n}{6}}{\sqrt{\frac{5n}{36}}} \doteq -1,64.$$

Tuto rovnici snadno vyřešíme s výsledkem $n = 96$. Neboli musíme hodit nejméně 96krát kostkou, abychom měli (přibližně) 95% pravděpodobnost, že padne alespoň deset šestek. $\triangle$

Příklad 11.4. Podle úmrtnostních tabulek je pravděpodobnost úmrtí 32leté ženy rovna 0,001 819. V případě úmrtí vyplatí pojišťovna rodině 1 milión korun. Pojišťovna pojišťuje 5 000 32letých žen. Jakou sazbu za pojištění má pojišťovna nastavit, aby pravděpodobnost, že pojišťovna vydělá alespoň 2 milióny korun, byla rovna 0,95.

Řešení. Označme x sazbu za pojištění. Počet žen, které zemřou, se řídí binomickým rozdělením, tedy $Y_n \sim \text{Bi}(n; p)$, kde $n = 5\,000$ a $p = 0,001\,819$. Ze zadání příkladu můžeme stanovit rovnici

$$\mathbb{P}(x \cdot 5\,000 - 1\,000\,000 \cdot Y_n > 2\,000\,000) = 0,95,$$

$$\mathbb{P} \left(Y_n < \frac{5\,000x - 2\,000\,000}{1\,000\,000} \right) = 0,95.$$

²Připomeňme, že kvantil náhodné veličiny, jak jsme jej zavedli v odstavci 6.2, udává polohu levého chvostu náhodné veličiny, kdežto nerovnice (11.3) udává polohu chvostu pravého.

Nyní nerovnost upravíme tak, abychom mohli použít de Moivreovu-Laplaceovu CLV.

$$\begin{aligned}\mathbb{P}\left(\frac{Y_n - 5\,000p}{\sqrt{5\,000p(1-p)}} < \frac{\frac{1}{200}x - 2 - 5\,000p}{\sqrt{5\,000p(1-p)}}\right) &= 0,95, \\ \mathbb{P}\left(Z_n < \frac{\frac{1}{200}x - 11,095}{3,013}\right) &= 0,95, \\ \frac{\frac{1}{200}x - 11,095}{3,013} &\doteq \Phi^{-1}(0,95) \doteq 1,64.\end{aligned}$$

$$= \text{NORM.S.INV}(0,95)$$

Odtud $x \doteq 3\,207,28$.

△

Zesílením de Moivreovy–Laplaceovy věty 11.5 je věta Lévyho³-Lindebergova⁴, která vyjadřuje konvergenci k normálnímu rozdělení za obecnějších podmínek. Náhodnou veličinou Y_n je v tomto případě součet n nezávislých náhodných veličin $X_1, X_2, \dots, X_n$, které jsou stejně rozdělené s konečnou střední hodnotou $\mathbb{E}X_i = \mu$ a konečným rozptylem $\text{var } X_i = \sigma^2$ pro $i = 1, 2, \dots, n$. Pak podle Lévyho-Lindebergovy věty platí pro normovanou náhodnou veličinu

$$Z_n = \frac{Y_n - n\mu}{\sqrt{n\sigma^2}}$$

rovněž limitní vztah

$$\lim_{n \rightarrow \infty} \mathbb{P}(Z_n \leq x) = \Phi(x), \quad -\infty < u < \infty.$$

Věta 11.6 (Lévyho-Lindebergova). *Nechť $X_1, X_2, \dots$ jsou nezávislé stejně rozdělené náhodné veličiny se střední hodnotou μ a konečným rozptylem σ^2 .*

Označme

$$Z_n = \frac{\sum_{k=1}^n X_k - n\mu}{\sqrt{n\sigma^2}}, \quad n = 1, 2, \dots$$

a označme $F_n(x)$ distribuční funkci Z_n . Pak

$$\lim_{n \rightarrow \infty} F_n(x) = \Phi(x), \quad -\infty < x < \infty,$$

kde $\Phi(x)$ je distribuční funkce $\mathbf{N}(0; 1)$.

³Paul Pierre Lévy (1886–1971), francouzský matematik.

⁴Jarl Waldemar Lindeberg (1876–1932), finský matematik.

Větu 11.6 uvádíme bez důkazu (je k němu potřeba teoretický aparát pravděpodobnostních charakteristických funkcí), avšak ukážeme si, jakým způsobem je de Moivreova–Laplaceova věta zobecněním věty Lévyho–Lindebergovy.

Důkaz věty 11.5. Vyjdeme z věty 11.6. Za X_n položíme veličiny s alternativním rozdělením s parametrem p . Pak Y_n má binomické rozdělení s parametry (n, p) a podle věty 11.6 tvrzení (11.1) platí. $\square$

Příklad 11.5. Pan Novák jezdí pravidelně do zaměstnání i zpět tramvají. Je známo, že doba čekání na příjezd tramvaje se pohybuje v mezích 0 až 3 minuty. Jaká je pravděpodobnost, že celková doba čekání pana Nováka během 23 pracovních dnů bude kratší než 80 minut?

Řešení. Doba čekání na příjezd tramvaje při i -té cestě, tj. náhodná veličina X_i , má rovnoměrné rozdělení s hustotou pravděpodobnosti

$$f(x_i) = \begin{cases} \frac{1}{3} & 0 < x_i < 3, \ i = 1, \dots, 46, \\ 0 & \text{jinak.} \end{cases}$$

Celková doba čekání na cestu do zaměstnání a zpět během 23 dnů je náhodná veličina

$$Y_{46} = \sum_{i=1}^{46} X_i.$$

Střední hodnota a rozptyl veličin X_i pro $i = 1, 2, \dots, 46$ jsou

$$\mathbb{E}X_i = \frac{3-0}{2} = \frac{3}{2}, \quad \text{var } X_i = \frac{3}{4}.$$

Jelikož jsou splněny podmínky Lévy–Lindebergovy věty, můžeme hledanou pravděpodobnost určit jako

$$\begin{aligned} \mathbb{P}(Y_{46} < 80) &= \mathbb{P}\left(\frac{\sum_{i=1}^{46} X_i - 46 \cdot \mathbb{E}X_i}{\sqrt{46 \cdot \text{var } X_i}} < \frac{80 - 46 \cdot \mathbb{E}X_i}{\sqrt{46 \cdot \text{var } X_i}}\right) = \\ &= \mathbb{P}\left(Z_{46} < \frac{80 - 69}{\sqrt{34,5}}\right) \doteq \mathbb{P}(Z_{46} < 1,87) \doteq \Phi(1,87) \doteq 0,969. \end{aligned}$$

= `NORM.S.DIST(STANDARDIZE(80;69;ODMOCNINA(34,5));PRAVDA)`

$\triangle$

Tabulka 11.1: Porovnání distribuční funkce Z_{12} s distribuční funkcí $N(0; 1)$

x	0	0,5	1,0	1,5	2,0	2,5	3,0
$\Phi(x)$	0,5	0,691 5	0,841 3	0,933 1	0,977 2	0,993 7	0,998 6
$F_{12}(x)$	0,5	0,689 4	0,839 3	0,932 6	0,977 7	0,994 4	0,999 0

Velmi obecně formuloval centrální limitní větu Alexandr Ljapunov⁵. Dokázal, že rozdělení součtu nezávislých náhodných veličin X_i konverguje pro $i = 1, 2, \dots, n$ k normálnímu rozdělení i tehdy, nejsou-li veličiny X_i stejně rozdělené. Tato věta se dle svého autora nazývá Ljapunovova. Další varianty CLV postulují asymptotické chování náhodných vektorů nebo dokonce celých náhodných procesů. Také podmínku nezávislosti veličin $X_1, X_2, \dots$ lze oslabit. Rigorózní znění těchto vět i jejich matematické důkazy jsou však poměrně obtížné, proto je v této učebnici pomineme a čtenáře odkážeme např. na literaturu [8].

Jak rychle se blíží rozdělení součtů nezávislých náhodných veličin k normálnímu rozdělení? Touto otázkou se nebudeme zabývat teoreticky, ale uvedeme jen numerické srovnání pro jeden speciální případ: nechť X_k , $1 \leq k \leq \leq 12$, jsou „náhodná čísla“, tedy nezávislé náhodné veličiny s rovnoměrným rozdělením na $[0, 1]$. Označme $Z_{12} = \sum_{k=1}^{12} X_k - 6$ (Z_{12} má nulovou střední hodnotu a jednotkový rozptyl). V tabulce 11.1 značí F_{12} distribuční funkci Z_{12} a Φ distribuční funkci $N(0; 1)$; vzhledem k symetrii rozdělení se stačí omezit na nezáporné hodnoty argumentu. Shoda je překvapivě dobrá, uvážíme-li malý počet sčítanců a značně odlišný tvar hustoty rovnoměrného a normálního rozdělení. Tak dobrou aproximaci však nelze vždy očekávat. Zhruba platí, že konvergence je rychlejší pro (kolem střední hodnoty) symetrická rozdělení, pomalejší pro asymetrická. Zmíňme se ještě o pravidlu užívaném v praxi, že totiž binomické rozdělení lze aproximovat normálním, je-li $np(1-p) \geq 9$.

⁵Alexandr Michajlovič Ljapunov (1857–1918), ruský matematik a fyzik.

11.4 Úlohy

11.1 Zásilka obsahuje 3 000 výrobků určitého typu. Je známo, že pravděpodobnost zhotovení vadného výrobku tohoto typu je 0,04.

- Odhadněte pravděpodobnost, že absolutní odchylka podílu vadných výrobků v zásilce a pravděpodobnost vyrobení vadného výrobku bude menší než 0,01.
- Jak se změní výsledek, jestliže pravděpodobnost vyrobení vadného výrobku bude 0,004 a jestliže zásilka bude obsahovat 30 000 výrobků?

11.2 Během zkoušky spolehlivosti se výrobek porouchá s pravděpodobností $p = 0,05$. Jaká je pravděpodobnost, že při zkoušení 100 výrobků se porouchá

- alespoň 7 výrobků,
- méně než 5 výrobků.

11.3 Pravděpodobnost výskytu jevu při jednom pokusu je 0,3. S jakou pravděpodobností lze tvrdit, že relativní četnost výskytu tohoto jevu ve 100 pokusech bude v mezích od 0,2 do 0,4?

11.4 Dlouhodobým pozorováním bylo zjištěno, že doba potřebná k objevení a odstranění poruchy stroje – náhodná veličina X – má střední hodnotu $\mathbb{E}X = 40$ minut a směrodatnou odchylku $\sqrt{\text{var } X} = \sigma = 30$ minut. Jakou dobu si vyžádá objevení a odstranění 100 poruch, jestliže žádáme, aby tato hodnota nebyla s pravděpodobností 0,95 překročena?

11.5 Hodíme 100krát hrací kostkou. Označme S_{100} součet dosažených ok. Užítím CLV určete přibližnou hodnotu pravděpodobnosti

$$\mathbb{P}(320 \leq S_{100} \leq 380).$$

11.6 Nechť v_n značí poměrnou četnost líců v n hodech mincí. Kolik musíme provést hodů, aby pravděpodobnost jevu $\left|v_n - \frac{1}{2}\right| \leq 0,05$ byla nejméně 0,95?

- Řešte pomocí Čebyševovy nerovnosti.
- Řešte pomocí CLV.



11.7 Pojišťovna pojišťuje 1 000 lidí stejného věku. Pravděpodobnost úmrtí během roku je pro každého z nich 0,01. Každý pojištěnec zaplatí 1 200 Kč. V případě úmrtí vyplatí pojišťovna rodině 80 000 Kč. Jaká je pravděpodobnost, že pojišťovna utrpí ztrátu? Použijte CLV.

Řešení úloh

11.1 a) 0,872; b) přibližně 1.

11.2 a) 0,179; b) 0,324.

11.3 0,971

11.4 4 493,5 minut

11.5 0,921 6

11.6 384 hodů

11.7 asi 0,065

Část II

Matematická statistika

Kapitola 12

Zpracování statistického materiálu

12.1 Základní pojmy

Dříve než se začneme zabírat základními statistickými metodami, definujeme základní pojmy z oblasti zpracování statistického materiálu.

Definice 12.1 (Základní pojmy statistiky)

- a) Statistickým souborem Z nazýváme množinu předmětů roztríděných z hlediska jejich určité společné vlastnosti zvané znak.
- b) Předměty patřící do statistického souboru nazýváme statistické jednotky nebo též jednoduše prvky souboru.
- c) Znak, jehož různé hodnoty, popř. obměny nacházíme u všech prvků daného souboru a jenž je zvolen za podklad pro třídění těchto prvků, budeme nazývat argumentem souboru.
- d) Celkový počet všech prvků uvažovaného souboru nazýváme rozsahem souboru.
- e) Součet všech hodnot argumentu (znaku) statistického souboru nazýváme úhrnem argumentu (úhrnem znaku).

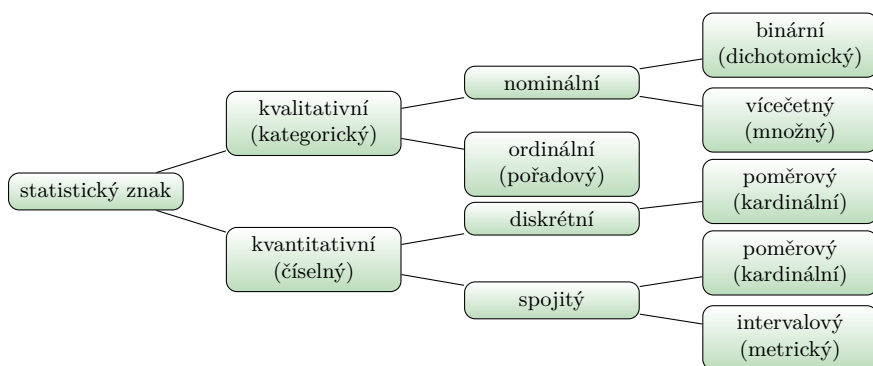


Příklad 12.1. Skupina n osob roztríděná z hlediska jejich věku tvoří statistický soubor rozsahu n . Jeho argument je věk osob. Táž skupina n osob tvoří jiný statistický soubor, zvolíme-li za argument výšku osoby (měřenu např. v centimetrech).

Různé druhy měření na jednotkách základního souboru produkují znaky různého charakteru:

- 1) kategorické znaky z nominálního měření, například zjišťování pohlaví jednotek statistického souboru (muž, žena), povolání, krevní skupiny apod., u kterých spočívá statistické zpracování především v
 - určování četností (absolutní a relativní četnosti),
 - určování charakteristik polohy (modus),
 - určování míry variability (nominální variance),
 - zpracování kontingenčních tabulek,
 - výpočet míry závislosti (koeficient kontingence) aj.
- 2) kategorické znaky z ordinálního (pořadového) měření, kdy se jednotkám souboru přiřazují čísla tak, že vyjadřují pořadí podle určitého kritéria, například ve skupině 20 zaměstnanců se přiřadí čísla 1–20 podle odolnosti jedince vůči stresu (1 znamená nejvíce odolný jedinec, 20 nejméně odolný jedinec) nebo hodnocení v dotazníkovém šetření pomocí Likertovy škály (ze znaku tedy nelze dovodit velikost rozdílu mezi jednotkami s různými hodnotami znaku); statistické zpracování pořadových dat spočívá většinou v
 - určení četností (včetně četností kumulativních),
 - určení charakteristik polohy (medián, kvantily),
 - určení měř variability (např. kvartilová odchylka),
 - určování míry závislosti (např. Spearmanova koeficientu pořadové korelace),
 - testování hypotéz založeném na výpočtu pořadových statistik (znaménkový test, Wilcoxonovy testy, Kruskalův–Wallisův test aj.).
- 3) spojité znaky z intervalového (metrického) měření, kdy se jednotkám přiřadí čísla tak, že vyjadřují velikost rozdílů mezi sebou (tedy mají rovnoměrné rozdíly mezi hodnotami, ale nemají nulovou hodnotu ve smyslu neexistence dané vlastnosti), například měření teploty ve stupních Celsia (0°C neznačí neexistenci teploty, ale pouze jednu z možných hodnot teploty); při statistickém zpracování dat intervalového měření většinou používáme
 - určení charakteristik polohy (aritmetický průměr),
 - určení měř variability (směrodatná odchylka),
 - parametrické testy (Studentovy t -testy, F -test, analýzy rozptylu),
 - určení míry závislosti (Pearsonův koeficient korelace) aj.





Obrázek 12.1: Kategorizace statistických znaků

- 4) diskretní či spojitý znaky z poměrového (kardinálního) měření, kdy jsou jednotkám přiřazena čísla vyjadřující „množství vlastnosti“, kterou měří (tedy mají rovnoměrné rozdíly mezi hodnotami, ale nulová hodnota značí, že daná proměnná u studovaného objektu neexistuje), například měření hmotnosti jedince (0 znamená neexistenci hmotnosti) či měření počtu pojistných událostí (0 znamená neexistence události); při statistickém zpracování dat poměrového měření lze využít postupy ze všech předchozích částí 1)–3).

Schéma na obrázku 12.1 prezentuje druhy proměnných, se kterými statistici nejčastěji pracují. Hodnoty proměnné, které výzkumník zjistí při sběru dat v reálném prostředí, se nazývají data.

12.2 Rozložení četností a třídění

U statistických souborů většího rozsahu často dochází k opakovanému výskytu týchž hodnot statistického znaku. Počet výskytů dané hodnoty znaku představuje tzv. *četnost* hodnoty znaku. Ta slouží především k úplnému popisu kategorických (nominálních a ordinálních) dat, případně též kvantitativních (intervalových a poměrových) dat, pokud jsou předem roztržena do tzv. třídních intervalů. V případě diskretních statistických znaků, které vykazují menší počet hodnot, volíme prosté (bodové) třídění. Jednotlivé hodnoty statistického znaku označujeme jako *třídy* statistického znaku.



Definice 12.2 (Druhy četností)

- a) Počet prvků souboru patřících do k -té třídy nazýváme absolutní četností argumentu v k -té třídě (také absolutní třídní četností nebo stručně četností k -té třídy) a značíme jej f_k .
- b) Je-li f_k absolutní třídní četnost k -té třídy a n rozsah uvažovaného souboru, pak
- $\frac{f_k}{n}$ nazýváme relativní četností k -té třídy,
 - $100 \frac{f_k}{n}$ nazýváme procentní relativní četností k -té třídy.
- c) Kumulativní (součtovou) absolutní četností F_k k -té třídy nazýváme součet všech četností f_j až do k -té třídy včetně, tj.

$$F_k = f_1 + f_2 + \cdots + f_k = \sum_{j=1}^k f_j.$$

- d) Kumulativní relativní četností R_k k -té třídy nazýváme součet

$$R_k = \frac{f_1}{n} + \frac{f_2}{n} + \cdots + \frac{f_j}{n} = \sum_{j=1}^k \frac{f_j}{n} = \frac{F_k}{n}.$$

Kumulativní četnosti (body c) a d)) mají samozřejmě smysl u dat, které lze přirozeným způsobem řadit, tedy speciálně pro ordinální (avšak nikoliv nominální) kategorické znaky. Pro četnosti platí některé přirozené vlastnosti sumarizované v následující větě.



Věta 12.1. Uvažujme statistický soubor rozsahu n , který je rozdělen do r tříd.

- a) Součet všech absolutních četností všech hodnot všech tříd je roven rozsahu souboru, neboli

$$f_1 + f_2 + \cdots + f_r = \sum_{k=1}^r f_k = n.$$

- 2) Kumulativní absolutní četnost všech tříd souboru se rovná rozsahu souboru, neboli

$$F_r = n,$$

- 3) součet všech relativních četností všech tříd souboru se rovná jedné, neboli

$$\frac{f_1}{n} + \frac{f_2}{n} + \cdots + \frac{f_r}{n} = \sum_{k=1}^r \frac{f_k}{n} = 1.$$

Tabulka 12.1: Tabulka rozložení četností pro příklad 12.2

Počet výzev	0	1	2	3	4	5	6	7	Celkem
f_k	8	17	16	10	6	2	0	1	60
F_k	8	25	41	51	57	59	59	60	
f_k/n	0,133	0,283	0,266	0,166	0,1	0,033	0,000	0,016	1
R_k	0,133	0,417	0,683	0,850	0,850	0,983	0,983	1	

Četnosti statistického souboru zpravidla předkládáme formou tabulky obsahující hodnoty statistického znaku s příslušnými absolutními a relativními četnostmi; říkáme ji tabulka rozložení (či rozdělení) četností. Pro vizualizaci těchto četností lze použít řadu grafických nástrojů, popsanych v následujícím výčtu.

- 1) *Sloupcový graf* rozložení absolutních, resp. relativních četností sestavíme tak, že na osu x vyneseme jednotlivé třídy (hodnoty kategorického statistického znaku) a nad každou třídou k sestrojíme obdélník s výškou rovnou příslušné absolutní četnosti f_k , resp. relativní četnosti $\frac{f_k}{n}$.
- 2) *Úsečkový diagram* (nebo graf) rozložení absolutních (resp. relativních) četností sestrojíme obdobně, jen obdélníky nahradíme úsečkami ve směru osy y .
- 3) *Polygon rozložení četností* (spojnicový diagram) dostaneme, jestliže koncové body úsečkového diagramu rozložení četností spojíme úsečkami a vytvoříme tak lomenou čáru, která pak představuje hledaný polygon neboli spojnicový diagram.
- 4) Sloupcový, úsečkový graf nebo polygon kumulativních četností dostaneme analogicky jako v bodech 1, 2 a 3. Polygon kumulativních relativních četností se nazývá *ogivní křivka* (stručně ogiva).

Příklad 12.2.  Na telefonní stanici zaznamenávali počet telefonních výzev za dobu 1 minuty. Během jedné hodiny bylo v určité denní době dosaženo těchto výsledků: 3, 2, 2, 3, 1, 1, 0, 4, 2, 1, 1, 4, 0, 1, 2, 3, 1, 2, 5, 2, 3, 0, 2, 4, 1, 2, 3, 0, 1, 2, 1, 3, 1, 2, 0, 7, 3, 2, 1, 1, 4, 0, 0, 1, 4, 2, 3, 2, 1, 2, 2, 3, 1, 4, 0, 2, 1, 1, 5, 3. Sestavte tabulku rozložení absolutních a relativních četností tohoto statistického souboru (i kumulativních). Dále sestavte graf relativních četností a graf kumulativních relativních četností tohoto souboru.

Řešení. V Excelu se hodnoty četností spočítají nejrychleji nástrojem kontingenční tabulka (Vložení > Kontingenční tabulka): pole „počet výzev“

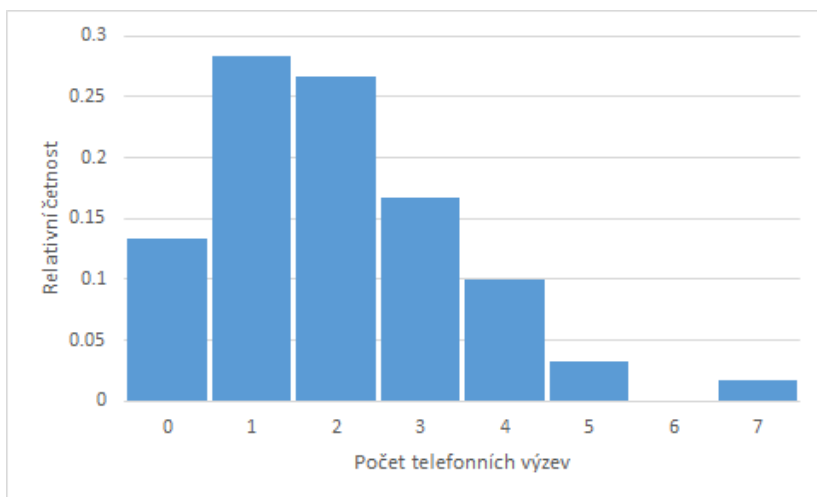
umístíme zároveň do oblasti řádků i hodnot, přičemž v oblasti hodnot upravíme nastavení pole tak, že zaměníme kritérium pole na počet – tím dostaneme rozložení absolutní četnosti. Upravme dále zobrazení hodnot na % ze součtu sloupce (příp. ještě formát čísla na číslo s třemi desetinnými místy), dostaneme hodnoty relativních četností. Kumulativní četnosti dostaneme volbou zobrazení na Mezisoučet v, resp. % mezisoučtu v. Jedno zdrojové pole je možné do oblasti hodnot kontingenční tabulky umístit opakovaně, všechny čtyři typy četností tak můžeme spočítat ve stejné kontingenční tabulce. Výsledek je patrný z tabulky 12.1.

Alternativním postupem je využít funkce Excelu pro podmíněný počet. K tomu je nejprve potřeba vytvořit blok buněk obsahující možné hodnoty kategoričké proměnné (v našem případě tedy hodnot od 0 do 7, dejme tomu v buňkách C3:C9). Funkce `= COUNTIF` vrátí pro každou z těchto hodnot počet výskytů v oblasti zdrojových dat (dejme tomu v buňkách A2:A61). Pro výskyt nul bude tedy vzorec mít podobu `= COUNTIF($C$2:$C$61;$E2)` a pro ostatní hodnoty jej jednoduše zkopírujeme. Výsledkem je tabulka absolutních četností. Relativní četnosti dostaneme prostým vydělením celkového počtu pozorování, kumulativní četnosti postupným přičítáním četností absolutních (přípravu vzorců již ponecháváme na čtenáři jako samostatný úkol).

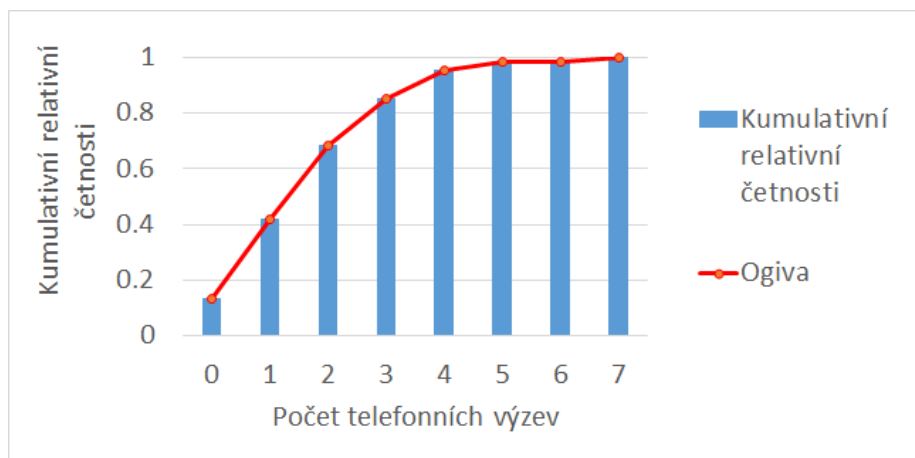
Grafické znázornění získáme pomocí vhodného typu grafu nad výše připravenými tabulkami. Jako zdroj je možné využít i jak kontingenční tabulku, tak ručně připravenou tabulku četností. Jako typ grafu volíme sloupcový, resp. spojnicový graf. Výsledné grafy mohou vypadat kupříkladu jako na obrázcích 12.2 a 12.3. △

Relativní četnosti aproximují pravděpodobnostní rozdělení náhodné veličiny a kumulativní relativní četnosti její distribuční funkci. Například hodnoty relativních četností v tabulce 12.1 představují odhady neznámých pravděpodobností diskrétní náhodné veličiny reprezentující počet telefonních výzev za náhodně zvolený minutový časový interval, a kumulativní relativní četnosti odhady hodnot její distribuční funkce.

Prosté třídění kvantitativních znaků nedává uspokojivé výsledky: většina hodnot takových znaků se v souboru vyskytuje obvykle jen jednou (nebo v malých četnostech, pokud je statistický soubor výsledkem měření s nízkou přesností). Před počítáním četností je tedy potřeba nejprve data kategorizovat do menšího počtu tříd (typicky třídních intervalů) – tomuto postupu se říká intervalové třídění.



Obrázek 12.2: Sloupcový graf relativních četností z příkladu 12.2



Obrázek 12.3: Graf kumulativních relativních četností a ogiva z příkladu 12.2

Definice 12.3 (Variační obor a intervalové třídění)

Nechť a je minimální hodnota argumentu X , b je maximální hodnota argumentu X daného statistického souboru, tj. $x_{\min} = a$, $x_{\max} = b$.¹

- Interval $[a; b]$ nazýváme variačním oborem (nebo též oborem variability, intervalem variability) argumentu X daného statistického souboru.
- Rozdíl $R = b - a$ nazýváme variačním rozpětím (šíře) argumentu X daného statistického souboru.

¹Maximální a minimální hodnotu má smysl uvažovat u metrického měření.

- c) *Variační obor $[a, b]$ rozkládáme na menší části nazývané třídy argumentu X . Nejčastějším případem je rozklad intervalu $[a; b]$ na posloupnost navazujících nepřekrývajících se třídních intervalů $(a_k; b_k]$.*
- d) *Šířkou (délkou) třídního intervalu $(a_k; b_k]$ nazýváme číslo $h_k = b_k - a_k$. Číslo $\frac{1}{2}(a_k + b_k)$ nazýváme středem třídy, číslo a_k dolní hranicí třídy, číslo b_k horní hranicí třídy.*
- e) *Vybranou hodnotu x_k argumentu X , která zastupuje všechny hodnoty patřící do této třídy, nazýváme třídním znakem k -té třídy. Třídní znak je nejčastěji střed příslušné třídy.*

Dvě pozorování považujeme za ekvivalentní, jakmile data (hodnoty znaku) zjištěná v obou měřeních padnou do téže třídy. Při rozkladu variačního oboru $[a; b]$ s n naměřenými hodnotami na třídní intervaly budeme dbát zpravidla těchto zásad:

- 1) Doporučený počet třídních intervalů je 8 až 20. Zvolený počet záleží na rozsahu souboru a účelu statistické tabulky. Počet r třídních intervalů můžeme také určit například pomocí Sturgesova² pravidla

$$r \approx 1 + \log_2 n \doteq 1 + 3,322 \log n,$$

Yuleova³ pravidla

$$r \approx 2,5 \sqrt[4]{n},$$

nebo jednodušších vztahů $r \approx \sqrt{n}$ či $r \approx 5 \log_{10} n$.

- 2) Počet tříd můžeme také určit pomocí volby šířky třídy. V těchto případech volíme obvykle stejnou šířku pro všechny třídní intervaly a obvykle nějaké přirozené hodnoty. Jako orientační hodnotu můžeme vzít například šířku

$$h_k = h \approx \frac{8}{100}(b - a).$$

- 3) Třídní intervaly v definici 12.3 jsou formalizovány jako zleva otevřené a zprava uzavřené. Nejedná se o rigorózní pravidlo, takto se například v Excelu chová funkce = ČETNOSTI a graf typu histogram. (Navíc první třídní interval i v těchto nástrojích je z podstaty uzavřený i zleva.) Jiné nástroje mohou používat alternativních způsobů zařazení hraničních hodnot: kategorizace s využitím funkce = SVYHLEDAT povede například na třídní intervaly zleva uzavřené a zprava otevřené. Hodnoty na hranici tříd obvykle necháváme zařadovat podle výchozího chování zvoleného nástroje.

²Herbert Arthur Sturges (1882–1958), německý matematik a statistik.

³George Udny Yule (1871–1951), anglický matematik a statistik.

Tabulka 12.2: Počty zemřelých na 100 tisíc obyvatel v jednotlivých týdnech roku (od 1. do 52. týdne: zleva doprava shora dolů)

50,08	44,55	42,5	40,68	39,42	38,63	43,45	36,97	39,50	42,26
38,23	38,87	38,55	37,92	38,79	38,00	37,44	39,73	33,41	35,86
32,47	34,05	32,94	32,47	37,13	31,28	33,26	37,52	32,23	34,76
32,78	37,84	34,76	37,29	33,89	35,23	34,05	36,97	30,65	36,34
38,55	35,31	39,50	36,50	39,15	38,79	38,47	43,84	38,47	39,34
43,21	42,18								

Zdroj dat: Český statistický úřad

- 4) Vyskytuje-li se v krajních třídách (první nebo poslední) velmi málo hodnot argumentu X , je vhodné tyto třídy spojit se sousední třídou v třídu jedinou. (Taková třída pak bude mít samozřejmě odlišnou šířku.)

Počet a šířka intervalů nejsou dány jednoznačně, vždy záleží na subjektivní volbě výzkumníka. Cílem je však zvolit takový počet intervalů s příslušnou šířkou, aby vytvořené intervalové rozdělení četnosti vypovídalo o základních charakteristikách sledovaného znaku. Jakmile jsou data roztříděna do třídních intervalů, můžeme nad nimi počítat četnosti a konstruovat četnostní grafy, jak jsme již činili v případě prostého třídění. Některé prvky lze dokonce konstruovat přímo nad daty (roztřídění do třídních intervalů je příslušným nástrojem provedeno automaticky). Konkrétně sloupcový graf absolutních či relativních četností třídních intervalů se nazývá *histogram*. Sestavuje se tak, že na ose x se vynášejí středy jednotlivých tříd a nad každou úsečkou zobrazující určitou třídu (šířky h_k) sestrojíme obdélník s výškou rovnou příslušné absolutní četnosti f_k , popř. relativní četnosti $\frac{f_k}{n}$. V Excelu se histogram absolutních četností vytváří jako graf typu histogram přímo nad daty. Histogram relativních četností však je nutné vytvořit buď jako klasický sloupcový graf nad tabulkou intervalového třídění, anebo je možné použít nástroj Analýza dat (konkrétně analytický nástroj Histogram).

Příklad 12.3.  Data v tabulce 12.2 ukazují počty zemřelých v ČR na 100 tisíc obyvatel k 1. lednu 2023 ve věku 65 až 74 let podle týdnů. Sestavte tabulku intervalového rozložení absolutních a relativních četností (včetně kumulativních) udaného statistického souboru a zkonstruujte histogram četností.



Řešení. Rozsah datového souboru je $n = 52$ pozorování. Odpovídá tomu $1 + \log_2(52) \doteq 6,7 \doteq 7$ třídních intervalů podle Sturgesova

va pravidla a stejný počet intervalů i podle Yuleova pravidla ($2,5 \sqrt[4]{52} \doteq 6,7$) či odmocninového pravidla ($\sqrt{52} \doteq 7,2$). Hodnota logaritmického pravidla je vyšší, $5 \log_{10}(52) \doteq 8,6$ odpovídá osmi nebo devíti třídním intervalům. Případně můžeme vyjít z variačního rozpětí $50,08 - 30,65 = 19,43$ a volit šířku přibližně $8/100 \cdot 19,43 \doteq 1,55$.

Na základě výše uvedené analýzy jsme se rozhodli data intervalově roztrdit do sedmi intervalů (to odpovídá šířce intervalu $19,43/7 \doteq 2,77 \doteq 3$) a ještě zvolíme hranice na „hezkých“ hodnotách počínaje hodnotou 30. Při této volně leží minimum a maximum uvnitř oblasti vymezené tříděním (mezi 30 a 51), ale i v případě, kdyby tomu tak nebylo, je možné případné nízké a vysoké hodnoty zahrnout do krajních třídních intervalů. Řešení ukazuje tabulka 12.3. Jestliže umístíme data do buněk A2:A53, vytvoříme tabulku absolutních četností takto: nejprve do prvního sloupce vytvářené tabulky (dejme tomu do buněk C2:C8) zadáme hodnoty hranic třídních intervalů, počínaje 33 a konče 48. Ve vedlejším sloupci označíme buňky D2:D9 (o jednu buňku více, než je počet hraničních buněk), stiskneme klávesu F2 a zadáme vzorec `= ČETNOSTI (A2:A53;C2:C8)` a ukončíme jej stiskem **Ctrl+Shift+ENTER** – vzorec se takto uloží jako maticový. V novějších verzích Excelu se vzorec vloží jednodušeji jako dynamický maticový vzorec – stačí jej vložit standardním způsobem do buňky D2, zbytek (roztažení do dalších buněk) již zařídí Excel automaticky.

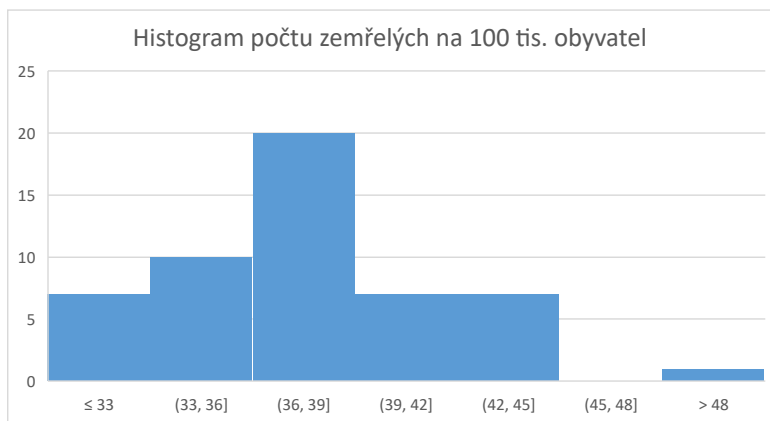
Nástroj histogram (Vložení > Grafy > Histogram) bude volit počet intervalů a šířku tříd automaticky, je možné ale prostřednictvím formátu osy x volby upravit (Formát osy > Možnosti osy). Volbou Interval dolní, resp. horní hranice můžeme upravit i polohu hranic mezi třídními intervaly (zadáva se hranice mezi prvním a druhým, resp. mezi předposledním a posledním třídním intervalem); tím dojde současně i k zahrnutí nízkých a vysokých hodnot znaku do krajních třídních intervalů. Výsledek může vypadat podobně jako na obrázku 12.4.

△

Argument statistického souboru je v pravděpodobnostním pohledu představován náhodnou veličinou X . Ze zákona velkých čísel (podrobněji viz větu 13.2) plyne, že relativní četnost $\frac{f_k}{n}$ udává (přibližně) pravděpodobnost, že X padne do k -té třídy, tedy $\mathbb{P}(a_k < X \leq b_k) \approx \frac{f_k}{n}$. Histogram relativních četností tedy aproximuje hustotu pravděpodobnosti náhodné veličiny X . Podobně kumulativní relativní četnosti znázorněné ogivou aproximují distribuční funkci této náhodné veličiny.

Tabulka 12.3: Tabulka intervalového třídění počtu zemřelých na 100 tisíc obyvatel (n_k – absolutní četnost, F_k – kumulativní absolutní četnost, f_k – relativní četnost, R_k – kumulativní relativní četnost)

k	interval	n_k	F_k	f_k	R_k
1	(30; 33]	7	7	0,135	0,135
2	(33; 36]	10	17	0,192	0,327
3	(36; 39]	20	37	0,385	0,712
4	(39; 42]	7	44	0,135	0,846
5	(42; 45]	7	51	0,135	0,981
6	(45; 49]	0	51	0,000	0,981
7	(48; 51]	1	52	0,019	1,000



Obrázek 12.4: Histogram četností z příkladu 12.3

12.3 Charakteristiky polohy

Pravděpodobnostní chování náhodné veličiny je jednoznačně a úplně určeno pravděpodobnostní funkcí, hustotou pravděpodobnosti nebo distribuční funkcí. Při řešení pravděpodobnostních úloh je mnohdy výhodné shrnout základní informace o rozdělení náhodné veličiny do několika číselných charakteristik, jak jsme učinili v kapitole 6 pro charakteristiky polohy (střední hodnota, modus, medián), variability (rozptyl, směrodatná odchylka) symetrie (šikmost) či koncentrace (špičatost). Takové charakteristiky se ve statistice označují jako teoretické. Podobným způsobem lze však charakterizovat i datové soubory: například relativní četnosti popisující kategorických dat jsou datovým protipólem pravděpodobnostní funkce náhodné veličiny a kumulativní četnosti datovým protipólem distribuční funkce. Stejně tak

je možné vlastnosti proměnných datových souborů shrnout pomocí jednoduchých číselných charakteristik. Protože většina datových souborů pochází z výběrů (empirických dat), označují se tyto charakteristiky jako výběrové (empirické). V tomto a následujících odstavcích se tedy budeme věnovat výběrovým charakteristikám polohy, variability, symetrie a koncentrace empirických souborů.

Definice 12.4 (Aritmetický, geometrický a harmonický průměr)

Nechť je dán statistický soubor, jehož argument x nabývá hodnot $x_1, x_2, \dots, x_n$, které jsou popř. roztrženy do r tříd, přičemž f_k značí absolutní četnost k -té třídy a x_k je vhodně zvolený reprezentant této třídy.

a) Aritmetický průměr $\bar{x}$ je definován vztahy

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \doteq \frac{1}{n} \sum_{k=1}^r f_k x_k. \quad (12.1)$$

b) Geometrický průměr $\bar{x}_g$ je definován pro kladné hodnoty $x_1, x_2, \dots, x_n$ ⁴ vztahem

$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n} \doteq \sqrt[n]{\prod_{k=1}^r x_k^{f_k}}. \quad (12.2)$$

c) Harmonický průměr $\bar{x}_h$ je definován vztahy

$$\bar{x}_h = \frac{1}{a}, \quad kde \quad a = \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i} \doteq \frac{1}{n} \sum_{k=1}^r \frac{f_k}{x_k}. \quad (12.3)$$

Ve vztazích (12.1), (12.2) až (12.3) jsou uvedeny dva tvary. První tvar odpovídá neroztříděnému a druhý tvar roztříděnému souboru.

Aritmetický průměr je zřejmě nejpoužívanější charakteristikou polohy souboru dat. Hodnotu aritmetického průměru však významně ovlivňují extrémní hodnoty zastoupené v souboru (říkáme jim odlehlá pozorování). Proto je někdy vhodné použít jako míru polohy jinou charakteristiku, například medián nebo useknutý průměr.

Geometrický průměr se uplatňuje v ekonomii a finanční matematice, například při výpočtu průměrného procentního zvýšení (snížení) cen zboží (průměrného tempa růstu či poklesu) v úsecích ve sledovaném časovém období. Geometrický průměr nelze použít, pokud argument X nabývá nulové nebo záporné hodnoty.

⁴Bude-li alespoň jedna z hodnot x_i rovna nule, bude také geometrický průměr roven nule.

Harmonický průměr lze použít tehdy, má-li smysl součet reciprokových hodnot, tj. tehdy, když průměrovaná veličina má charakter části z celku, tedy máme vypočítat průměr tzv. poměrných čísel, například při výpočtu průměrné rychlosti $\bar{v}$ auta v km/h, jsou-li dány dílčí dráhy s_i a jim odpovídající rychlosti v_i , kde čas je roven $t = \frac{s}{v}$.

Narozdíl od grafů rozložení četností, které poskytují poměrně obsáhlou informaci o statistickém souboru, dává nám každá charakteristika polohy jen omezenou dílčí informaci o „průměrné“ hodnotě jednotky statistického souboru.

Následující dvě věty uvádíme bez důkazů.

Věta 12.2 (o vlastnostech aritmetického průměru).

a) *Součet všech absolutních odchylek hodnot znaku X_i od aritmetického průměru všech hodnot souboru je roven nule, neboli*

$$\sum_{i=1}^n (x_i - \bar{x}) = 0.$$

b) *Jsou-li všechny hodnoty znaku x_i rovny téže konstantě c , pak je aritmetický průměr všech hodnot znaku roven téže konstantě c , tj.*

$$\bar{x} = \sum_{i=1}^n \frac{x_i}{n} = \sum_{i=1}^n \frac{c}{n} = c.$$

c) *Přičteme-li ke všem hodnotám znaku x_i touž konstantu c , zvětší se o tuto konstantu c i aritmetický průměr všech hodnot znaku, tj. platí*

$$\sum_{i=1}^n \frac{x_i + c}{n} = \bar{x} + c.$$

d) *Vynásobíme-li všechny hodnoty znaku x_i touž konstantou c , bude aritmetický průměr těchto hodnot vynásoben stejnou konstantou c , neboli*

$$\sum_{i=1}^n \frac{c \cdot x_i}{n} = c \cdot \bar{x}.$$

Věta 12.3 (o vztahu mezi průměry, bez důkazu). *Pro libovolný statistický soubor se znakem X platí:*

$$\bar{x}_h \leq \bar{x}_g \leq \bar{x}.$$

Příklad 12.4.  Uvažujme akcii, jejíž cena v prvním roce rostla o 10 %, v druhém roce klesla o 20 % a v třetím roce opět klesla o 15 %. Jaký byl průměrný procentní růst ceny dané akcie?

Řešení. Průměrné tempo růstu cen akcie je geometrickým průměrem dat o růstu (poklesu) její ceny. Můžeme psát:

$$\bar{x}_g = \sqrt[3]{(1 + 0,1) \cdot (1 - 0,2) \cdot (1 - 0,15)} \doteq 0,908,$$

= **GEOMEAN(A1:A3)**

tedy průměrný pokles ceny akcie byl po zaokrouhlení 9,1%. Číslo 9,1% vyjadřuje hodnotu průměrného konstantního poklesu ceny akcie ve sledovaných letech, tedy platí $0,908^3 \doteq 1,1 \cdot 0,8 \cdot 0,85$. Můžeme to též interpretovat tak, že celkový pokles ceny této akcie by byl po třech letech stejný, jako kdyby každým rokem poklesla cena o 9,1%. $\triangle$

Příklad 12.5.  Pobočka České pošty má v určitou dobu celkem čtyři zaměstnance u přepážky. První zaměstnanec odbaví zákazníka za dvě minuty, druhý za tři minuty, třetí za pět minut a čtvrtý za šest minut. Všichni pracovníci pracují na stejně náročném úkolu. Jaká je průměrná doba potřebná k odbavení jednoho zákazníka?

Řešení. První zaměstnanec odbaví za jednu hodinu 30 zákazníků, druhý zaměstnanec za stejnou dobu 20 zákazníků, třetí 12 zákazníků a čtvrtý 10 zákazníků. Za celkový čas 4 hodiny odbaví všichni čtyři zaměstnanci dohromady 72 zákazníků. Průměrná doba potřebná na odbavení jednoho zákazníka je $\frac{4}{72} = \frac{1}{18}$ hodiny, což je asi 3,3 minuty. Zlomek $\frac{4}{72}$ je prostým harmonickým průměrem časů potřebných na odbavení jednoho zákazníka u jednotlivých zaměstnanců.

$$\bar{t}_h = \frac{4}{\frac{1}{t_1} + \frac{1}{t_2} + \frac{1}{t_3} + \frac{1}{t_4}} = \frac{4}{\frac{1}{2} + \frac{1}{3} + \frac{1}{5} + \frac{1}{6}} = \frac{4}{\frac{72}{60}} \doteq 3,3 \text{ min.}$$

= **HARMEAN(A1:A4)**

Průměrná doba potřebná na odbavení jednoho zákazníka je asi 3,3 min. $\triangle$

Má-li výzkumník podezření, že je hodnota vypočítaného aritmetického průměru do značné míry ovlivněna extrémními hodnotami souboru (minimem a maximem hodnot), je často běžnou praxí, že se tyto extrémní hodnoty (nebo určité procento všech hodnot) ze souboru dat vypustí, pak mluvíme o tzv. *useknutém průměru*, nebo se nahradí hodnotami bližšími zbylým hodnotám v souboru, pak jde o tzv. *winsorizovaný průměr*. Nejčastěji se ovšem problém extrémních hodnot řeší robustní charakteristikou polohy, kterou je medián.

Nechť je dán statistický soubor, jehož argument x nabývá hodnot $x_1, x_2, \dots, x_n$. Seřadíme-li hodnoty podle velikosti, dostaneme tzv. *uspořádaný statistický soubor*

$$x_{(1)}, x_{(2)}, \dots, x_{(n)},$$

kde $x_{(1)}$ označuje nejmenší hodnotu, $x_{(2)}$ označuje druhou nejmenší hodnotu atd. Obecně $x_{(i)}$ označuje i -tou pořadovou hodnotu.

Definice 12.5 (Medián)

Medián je definován dvěma způsoby v závislosti na počtu prvků statistického souboru. V případě lichého počtu hodnot vezmeme za medián $\tilde{x}$ prostřední hodnotu uspořádaného souboru

$$\tilde{x} = x_{(\lfloor \frac{n}{2} \rfloor + 1)}.$$

Pokud X má sudý počet hodnot, vezmeme za medián $\tilde{x}$ aritmetický průměr prostředních dvou hodnot

$$\tilde{x} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}.$$

Medián je speciálním případem výběrového kvantilu. Výběrovým kvantilem nazýváme hodnotu zvolenou tak, že pozorování, která jsou menší než tato hodnota, tvoří předepsaný díl výběru (např. 10% výběrový kvantil označuje hodnotu, která je větší než 10 % hodnot statistického souboru a menší než 90 % hodnot statistického souboru).

Definice 12.6 (Kvantil)

Výběrový kvantil $\tilde{x}_p$ je reálné číslo, pro které platí, že $100p$ % jednotek uspořádaného souboru má hodnotu menší nebo rovnou $\tilde{x}_p$ a $100(1-p)$ % jednotek má hodnotu větší nebo rovnou x_p , kde $0 < p < 1$.

Rozeznáváme tři speciální případy výběrového kvantilu: 25% výběrový kvantil, označený $\tilde{x}_{25}$, se nazývá *dolní* (také *první*) *výběrový kvartil* a je to takové číslo, pod nímž leží čtvrtina všech hodnot souboru, neboli 25 % všech hodnot je menších nebo rovno tomuto kvantilu. Dále 50% výběrový kvantil nazýváme druhý kvartil a je jím ve skutečnosti medián, označený v tomto kontextu $\tilde{x}_{50}$. Konečně 75% výběrový kvantil, označený $\tilde{x}_{75}$, se nazývá *horní* (třetí) *výběrový kvartil* a leží pod ním tři čtvrtiny všech hodnot souboru. Dalšími užívanými názvy kvantilů jsou decily $\tilde{x}_{10}, \tilde{x}_{20}, \dots, \tilde{x}_{100}$ a percentily $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_{100}$.

Kvantily (včetně mediánu a kvartilů) nejsou svojí definicí určeny jednoznačně. Například medián pro sudý rozsah statistického souboru může být

Tabulka 12.4: Tabulka absolutních četností z příkladu 12.6

Znak x	x_1	x_2	x_3	x_4
Hodnota znaku	10	16	28	50
Absolutní četnost n_i	30	10	3	1

v principu kterékoli číslo mezi dvěma „prostředními“ hodnotami. Vzorec v definici 12.5 (průměrující tyto dvě hodnoty) je pouze nejobvyklejší volbou hodnoty mediánu. U kvantilů a kvartilů je situace méně jednotná: různý statistický software může vrátit různé (i když podobné) hodnoty výběrových kvantilů a Microsoft Excel má dokonce dvě odlišně implementované funkce pro výpočet kvantilů (`= PERCENTIL.INC` a `= PERCENTIL.EXC`). Pro větší rozsahy souborů jsou však rozdíly z praktického hlediska nepodstatné.

Definice 12.7 (Modus)

Nechť argument statistického souboru může nabývat pouze konečně mnoha hodnot. Pak modus je hodnota argumentu s největší absolutní četností.

Modus, označovaný nejčastěji $\hat{x}$, nemusí být určen jednoznačně: může se stát, že dvě či více hodnot argumentu mají stejnou absolutní četnost. U spojitých znaků pak není modus obvykle definován vůbec (většina hodnot má obvykle pouze jeden nebo žádný výskyt v souboru). V těchto případech je ale možné statistický soubor intervalově roztřídit a určit modus třídních intervalů. Takto určený modus se označuje jako *modální interval*.

Pro výpočet modu v MS Excel lze využít funkci `= MODE.MULT` (případně `= MODE.SNGL`), pokud je modus jednoznačný). Jejich argumentem je oblast všech zpracovávaných dat a výstupem obou funkcí je hodnota, která se v souboru dat vyskytuje nejčastěji, tedy minimálně dvakrát. V případě, že v oblasti je každá hodnota zastoupena jen jednou, funkce vrátí chybu. Je-li modus vícenásobný, všechny takové výskyty vrátí první z výše jmenovaných funkcí jako dynamickou matici.

Modus, medián a aritmetický průměr nabývají většinou různých hodnot. Blízké hodnoty mediánu a průměru indikují symetrické rozdělení četností. Ve statistice se také často zpracovávají i soubory dat s více maximy, resp. vrcholy. Rozdělení s jedním vrcholem se nazývají unimodální, se dvěma vrcholy bimodální atd.

Příklad 12.6.  Tabulka 12.4 ukazuje absolutní četnosti čtyř hodnot nějakého statistického znaku. Vypočítejte aritmetický průměr, medián a modus. Určete hodnotu dolního a horního výběrového kvartilu.

Řešení. Aritmetický průměr je

$$\bar{x} = \frac{30 \cdot 10 + 10 \cdot 16 + 3 \cdot 28 + 1 \cdot 50}{44} = 13,50.$$

= PRŮMĚR(A1:A44)

Jelikož máme 44 hodnot, bude medián průměr 22. a 23. pořadové hodnoty, tedy $\tilde{x} = 10$.

= MEDIAN(A1:A44)

Modus je nejčtenější hodnota znaku, podle tabulky 12.4 je to číslo 10, lze ověřit, že také $\hat{x} = \text{MODE.MULT}(10; \dots; 50) = 10$.

Dolní výběrový kvartil bude průměr 11. a 12. pořadové hodnoty, tj. 10, a horní výběrový kvartil je 16.

$$\tilde{x}_{25} = \text{QUARTIL.EXC}(A1:A44; 1) = 10,$$

$$\tilde{x}_{75} = \text{QUARTIL.EXC}(A1:A44; 3) = 16. \quad \triangle$$

12.4 Charakteristiky variability (měnlivosti)

Definice 12.8 (Charakteristiky variability)

a) Variační šíří (variačním rozpětím) R nazýváme rozdíl největší a nejmenší hodnoty znaku (v tomto pořadí), tedy

$$R = x_{(n)} - x_{(1)}.$$

b) Rozptylem (disperzí, také populačním rozptylem) σ_n^2 statistického souboru s rozsahem n nazýváme aritmetický průměr kvadratických odchylek $(x_k - \bar{x})^2$ hodnot argumentu X od aritmetického průměru $\bar{x}$:

$$\sigma_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{k=1}^r f_k (x_i - \bar{x})^2. \quad (12.4)$$

c) Výběrovým rozptylem s_n^2 statistického souboru s rozsahem n nazýváme průměr kvadratických odchylek $(x_k - \bar{x})^2$ hodnot argumentu X od aritmetického průměru $\bar{x}$, korigovaný o tzv. počet stupňů volnosti:

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \sum_{k=1}^r f_k (x_i - \bar{x})^2. \quad (12.5)$$

d) Populační, resp. výběrovou směrodatnou odchylkou σ_n , resp. s_n nazýváme

$$\sqrt{\sigma_n^2} = \sigma_n \geq 0. \quad (12.6)$$

pro populační verzi rozptylu, resp.

$$\sqrt{s_n^2} = s_n \geq 0 \quad (12.7)$$

pro výběrovou verzi rozptylu.

- e) Průměrnou odchylkou $\bar{d}$ nazýváme aritmetický průměr absolutních hodnot odchylek od aritmetického průměru $\bar{x}$, tj.

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}| = \frac{1}{n} \sum_{k=1}^r f_k |x_k - \bar{x}|. \quad (12.8)$$

- f) Variační koeficient⁵ v statistického souboru je definován jako

$$v = \frac{\sigma_n}{\bar{x}} \text{ nebo } v = \frac{s_n}{\bar{x}} \quad (12.9)$$

(podle verze použité směrodatné odchylky).

- g) Variační poměr (pro nominální data) se definuje

$$VP = 1 - \frac{n_m}{n}, \quad (12.10)$$

kde n_m je četnost odpovědí v modální kategorii a n je celkový rozsah souboru.

- h) Nominální variance se definuje předpisem

$$NV = \frac{n^2 - \sum n_k^2}{n^2}, \quad (12.11)$$

kde n_k je četnosti příslušné kategorie a n rozsah souboru.

Charakteristiky f) a g) se využívají u nominálních dat.

Variační šíře je „nejhrubší“ mírou variability, která udává délku intervalu, v němž se nacházejí všechny hodnoty znaku. Nevýhoda variační šíře spočívá v tom, že závisí pouze na extrémních hodnotách, které nemusí být pro daný znak charakteristické. Proto se v rámci vyšetřování variability (měnlivosti) dat určují také tzv. *kvantilová rozpětí*, speciálně

- (mezi)kvartilové rozpětí jako rozdíl horního a dolního kvartilu,

$$R_Q = \tilde{x}_{75} - \tilde{x}_{25},$$

- decilové rozpětí

$$R_D = \tilde{x}_{90} - \tilde{x}_{10},$$

⁵Označovaný také CV z anglického *coefficient of variation*.

- percentilové rozpětí

$$R_C = \tilde{x}_{99} - \tilde{x}_1,$$

a (mezi)kvartilové odchytky

- (mezi)kvartilová odchytky $Q = \frac{R_Q}{2}$,
- decilová odchytky $D = \frac{R_D}{8}$,
- percentilová odchytky $C = \frac{R_C}{98}$.

Například kvartilová odchytky udává průměrnou délku dvou prostředních kvartilových intervalů, podobně decilová a percentilová odchytky vyjadřuje průměrnou vzdálenost mezi sousedními decily, resp. percentily. Mezikvartilové rozpětí a kvartilová odchytky se obvykle pojí se situací, kdy je za míru polohy statistického souboru volen medián.

Variační koeficient slouží k porovnání relativního významu směrodatné odchytky od aritmetického průměru, tj. kolik procent průměru představuje směrodatná odchytky. Tento ukazatel je bezrozměrná veličina většinou vyjádřená v procentech a využívá se pro srovnání variabilit dvou a více proměnných, které používají různé jednotky měření. Variační koeficient se nehodí v případech, kdy je průměrná hodnota blízko nule. Naopak je vhodný, jsou-li všechna data kladná. Hodnota $v = 1$ představuje 100 %.

Rozptyl je definován vzorcem (12.4), který by se uplatnil při manuálním výpočtu v MS Excel. Pro výpočet za pomoci kalkulátoru bychom využili tzv. výpočetního tvaru

$$\sigma_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i^2) - \bar{x}^2 = \frac{1}{n} \sum_{k=1}^r f_k x_i^2 - \bar{x}^2. \quad (12.12)$$

Máme-li však k dispozici podkladová data, nejrychlejší cestou k výpočtu rozptylu by byl vzorec **=VAR.P**.

Následující větu opět uvádíme bez důkazu.

Věta 12.4 (o vlastnostech rozptylu).

- a) Jsou-li všechny hodnoty statistického znaku totožné konstanty c , pak je jejich rozptyl roven nule, tj.

$$\sigma_n^2 = \sum_{i=1}^n \frac{(c - c)^2}{n} = 0.$$

- b) Přičteme-li k jednotlivým hodnotám znaku x_i touž konstantu c , hodnota rozptylu se nezmění, tedy platí

$$\sum_{i=1}^n \frac{[(x_i - c) - (\bar{x} + c)]^2}{n} = \sigma_n^2.$$

c) *Násobíme-li jednotlivé hodnoty znaku X_i konstantou c , je rozptyl násoben druhou mocninou této konstanty, tedy*

$$\sum_{i=1}^n \frac{(c \cdot x_i - c \cdot \bar{x})^2}{n} = c^2 \cdot \sigma_n^2.$$

Hodnoty argumentu statistického souboru jsou nejčastěji realizace nějaké náhodné veličiny. Například počet telefonních hovorů na ústředně za 1 minutu (viz příklad 12.2) je náhodná veličina, která se řídí Poissonovým rozdělením $X \sim \text{Po}(\lambda)$, v tuto chvíli s neznámým parametrem λ . Jedním z úkolů statistické analýzy (konkrétně teorie odhadu) je odhadnout skutečnou hodnotu parametru λ . Máme-li k dispozici data, pak k odhadu neznámé střední hodnoty $\mathbb{E}X = \lambda$ této náhodné veličiny použijeme aritmetického průměru statistického souboru. K aproximaci rozptylu náhodné veličiny $\text{var } X = \lambda$ se v podobném duchu používá rozptyl statistického souboru. Rozptyl uvedený ve vzorcích (12.4) a (12.12) však rozptyl podkladové náhodné veličiny ve skutečnosti podhodnocuje, proto se k odhadu používá výběrová verze rozptylu (12.5), příp. výpočetní vzorec

$$s_n^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k^2) - \frac{n}{n-1} \bar{x}^2 = \frac{1}{n-1} \sum_{i=1}^r f_i x_i^2 - \frac{n}{n-1} \bar{x}^2.$$

V Excelu k tomu rovněž slouží funkce `=VAR.S`, resp. `=SMODCH.VÝBĚR.S` pro výpočet výběrové směrodatné odchylky. Takto definovaný rozptyl teoretickou hodnotu rozptylu podkladové náhodné veličiny již nepodhodnocuje (podrobněji viz větu 13.1 na straně 263).

Příklad 12.7.  Vypočítejte charakteristiky variability, variační šíři, horní a dolní kvartil, kvartilové rozpětí, kvartilovou odchylku a variační koeficient dat (viz tabulku 12.5), která vyjadřují spotřebu paliva v mpg (míle na galon) osobních automobilů různých světových značek.

Řešení. Rozptyl σ_n^2 je aritmetický průměr kvadratických odchylek hodnot od aritmetického průměru:

$$\begin{aligned} \sigma_{82}^2 &= \frac{1}{82} \cdot \sum_{k=1}^{82} (x_k - 33,78)^2 = \\ &= \frac{1}{82} \cdot ((23,1 - 33,78)^2 + \dots + (23,4 - 33,78)^2) \doteq 98,87. \end{aligned}$$

`=VAR.P(D4:D85)`

Směrodatná odchylka $\sigma_{82} = \sqrt{s_{82}^2} \doteq \sqrt{98,87} \doteq 9,94$.

Tabulka 12.5: Soubor dat pro příklad 12.7

23,1	16,7	22,9	23,6	22,9	19,5	46,5	43,4	39,3	31,5	28,0
31,5	31,4	35,1	42,2	36,1	36,1	28,0	31,4	25,6	56	55,9
65,4	35,4	45,4	46,2	38,8	40,9	40,9	38,4	40,9	38,8	38,4
59,2	53,3	38,4	32,2	25,3	32,2	28,0	23,6	40,0	17,0	23,5
23,9	36,3	18,1	17,2	23,4	35,1	35,0	23,6	30,4	28,9	23,6
23,6	40,7	32,2	13,2	28,0	41,1	40,4	39,6	39,3	29,5	49,0
33,7	32,6	31,3	28,0	31,3	33,2	32,9	31,2	35,3	32,3	32,2
38,9	38,2	46,9	27,7	23,4						

= `SMODCH.P(D4:D85)`

Výběrový rozptyl s^2 se liší pouze normovacím jmenovatelem:

$$s_{82}^2 = \frac{82}{81} \sigma_{82}^2 \doteq 100,09.$$

= `VAR.S(D4:D85)`

Výběrová směrodatná odchylka $s_{82} = \sqrt{s_{82}^2} = \sqrt{100,09} \doteq 10$.

= `SMODCH.VÝBĚR.S(D4:D85)`

Průměrná absolutní odchylka

$$\begin{aligned} \bar{d} &= \frac{1}{82} \sum_{k=1}^{81} |x_k - 33,78| = \\ &= \frac{1}{82} \cdot (|23,1 - 33,78| + \dots + |23,4 - 33,78|) \doteq 7,73. \end{aligned}$$

= `PRŮMODCHYLKA(D4:D85)`

Variační koeficient

$$v = \frac{\sigma}{\bar{x}} \doteq \frac{10}{33,78} \doteq 0,296.$$

Kvartily $x_{25} \doteq 27,175$, $x_{75} \doteq 39,375$.

x_{25} = `QUARTIL.EXC(D4:D85;1)`

x_{75} = `QUARTIL.EXC(D4:D85;3)`

Kvartilové rozpětí $R_Q = 12,2$, kvartilová odchylka $Q = \frac{12,2}{2} = 6,1$. △

12.5 Obecné výběrové momenty*

Ve statistice slouží tzv. momenty k popisu charakteristik rozložení pravděpodobnosti nebo rozdělení dat. Pod pojmem moment se v matematice

rozumí charakteristika (vlastnost) pravděpodobnostního rozdělení nebo rozdělení dat, která může poskytnout informace o tvaru a charakteru tohoto rozdělení. Tyto momenty se mohou lišit ve své schopnosti popisovat rozložení dat, a proto jsou užitečné při analýze dat a porovnávání různých souborů dat.

Tzv. *obecný moment k -tého řádu* je definován vztahem

$$m'_k = \frac{1}{n} \sum_{i=1}^n x_i^k, \quad k = 0, 1, \dots \quad (12.13)$$

Tzv. *centrální moment k -tého řádu* je definován vztahem

$$m_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k, \quad k = 0, 1, \dots \quad (12.14)$$



Mezi momenty patří například průměr a rozptyl:

- průměr je obecným momentem prvního řádu ($m'_1 = \bar{x}$),
- rozptyl je centrálním momentem druhého řádu ($m_2 = \sigma_n^2$).

Na momentech třetího a čtvrtého řádu jsou založeny též charakteristiky šikmosti a špičatosti, které zavedeme v následujícím odstavci.

12.6 Charakteristiky symetrie a koncentrace

Při statistické analýze dat často nevystačíme s charakteristikami polohy a variability. Další informace o souboru poskytují charakteristiky symetrie, resp. koncentrace („hustoty“) dat. Na úrovni náhodné veličiny jsme se s nimi seznámili v odstavci 6.4. Nyní zavedeme koeficienty šikmosti a špičatosti jako charakteristiku statistického souboru dat.



Definice 12.9 (Koeficient šikmosti, symetrie)

Koeficient šikmosti je dán vztahem

$$a_3 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\sigma_n^3} = \frac{m_3}{m_2^{3/2}},$$

kde m_2 je druhý centrální moment (rozptyl), m_3 je třetí centrální moment a s_n směrodatná odchylka statistického souboru.

Šikmost je možné nad datovým souborem zavést několika různými způsoby, přičemž všechny odhadují šikmost podkladové náhodné veličiny, avšak žádný není obecně preferován před jiným. Například funkce `= SKEW` v MS Excel počítá šikmost dle vzorce (21.4) uvedeného na straně 443.

Je-li $a_3 = 0$, jsou v souboru hodnoty ležící pod a nad průměrem souboru zastoupeny stejně, což vede k souměrnému rozložení četností souboru. Je-li $a_3 > 0$, je zastoupení malých hodnot větší než velkých hodnot; takové rozložení četností označujeme jako zešikmené doleva. Je-li $a_3 < 0$, bude rozložení četností zešikmené doprava.

Definice 12.10 (Koeficient špičatosti, koncentrace)

Koeficient špičatosti je definován vztahem

$$a_4 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\sigma_n^4} - 3 = \frac{m_4}{m_2^2} - 3,$$

kde m_2 je druhý centrální moment (rozptyl), m_4 je čtvrtý centrální moment a s_n směrodatná odchylka statistického souboru.

I koeficient špičatosti se zavádí vícero způsoby, které všechny odhadují špičatost (přesněji „nadměrnou“ špičatost, excess kurtosis) podkladové náhodné veličiny. V Excelu používá funkce `= KURT` upraveného vzorce (21.3).

Je-li $a_4 = 0$, půjde o normální špičatost odpovídající koncentraci normálního rozdělení. Takové rozdělení označujeme jako mesokurtické. Je-li $a_4 > 0$, jsou hodnoty statistického souboru více koncentrovány kolem průměrné hodnoty. Rozdělení s výraznější špičatostí, než má hustota pravděpodobnosti normálního rozdělení, se nazývá leptokurtické. Bude-li naopak $a_4 < 0$, větší množství dat se nachází dále od průměru a jejich rozdělení nabývá ploššího charakteru. Rozdělení s menší špičatostí než u normálního rozdělení se nazývá platykurtické rozdělení.

Příklad 12.8.  Určete koeficienty šikmosti a špičatosti pro data z příkladu 12.7.

Řešení. $\bar{x} = 33,78$, $\sigma_n \doteq 9,94$,

$$a_3 = \frac{\sum_{i=1}^{82} (x_i - 33,78)^3}{82 \cdot 9,94^3} \doteq 0,595.$$

 = SKEW.P(A2:A45)

Protože je $a_3 > 0$, očekáváme při rozložení četností zešikmení doleva. Modifikovaná verze šikmosti je pomocí funkce  = SKEW.P spočtena v hodnotě 0,606.

$$a_4 = \frac{\sum_{i=1}^{82} (x_i - 33,78)^4}{82 \cdot 10^4} \doteq 0,749.$$

 = KURT(A2:A45)

Hodnoty statistického souboru jsou více koncentrovány kolem průměru, než by odpovídalo standardnímu normálnímu rozdělení. $\triangle$

12.7 Úlohy

12.1 Vytvořte graf rozdělení absolutních četností, histogram relativních četností a ogivu pro data z příkladu 12.3.

12.2 Vypočítejte charakteristiky polohy a variability, včetně hodnot kvartilů, kvartilového rozpětí a kvartilové odchylky pro následující soubor dat: 2, 7, 1, 9, 2, 4, 4, 5, 2, 8, 2, 6.

12.3 Údaje o skupinách aut rozdělených podle stáří a jejich počtu ukazují tabulka 12.6. Soubor popište pomocí základních charakteristik, sestavte histogram a srovnajte s normálním rozdělením.

Tabulka 12.6: Tabulka rozdělení četnosti aut podle stáří

Stáří auta	1	2	3	4	5	6	7	8	9	10
Počet	10	19	27	10	9	7	8	5	3	2

12.4 Při výstupní kontrole bylo náhodně vybráno deset balení rýže s deklarovanou hmotností 480 g a deset balení čočky každé o hmotnosti 500 g, z nichž každé bylo očíslováno a po jedné zváženo. Výsledky vážení jsou v tabulce 12.7. Vypočítejte rozptyl hmotností balení rýže a čočky a určete směrodatné odchylky hmotnosti balení. Porovnejte přesnost obou strojů, se kterou plní stroje balení rýží a čočkou.

Tabulka 12.7: Hmotnosti jednotlivých balení rýže

Č. balení	1	2	3	4	5	6	7	8	9	10
m [g] rýže	485	482	487	479	476	480	482	488	477	488
m [g] rýže	485	495	500	510	520	512	489	491	485	488

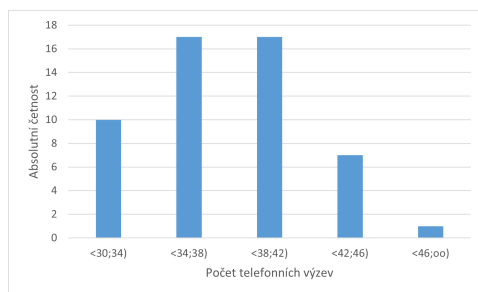
12.5 Vypočítejte $\bar{x}$, R , s_n^2 , s_n , $\bar{d}$ a v pro tři soubory dat v tabulce 12.8 a porovnejte variabilitu dat všech souborů.

Tabulka 12.8: Tři soubory dat

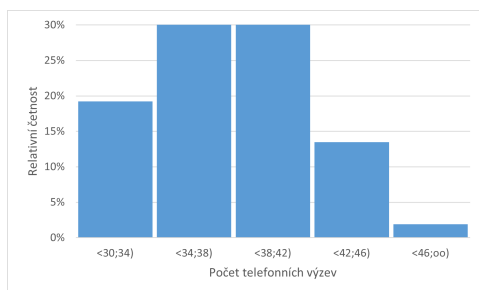
S_1	10	10	10	10	10	10	10	10	10
S_2	6	7	8	9	10	11	12	13	14
S_3	2	4	6	8	10	12	14	16	18

Řešení úloh

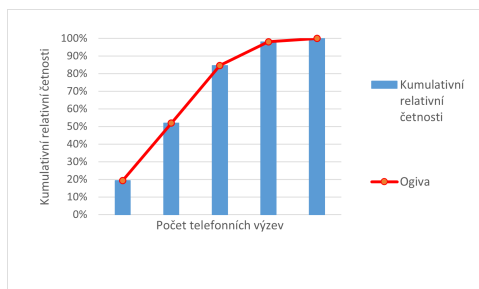
12.1



Obrázek 12.5: Rozložení absolutních četností



Obrázek 12.6: Histogram



Obrázek 12.7: Rozložení kumulativních relativních četností a ogiva

12.2 $\bar{x} = 4, \bar{3}, \bar{x}_g \doteq 3,54, \bar{x}_h \doteq 2,83, \tilde{x} = 4, \hat{x} = 2, x_{0,25} = 2, x_{0,75} = 6,75, R_Q = 4,75, Q = 2,375, R = 8, \sigma_{12}^2 = 6,5, \sigma_{12} \doteq 2,56, s_{12}^2 = 7,15, s_{12} \doteq 2,67, \bar{d} = 2,2.$

12.3 $\bar{x} \doteq 3,99, \tilde{x} = 3, \hat{x} = 3, \sigma_{10} \doteq 2,31, \text{variační koeficient } v \doteq 57,93\%, \text{ což svědčí o velké rozptýlenosti dat souboru, koeficient šikmosti } a_3 \doteq 0,8135, \text{ koeficient špičatosti } a_4 \doteq -0,200235.$

12.4 Balení rýže: $\bar{x} = 482,4 \text{ g}, s_{10}^2 = 19,82, s_{10} \doteq 4,45, v \doteq 0,0092$, tj. asi 0,9 %, což představuje vysokou přesnost práce stroje; balení čočky: $\bar{x} = 497,5 \text{ g}, s_{10}^2 = 155,83, s_{10} \doteq 12,48, v \doteq 0,025$, tj. 2,5 %, tedy stroj pracuje také s vysokou přesností, přesto s nižší přesností než stroj, který balí rýži.

Tabulka 12.9: Popisné statistiky k cvičení 12.5

12.5

	$\bar{x}$	R	s_n^2	s_n	$\bar{d}$	v
S_1	10	0	0	0	0	0
S_2	10	8	7,5	2,74	2,2	0,27
S_3	10	16	30	5,48	4,4	0,55

Kapitola 13

Stavební kameny statistiky

13.1 Základní a výběrový soubor

V mnoha případech nemůžeme při statistickém zpracování dat vycházet ze základního souboru Z (např. má-li soubor úplný nebo značně velký rozsah) a musíme se omezit na nějaký podsoubor souboru Z . Statistické výsledky získané zpracováním statistického podsouboru pak můžeme využít buď

- 1) pro jejich zobecnění na celý základní statistický soubor (na celou populaci) Z (toto zobecnění nazýváme statistickou indukci), nebo
- 2) pro testování hypotéz o celé populaci (neboli statistickou inferenci).

Znamená to tedy, že vyšetřujeme jen určitou část prvků zkoumaného souboru, kterou nazýváme výběrovým souborem. Statistická indukce ani statistická inference nám nedává závěry o populaci s naprostou jistotou, ale jen s předem danou pravděpodobností. Základem je teorie náhodných výběrů, které se budeme věnovat v této kapitole.

Náhodné výběry je potřeba chápat ze dvou různých pohledů: buď se bude jednat o kolekci (posloupnost) náhodných veličin označených velkými tiskacími písmeny, nebo půjde o kolekci realizací těchto náhodných veličin, označených malými písmeny. V druhém případě tedy jde o konkrétní čísla – často mluvíme o hodnotách, realizaci náhodného výběru, o instanci náhodného výběru, či jednoduše o datech.

Náhodné výběry můžeme dělit například podle způsobu jejich provedení:

- 1) *Prostý náhodný výběr s vrácením* je takový výběr, při němž se každý prvek základního souboru vrátí po vybrání zpět do souboru a další prvek se vybírá opět z celého základního souboru.
- 2) *Prostý náhodný výběr bez vrácení* je takový výběr, při němž se vybraný prvek nevrací zpět do základního souboru.

- 3) *Oblastní (stratifikovaný) výběr* spočívá v tom, že základní výběr rozdělíme na stejnorodé disjunktní části a v každé z nich pak provedeme náhodný výběr. O základním souboru ovšem musíme mít dostatečné informace umožňující správnou volbu jednotlivých oblastí. Stratifikovaný výběr realizují například agentury zjišťující předvolební průzkumy: ve výběru chceme mít poměrnou část mladých lidí, lidí ve středním věku a v důchodovém věku, nebo lidí žijících ve městě a lidí žijících na venkově apod., a to v poměrech odpovídající zkoumané populaci.
- 4) *Systematický (mechanický) náhodný výběr* spočívá v tom, že pravidla výběru jsou pevně dána a při opakování procesu výběru na stejné populaci a při stejné volbě počátečního prvku obdržíme identický výběr. Kupříkladu můžeme základní statistický soubor seřadit dle určeného kritéria (například podle abecedy) a následně vybírat každý k -tý prvek. Náhodnost v tomto případě spočívá pouze ve výběru počátečního prvku mezi prvními k prvky.

Jiné dělení náhodných výběrů můžeme provést podle jejich rozsahu:

- 1) náhodný výběr *malého rozsahu* – o rozsahu výběru $n < 30$,
- 2) náhodný výběr *velkého rozsahu* – o rozsahu výběru $n \geq 30$.

Obecně jako velkého rozsahu označujeme náhodné výběry, ve kterých je již patrná platnost centrální limitní věty. Číselná hranice mezi výběrem malého a velkého rozsahu je tedy orientační a záleží na vlastnostech prostředí, ve kterém je výběr realizován. Například v ekonomických studiích jsou typické výběry observačního typu (data jsou v nich pouze pozorována a prostředí, ve kterých jsou pozorována, nejsou pro všechna pozorování identická) a typická hranice mezi výběrem malého a velkého rozsahu se v nich posouvá ke 120 pozorováním.

V dalším textu budeme nadále pracovat pouze s prostým náhodným výběrem s vracením. Z hlediska teorie pravděpodobnosti definujeme náhodný výběr následujícím způsobem.

Definice 13.1 (Náhodný výběr)

Nechť Z je statistický soubor, jehož argument představuje náhodnou veličinu X . Náhodným výběrem z rozdělení náhodné veličiny X budeme nazývat posloupnost n nezávislých realizací pokusu, danou náhodnými veličinami $X_1, X_2, \dots, X_n$, které mají totéž rozdělení jako náhodná veličina X a jsou sdruženě nezávislé.

Náhodným výběrem tedy nazýváme takový výběr, který poskytuje každému prvku základního statistického souboru stejnou a nezávislou pravděpodobnost, že bude zahrnut do výběru.

Tabulka 13.1: Porovnání teoretických pravděpodobností s relativními četnostmi

k	0	1	2	3	4	5	6	7
$\mathbb{P}(X = k)$	0,135	0,271	0,271	0,180	0,090	0,360	0,012	0,003
f_k	0,133	0,283	0,266	0,166	0,100	0,033	0	0,016

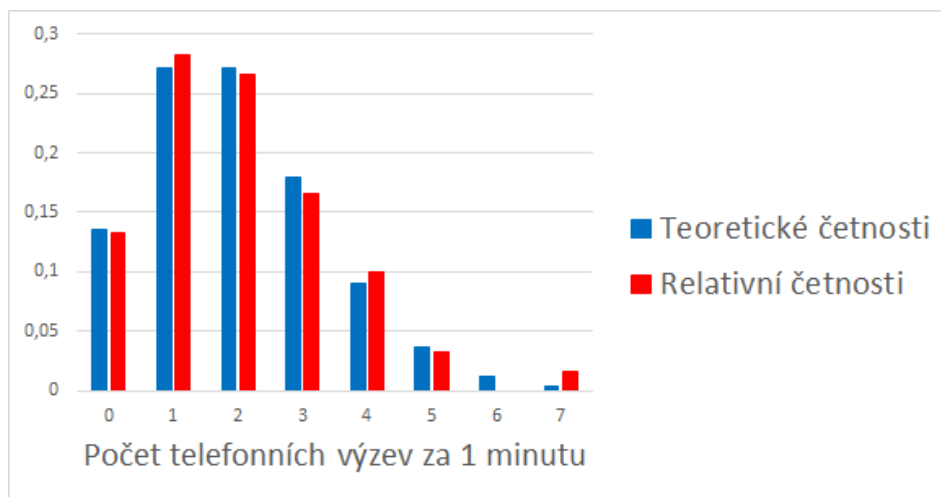
Definice 13.2 (Teoretické a empirické charakteristiky)

Charakteristiku základního souboru $\mathbf{Z}$ reprezentovanou náhodnou veličinou X budeme nazývat teoretickou charakteristikou. Charakteristiku odvozenou z náhodného výběru budeme nazývat empirickou (výběrovou) charakteristikou.

Teoretické charakteristiky jsou typicky střední hodnota, medián, modus, rozptyl, směrodatná odchylka a obecně jakákoliv další charakteristika náhodné veličiny X , která sama o sobě charakterizuje zvolený znak základního souboru. Z povahy věci tedy teoretické charakteristiky základního souboru jsou vždy určitá (nenáhodná) čísla, zatímco empirické charakteristiky jsou náhodné veličiny, neboť jejich hodnoty se mění od jedné realizace náhodného výběru k druhé. Empirické charakteristiky také nazýváme krátce *statistikami*. Jestliže známe typ rozdělení náhodné veličiny X představující argument základního statistického souboru $\mathbf{Z}$, můžeme za určitých předpokladů použít empirických charakteristik k určení odpovídajících teoretických charakteristik.

Příklad 13.1. Statistický soubor představují všichni muži České republiky. Argumentem je jejich věk. Náhodná veličina X určuje věk náhodného vybraného muže z ČR. Pro určení charakteristik náhodné veličiny X můžeme buď zkoumat celou populaci mužů v ČR (například z registrů obyvatel), což může být dost nákladné, anebo provedeme náhodný výběr o menším rozsahu n . Věk každého vybraného muže je realizací náhodné veličiny X_i , kde $i = 1, \dots, n$ je index každého muže. Empirické charakteristiky zjištěné na základě výběru $X_1, \dots, X_n$ pak mohou odhadovat teoretické charakteristiky náhodné veličiny X .

Příklad 13.2.  Necht' je X je náhodná veličina udávající počet telefonních výzev za dobu 1 minuty na telefonní stanici zmíněné v příkladu 12.2. Data zaznamenaná v příkladu 12.2 reprezentují realizaci náhodného výběru z rozdělení X . Určete empirické charakteristiky střední hodnoty a rozptylu a porovnejte je s teoretickými charakteristikami náhodné veličiny X .



Obrázek 13.1: Porovnání relativních a teoretických četností z příkladu 12.2

Řešení. Na otázky zmíněné v zadání příkladu 13.2 a na mnohé další dává odpověď matematická statistika, kterou se budeme zabývat podrobně v následujících kapitolách. Zde prozatím uvedeme náznak běžných technik, které matematická statistika používá.

Empirickou střední hodnotu určujeme typicky aritmetickým průměrem (12.1) a empirický rozptyl vzorcem (12.5). Čtenář sám výpočtem ověří, že $\bar{x} = 2$ a $s_{60}^2 = 2,136$. Tato čísla slouží ve statistice jako odhadnuté hodnoty neznámých teoretických charakteristik střední hodnoty a rozptylu náhodné veličiny X .

Předpokládejme dále, že $X \sim \text{Po}(\lambda)$. Z teorie pravděpodobnosti víme, že v takovém případě platí $\mathbb{E}X = \lambda = \text{var } X$ (viz odstavec 7.8). Položme si například otázku, zda empirická data odpovídají hypotéze $X \sim \text{Po}(2)$ o teoretickém parametru náhodné veličiny X . Můžeme to udělat orientačně tak, že porovnáme relativní četnosti (máme je již v tabulce 12.1) s teoretickými pravděpodobnostmi $\mathbb{P}(X = k)$ dané rozdělením náhodné veličiny $\text{Po}(2)$. Výsledek tohoto orientačního testu je zaznamenán prostřednictvím tabulky 13.1 a sloupcového grafu na obrázku 13.1. Vidíme, že relativní četnosti a teoretické pravděpodobnosti nejsou od sebe příliš vzdáleny, na ověření souladu empirických dat s hypotézou $X \sim \text{Po}(2)$ bude však nutné použít rigoróznější techniky matematické statistiky, a to konkrétně statistického testu. $\triangle$

13.2 Rozdělení výběrových statistik

Definice 13.3 (Výběrový průměr, rozptyl a směrodatná odchylka) *Nechť $X_1, \dots, X_n$ je náhodný výběr z rozdělení nějaké náhodné veličiny X . Zavedme náhodné veličiny*

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2, \quad S = \sqrt{S^2}.$$

$\bar{X}$ nazýváme výběrovým průměrem, S^2 výběrovým rozptylem a S výběrovou směrodatnou odchylkou.

Věta 13.1. *Nechť $X_1, \dots, X_n$ je náhodný výběr z rozdělení, které má střední hodnotu μ a konečný rozptyl σ^2 , pak*

$$\mathbb{E}\bar{X} = \mu, \quad \text{var } \bar{X} = \frac{\sigma^2}{n}, \quad \mathbb{E}S^2 = \sigma^2.$$

Náhodné veličiny $\bar{X}$, S^2 a S reprezentují empirické charakteristiky z pohledu teorie pravděpodobnosti, které odhadují teoretické charakteristiky střední hodnoty, rozptylu a směrodatné odchylky náhodné veličiny X . Hodnoty aritmetického průměru $\bar{x}$, výběrového rozptylu s_n^2 a směrodatné odchylky s_n spočítané z datového souboru jsou pak v tomto kontextu realizacemi náhodných veličin $\bar{X}$, S^2 a S – říkáme jim také odhady nebo přesněji odhadnuté hodnoty. Věta 13.1 pak bude v kapitole 14 podkladem pro vyhodnocení kvality těchto odhadů.

Věta 13.2 (silný zákon velkých čísel). *Nechť $X_1, \dots, X_n$ je náhodný výběr z rozdělení, které má střední hodnotu μ a konečný rozptyl σ^2 , pak*

$$\bar{X} \rightarrow \mu \quad \text{skoro jistě.}$$

Konvergence skoro jistě znamená, že existuje pouze množina ($A \subset \Omega$) pravděpodobnosti 0 ($\mathbb{P}(A) = 0$), pro kterou výraz nekonzverguje.

Věta 13.3 (o náhodném výběru z normálního rozdělení). *Nechť $X_1, \dots, X_n$ je náhodný výběr z $\mathcal{N}(\mu; \sigma^2)$, kde $\sigma^2 > 0$. Pak platí následující tvrzení:*

- $\bar{X} \sim \mathcal{N}(\mu; \frac{\sigma^2}{n})$.
- Je-li $n \geq 2$, pak $(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2$.
- Je-li $n \geq 2$, pak $\bar{X}$ a S^2 jsou nezávislé.
- Je-li $n \geq 2$, pak $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$.

Důkazy výše uvedených vět může čtenář nalézt například v [2].

Tabulka 13.2: Značení kvantilů vybraných náhodných veličin

Rozdělení	Značení kvantilu	Hodnota kvantilu ³	Funkce Excelu
normované normální $N(0; 1)$	$u(\alpha)$	$\Phi^{-1}(\alpha)$	= NORM.S.INV
Pearsonovo χ_k^2	$\chi_k^2(\alpha)$	$F_{Q_k}^{-1}(\alpha)$	= CHISQ.INV
Studentovo t_k	$t_k(\alpha)$	$F_T^{-1}(\alpha)$	= T.INV
Fisherovo–Snedecorovo F_{ν_1, ν_2}	$F_{\nu_1, \nu_2}(\alpha)$	$F_F^{-1}(\alpha)$	= F.INV

13.3 Kritické hodnoty rozdělení

Kritickou hodnotou rozdělení se ve starších učebnicích obvykle označovala hodnota, kterou náhodná veličina překročí se zadanou pravděpodobností α . Kritickou hodnotou vybraného rozdělení je tedy jeho $(1 - \alpha)$ -kvantil. Například 5% kritická hodnota standardního normálního rozdělení je přibližně hodnota 1,64, tedy 95% kvantil této náhodné veličiny¹.

Pro určení kritických hodnot rozdělení se používaly statistické tabulky s ustáleným značením, které ovšem již zastaralo² a je v kolizi se značením kvantilů, které jsme zavedli v kapitole 8. Proto pojem kritické hodnoty nadále vyhradíme pouze pro vlastnost statistického testu, zavedeme jej v kapitole 15), a pojmu kritická hodnota rozdělení se budeme nadále vyhýbat. Namísto toho v tabulce 13.2 shrnujeme značení kvantilů, které jsme zavedli pro vybrané náhodné veličiny a které využijeme v kapitole 15 o testování statistických hypotéz. Připomeňme též dvě důležité vlastnosti kvantilů, které se rovněž budou hodit při testování hypotéz.

- 1) Pro kvantily symetrických rozdělení, tedy pro normované normální a Studentovo rozdělení platí

$$u(1 - \alpha) = -u(\alpha), \text{ resp. } t_k(1 - \alpha) = -t_k(\alpha), \quad (13.1)$$

tedy kvantily příslušející oběma chvostům rozdělení se neliší hodnotou, ale pouze znaménkem.

- 2) Kvantil Fisherova–Snedecorova rozdělení s prohozenými stupni volnosti je převrácenou hodnotou kvantilu rozdělení s původním pořadím

¹Platí tedy $\mathbb{P}(X > 1,64) = 1 - \Phi(1,64) \doteq 1 - 0,95 = 0,05$.

²Například kritická hodnota odpovídající 97,5% kvantilu Studentova rozdělení se v těchto tabulkách typicky značila $t_k(0,05)$.

³ F^{-1} je kvantilová funkce příslušného rozdělení, Φ^{-1} je kvantilová funkce standardního normálního rozdělení.

stupňů volnosti na jeho opačném chvostu:

$$F_{\nu_2, \nu_1}(\alpha) = \frac{1}{F_{\nu_1, \nu_2}(1 - \alpha)}. \quad (13.2)$$

Příklad 13.3.  Firma vyrábějící plastové láhve zjistila, že hmotnost láhví je normálně rozložena s průměrem $\mu = 50$ g a směrodatnou odchylkou $\sigma = 2$ g. Automatický kontrolní systém označí vyrobené láhve za nekvalitní, pokud jejich hmotnost spadá mezi 5% nejlehčích a 5% nejtěžších láhví. Jaké budou hodnoty mezních hmotností jedné láhve?

Řešení. Náhodná veličina X udávající hmotnost vyrobené láhve se řídí normální rozdělením s parametry $\mu = 50$ a $\sigma^2 = 4$. Mezní hodnoty x_1 a x_2 jsou určeny nerovnostmi

$$\mathbb{P}(X \leq x_1) \leq 0,05, \quad \text{resp.} \quad \mathbb{P}(X \geq x_2) \geq 0,05.$$

Hodnoty x_1 , x_2 jsou tedy 5%, resp. 95% kvantilem předepsaného rozdělení, které snadno spočteme pomocí funkcí

$$\begin{aligned} x_1 &= \boxed{= \text{NORM.INV}(0,05;50;2)} \doteq 46,71, \\ x_2 &= \boxed{= \text{NORM.INV}(0,95;50;2)} \doteq 53,29. \end{aligned}$$

Kvalitně vyrobená láhev musí mít hmotnost aspoň 46,71 g a nejvýše 53,29 g.

△

13.4 Úlohy

13.1 Stanovte hodnoty kvantilů pro $\alpha = 0,05$, $\nu = 8$, $\mu = 11$:

- a) $u(\alpha)$, $u(1 - \alpha)$, $u(\frac{\alpha}{2})$, $u(1 - \frac{\alpha}{2})$,
- b) $t_\nu(\alpha)$, $t_\nu(1 - \alpha)$, $t_\nu(\frac{\alpha}{2})$, $t_\nu(1 - \frac{\alpha}{2})$,
- c) $\chi_\nu^2(\alpha)$, $\chi_\nu^2(1 - \alpha)$, $\chi_\nu^2(\frac{\alpha}{2})$, $\chi_\nu^2(1 - \frac{\alpha}{2})$,
- d) $F_{\nu,\mu}(\alpha)$, $F_{\mu,\nu}(\alpha)$, $F_{\nu,\mu}(1 - \alpha)$, $F_{\nu,\mu}(\frac{\alpha}{2})$, $F_{\nu,\mu}(1 - \frac{\alpha}{2})$.

13.2 Pro některé ze spojitých rozdělení načrtněte při zvolených parametrech graf hustoty pravděpodobnosti a graf distribuční funkce a vyznačte hodnotu 5% a 95% kvantilu.

13.3 Předpokládejme, že denní ztráta portfolia X se při finanční krizi řídí normálním rozdělením se střední hodnotou 0,015 a rozptylem 0,010. Spočítejte 1% a 5% hodnotu v riziku tohoto portfolia, tedy výši ztráty, která bude překročena s pravděpodobností 1 %, resp. 5 %.

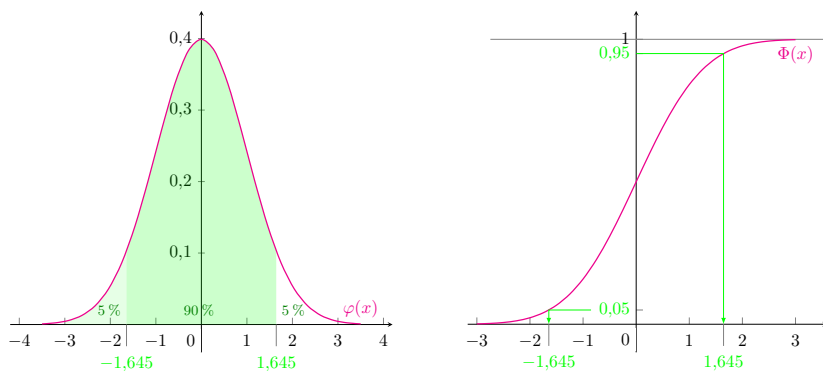
13.4 Vypočítejte z -skór pro hodnotu $x = 120$ z normálního rozdělení se střední hodnotou 110 a rozptylem 100.

Řešení úloh

13.1 Hodnoty kvantilů po zaokrouhlení na tisíce:

- a) $-1,645$; $1,645$; $-1,960$; $1,960$;
- b) $-1,860$; $1,860$; $-2,306$; $2,306$;
- c) $2,733$; $15,507$; $2,180$; $17,535$;
- d) $0,302$; $0,339$; $2,948$; $0,236$; $3,664$.

13.2 Obrázek 13.2 ukazuje 5% a 95% kvantil pro $X \sim N(0; 1)$.



Obrázek 13.2: 5% a 95% kvantil standardního normálního rozdělení

13.3 0,249; 0,179.

13.4 $z = 1$

Kapitola 14

Odhady parametrů

Charakteristiky základního statistického souboru jsou vyjádřeny vlastnostmi rozdělení argumentu statistického souboru. Například jednorozměrné rozdělení může být charakterizováno mírou polohy (typicky střední hodnotou), variabilitou rozdělení (směrodatnou odchylka), těsnost závislosti dvou- či vícerozměrného rozdělení lze charakterizovat koeficientem korelace apod. Rozdělení znaku základního souboru je dále specifikováno zpravidla parametry tohoto rozdělení. Například pro rekonstrukci normálního rozdělení $N(\mu; \sigma^2)$ je potřeba a stačí odhadnout jeho parametry μ a σ^2 , tedy střední hodnotu a rozptyl. Některé parametry rozdělení jsou současně jeho charakteristikami, jako v případě zmíněného normálního rozdělení.

Cílem *statistické indukce* je odhad charakteristik a parametrů základního statistického souboru prostřednictvím odhadu parametrů rozdělení náhodné veličiny představující argument základního souboru. Odhady konstruuje na výběrovém souboru, přičemž ke každé charakteristice či parametru rozdělení lze na výběrovém souboru určit tzv. výběrový protějšek (viz tabulku 14.1).

Rozlišujeme dva druhy odhadů:

- bodové odhady jsou takové odhady, při kterých je odhadem charakteristiky či parametru základního souboru výběrová charakteristika či parametr, tedy jedno číslo, například střední hodnotu podkladové náhodné veličiny μ odhadujeme pomocí průměru $\bar{x}$ výběrového souboru,
- intervalové odhady neboli intervaly spolehlivosti jsou odhady, při kterých odhadujeme charakteristiku či parametr základního souboru pomocí intervalu, odhadem je tedy interval, resp. jeho meze.

Formálně lze definovat odhad T parametru θ jako (abstraktní) funkci náhodného výběru. Na odhady je potřeba nahlížet ze dvou pohledů, podobně

Tabulka 14.1: Některé populační parametry a jejich výběrové protějšky

Parametr	Populační parametr	Výběrový parametr
Střední hodnota	μ	$\bar{x}$
Rozptyl	σ^2	s_n^2
Směrodatná odchylka	σ	s_n
Medián	$\tilde{\mu}$	$\tilde{x}$
Modus	$\hat{\mu}$	$\hat{x}$
Korelační koeficient	ρ	r

jako jsme tak učinili u náhodného výběru na straně 259: buď je odhadem číslo, které je spočtené z dat, nebo odhad označuje náhodnou veličinu odvozenou z náhodného výběru.

Definice 14.1 (Odhad parametru)

Nechť $X_1, \dots, X_n$ je náhodný výběr z rozdělení, které je charakterizováno parametrem θ . Odhad T parametru θ je libovolná známá funkce náhodného výběru

$$T = h(X_1, \dots, X_n).$$

Realizovaná hodnota náhodné veličiny T se nazývá odhadnutou hodnotou.

Odhadnutá hodnota se v češtině označuje krátce rovněž výrazem odhad, ale jedná se o dva různé matematické objekty. Význam je však vždy zřejmý z kontextu: odhad je náhodná veličina, odhadnutá hodnota spočtené číslo.

14.1 Kvalita odhadu

Odhadem parametru náhodné veličiny je dle definice 14.1 kterákoliv funkce náhodného výběru. Odhadem střední hodnoty náhodné veličiny tedy může být i kupříkladu maximum výběrového souboru. Je ovšem zřejmé, že toto maximum neodhaduje střední hodnotu příliš dobře – maximum výběrového souboru není dobře zvoleným odhadem střední hodnoty. Kvalitu odhadu lze posuzovat z několika hledisek, které popíšeme v tomto odstavci.

Definice 14.2 (Nestrannost odhadu)

Nechť T je odhadem parametru θ . Jestliže platí

$$\mathbb{E}T = \theta,$$

potom takový odhad nazýváme nestranným (též nevychýleným) odhadem parametru θ . Odhad, který není nestranný, nazýváme odhadem vychýleným a rozdíl $\mathbb{E}T - \theta$ se nazývá výchylka odhadu.

Nestranné odhady při opakovaných výběrech kolísají kolem teoretické hodnoty, a to tak, že „v průměru“ (ve střední hodnotě) dávají tuto teoretickou hodnotu. Nestrannost tedy není vlastnost konkrétní odhadnuté hodnoty, ale vlastnost rozdělení náhodné veličiny T , konkrétně informace o tom, že funkce h odhad T definující je určena v jistém smyslu korektně.

Věta 13.1 nám říká, že výběrový průměr $\bar{X}$ je nestranným odhadem střední hodnoty μ , jelikož $\mathbb{E}\bar{X} = \mu$. Podobně výběrový rozptyl S^2 je nestranný odhad populačního rozptylu σ^2 , jelikož $\mathbb{E}S^2 = \sigma^2$.

Nestrannost je velmi silná vlastnost: zdaleka ne všechny rutinně používané odhady jsou nestranné. Například výběrová směrodatná odchylka S není nestranným odhadem populační směrodatné odchylky σ , přestože pro rozptyl to platí. Navzdory tomu je S standardně používaným odhadem směrodatné odchylky. Minimální požadavek pro použitelnost odhadu je stanoven následující definicí.

Definice 14.3 (Konzistence odhadu)

Nechť T_n je odhadem parametru θ založeným na náhodném výběru $X_1, \dots, X_n$ o rozsahu n . T_n nazýváme konzistentním odhadem, jestliže pro libovolné $\varepsilon > 0$ platí

$$\mathbb{P}\{|T_n - \theta| > \varepsilon\} \xrightarrow{n \rightarrow \infty} 0. \quad (14.1)$$

Vlastnost (14.1) v definici 14.3 říká, že odhad T_n konverguje k teoretické hodnotě parametru θ podle pravděpodobnosti ve smyslu definice 11.1, část b). Z hlediska interpretace definice říká, že s rostoucím počtem pozorování se rozdělení odhadu T_n více koncentruje kolem θ : pro zvyšující se rozsah výběru je stále méně a méně pravděpodobné, že odhadnutá hodnota bude daleko (dále než ε) od teoretického parametru θ .

Každý užitečný odhad by měl být konzistentní. Nekonzistentní odhady nijak nepomáhají k odhadnutí hodnoty θ , a to ani s nekonečným množstvím dat. Výše zmíněné maximum výběrového souboru se s rostoucím počtem pozorování dokonce místo přibližování teoretické hodnotě od odhadované střední hodnoty vzdaluje. Pro ověření konzistence odhadu pomáhá užitečné kritérium, jež uvádíme bez důkazu.

Věta 14.1 (Konzistence odhadů). *Je-li T_n nestranný odhad parametru θ a $\text{var } T_n \xrightarrow{n \rightarrow \infty} 0$, potom je T_n konzistentním odhadem.*

Důsledek 14.2. *Výběrový průměr $\bar{X}$ je konzistentním odhadem střední hodnoty μ .*

Důkaz. Plyne z věty 13.1: $\mathbb{E}\bar{X} = \mu$ a $\text{var } \bar{X} = \frac{1}{n}\sigma^2 \rightarrow 0$. □

V definici 12.8 jsme ve vzorci (12.4) zavedli popisnou statistiku (populační) rozptyl. Lze ukázat, že náhodná veličina $\Sigma^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ není nestranným odhadem parametru σ^2 , ale je odhadem konzistentním. Stejně tak výběrová i populační směrodatná odchylka S , resp. Σ jsou konzistentními odhady teoretické směrodatné odchylky σ .

Zmínili jsme, že kvalitní odhady nemusí být nestranné (ale jsou vždy konzistentní). Na druhou stranu, nestrannost nemusí vždy nutně znamenat

kvalitní odhad: střední hodnotu teoretického lze odhadnout i pomocí prvního prvku náhodného výběru (tj. $T_n = X_1$). T_n je nestranný odhad ($\mathbb{E}T_n = \mu$ z definice náhodného výběru), avšak je jasné, že dostatečně nevytěžuje informaci, kterou náhodným výběrem dostáváme (zahodili jsme všechna pozorování vyjma prvního)¹. Chybu odhadu můžeme nejsnáze charakterizovat pomocí rozptylu odhadu.

Definice 14.4 (Vydatnost odhadu)

Vydatnost (eficience) odhadu T je charakterizována prostřednictvím rozptylu $\text{var } T$. Odhad je vydatný (eficientní), jestliže ve skupině uvažovaných odhadů má nejnižší rozptyl.

Pokračujeme-li s odhadem $T_n = X_1$, jeho rozptyl je $\text{var } T_n = \text{var } X_1 = \sigma^2$, kdežto odhad definovaný aritmetickým průměrem má rozptyl $\text{var } \bar{X} = \frac{1}{n}\sigma^2$. Například při deseti pozorováních je rozptyl aritmetického průměru desetkrát nižší než rozptyl X_1 a při rostoucím počtu pozorování se nadále rozptyl aritmetického průměru snižuje.

14.2 Bodové odhady střední hodnoty a rozptylu

Zopakujme, že věta 13.1 nám říká, že $\bar{X}$ je nestranným odhadem střední hodnoty μ a S^2 je nestranným odhadem teoretického rozptylu σ^2 . Spolu s jejich konzistencí máme důvod domnívat se, že $\bar{X}$ a S^2 by byly dobrými bodovými odhady parametrů μ a σ^2 a skutečně je tak v praxi stanovujeme. Stejně tak za bodový odhad teoretické směrodatné odchylky σ klademe výběrovou směrodatnou odchylku S , i když se nejedná² o nestranný odhad σ . S dostatečně velkým rozsahem výběru však i zde převáží vlastnost konzistence a poskytne poměrně přesný odhad teoretické směrodatné odchylky.

14.3 Intervalové odhady

Hodnota bodového odhadu je po realizaci náhodného výběru jediným číslem. Je-li odhad konzistentní, víme že při dostatečně velkém rozsahu výběru leží blízko skutečné teoretické hodnoty. Známe-li rozptyl odhadu, jsme

¹Šlo by ukázat, že X_1 není konzistentní, ale cíl této pasáže je jiný.

²Ve skutečnosti nestranný odhad směrodatné odchylky neexistuje. V literatuře lze nalézt alternativní konstrukce odhadů pro teoretickou směrodatnou odchylku s cílem snížit výchylku odhadu, ale jsou zatíženy řadou omezujících podmínek na populaci a vlastnosti náhodného výběru a nikdy se nestaly rutinně používanými.

zhruba schopni říci i jak blízko (použijeme k tomu interpretaci směrodatné odchylky odhadu). K ještě lepšímu určení (ne)přesnosti odhadu používáme *intervalový odhad*, což je ve interval odvozený z náhodného výběru³ pokrývající teoretickou hodnotu parametru s předepsanou pravděpodobností.

Definice 14.5 (Interval spolehlivosti, intervalový odhad)

Nechť jsou Θ_L, Θ_U takové statistiky odvozené z náhodného výběru z rozdělení s parametrem θ , že pro dané $\alpha \in (0; 1)$ platí

$$\mathbb{P}(\Theta_L \leq \theta \leq \Theta_U) = 1 - \alpha.$$

Potom interval $[\Theta_L, \Theta_U]$ nazýváme intervalem spolehlivosti (konfidenčním intervalem) pro parametr θ o spolehlivosti $1 - \alpha$. Interval $[\hat{\theta}_L, \hat{\theta}_U]$ určený realizacemi obou statistik se nazývá $100(1 - \alpha)\%$ intervalový odhad parametru θ .

Nejčastěji se α volí 0,05 nebo 0,01; ve výjimečných případech se volí hodnoty nižší, když potřebujeme mít zaručenou vysokou jistotu pokrytí skutečné teoretické hodnoty parametru. Cenou za takové pokrytí je ovšem šířka výsledného intervalu. Používá se také názvu interval $100(1 - \alpha)\%$ spolehlivosti pro parametr θ nebo názvu konfidenční interval pro parametr θ se $100(1 - \alpha)\%$ spolehlivostí.

Konstrukce intervalu spolehlivosti, narozdíl od bodových odhadů, vyžaduje znalost třídy rozdělení podkladové náhodné veličiny. V následujících odstavcích představíme intervalové odhady pro střední odchylku, resp. směrodatnou odchylku normálního rozdělení.

14.3.1 Intervalový odhad střední hodnoty normálního rozdělení

Uvažujme, že jsme získali náhodný výběr $X_1, \dots, X_n$ pocházející z populace, která má normální rozdělení $N(\mu; \sigma^2)$. Odhadujeme parametr střední hodnoty μ a předpokládáme, že parametr $\sigma^2 > 0$ není znám. Pak podle věty 13.3 platí

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

tudíž podle definice kvantilů Studentova rozdělení je

$$\mathbb{P}\left(-t_{n-1}\left(1 - \frac{\alpha}{2}\right) \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{n-1}\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha.$$

³Jeho meze jsou tedy rovněž náhodné veličiny.

Přeuspořádáním a využitím vlastnosti (13.1) dostaneme oboustranný intervalový odhad pro střední hodnotu μ normálního rozdělení o spolehlivosti $1 - \alpha$

$$\left[\bar{X} - t_{n-1} \left(1 - \frac{\alpha}{2} \right) \frac{S}{\sqrt{n}}, \bar{X} + t_{n-1} \left(1 - \frac{\alpha}{2} \right) \frac{S}{\sqrt{n}} \right]. \quad (14.2)$$

Podobně lze odvodit také jednostranné intervalové odhady μ :

- levostranný intervalový odhad μ

$$\left[\bar{X} - t_{n-1}(1 - \alpha) \frac{S}{\sqrt{n}}, \infty \right), \quad (14.3)$$

- pravostranný intervalový odhad μ

$$\left(-\infty, \bar{X} + t_{n-1}(1 - \alpha) \frac{S}{\sqrt{n}} \right]. \quad (14.4)$$

Výraz $\frac{S}{\sqrt{n}}$ se nazývá *směrodatná chyba*, výraz $t_{n-1}(1 - \alpha) \frac{S}{\sqrt{n}}$ se nazývá *přípustná chyba* a často se značí Δ . Oboustranný intervalový odhad střední hodnoty má pak zjednodušený tvar $[\bar{X} - \Delta; \bar{X} + \Delta]$.

Příklad 14.1.  U dvaceti náhodně vybraných zaměstnanců firmy byly zjišťovány údaje o délce doby (v minutách), kterou musí denně věnovat dojíždění do práce. Zjištěné doby jsou následující: 45, 38, 50, 42, 47, 40, 43, 49, 41, 44, 45, 38, 50, 42, 47, 40, 43, 49, 41, 44. Předpokládejme normální rozdělení doby dojíždění a nezávislost této doby mezi jednotlivými zaměstnanci. Vypočtete oboustranný intervalový odhad pro průměrnou dobu dojíždění do práce s 95% spolehlivostí.

Řešení. Rozsah souboru $n = 20$, průměrná hodnota znaku $\bar{x} = 43,9$, směrodatná odchylka $s \doteq 3,796$, koeficient spolehlivosti $1 - \alpha = 0,95$, počet stupňů volnosti $n - 1 = 19$, kvantil Studentova rozdělení $t_{19}(0,025) \doteq -2,093$, směrodatná chyba $\frac{s}{\sqrt{n}} = \frac{3,796}{\sqrt{20}} \doteq 0,849$, přípustná chyba $\delta \doteq 1,777$, hledaný intervalový odhad je $[42,12; 45,68]$. Průměrná doba dojíždění zaměstnanců firmy bude s pravděpodobností 95 % ležet ve zjištěném intervalu. $\triangle$

14.3.2 Odhad rozsahu výběru pro danou šířku intervalového odhadu střední hodnoty

Někdy je potřeba odhadnout rozsah výběru n tak, abychom dostali šířku intervalového odhadu střední hodnoty v požadované délce.

Délka d intervalu je rozdílem horní a dolní meze (v tomto pořadí), proto ze vzorce (14.2) dostaneme pro šířku d oboustranného intervalového odhadu o spolehlivosti $1 - \alpha$ dostaneme

$$d = 2t_{n-1} \left(1 - \frac{\alpha}{2}\right) \frac{s}{\sqrt{n}},$$

který dále umocněním upravíme do tvaru

$$\begin{aligned} d^2 &= 4t_{n-1}^2 \left(1 - \frac{\alpha}{2}\right) \frac{s^2}{n}, \\ n &= 4t_{n-1}^2 \left(1 - \frac{\alpha}{2}\right) \frac{s^2}{d^2}. \end{aligned} \quad (14.5)$$

Podobně by se postupovalo i případě intervalového odhadu rozptylu, který představíme později.

Je-li rozsah souboru větší než 30, je možné nahradit $t_{n-1}(1 - \frac{\alpha}{2})$ kvantilem normovaného normálního rozdělení $u(1 - \frac{\alpha}{2})$.

Příklad 14.2. Vraťte se k příkladu 14.1 na straně 275. Vypočítejte, kolik musíme udělat měření, aby při stejné spolehlivosti byla šířka výsledného intervalového odhadu 1 minuta.

Řešení. Necht požadovaná šířka d intervalu spolehlivosti $1 - \alpha = 0,95$ je 1. V příkladu 14.1 jsme provedli dvacet měření, z nichž jsme odhadli $s \doteq 3,796$ a $\bar{x} = 43,9$. Použijeme kvantil Studentova rozdělení $t_{19}(0,975) \doteq 2,093$.

Po dosazení do vzorce 14.5 dostaneme

$$n \doteq 4t_{19}(0,975)^2 \frac{s^2}{d^2} = 4 \cdot 2,093^2 \cdot \frac{3,796^2}{1} \doteq 252,5.$$

Je tudíž nutné provést nejméně 253 měření, aby šířka výsledného intervalového odhadu byla přibližně 1.

V odstavci 8.8 jsme uvedli, že Studentovo rozdělení se zvětšujícím se stupněm volnosti n konverguje k normálnímu rozdělení. Je-li $n \geq 30$, lze nahradit kvantil $t_{n-1}(0,975)$ kvantilem normálního rozdělení $u(0,975) \doteq 1,96$; v tomto případě by výpočet vedl k výsledku 222 pozorování. V našem výběru ovšem máme jen 20 pozorování a nahrazení kvantilu ve výše uvedeném smyslu by vedlo k podcenění potřebného počtu pozorování. $\triangle$

14.3.3 Intervalový odhad rozptylu σ^2 normálního rozdělení

Při konstrukci intervalového odhadu pro rozptyl σ^2 normálního rozdělení vycházíme (obdobně jako v případě střední hodnoty) ze znalosti rozdělení náhodné veličiny

$$(n-1) \frac{S^2}{\sigma^2} \sim \chi_{n-1}^2.$$

Jelikož

$$\mathbb{P} \left(\chi_{n-1}^2 \left(\frac{\alpha}{2} \right) \leq (n-1) \frac{S^2}{\sigma^2} \leq \chi_{n-1}^2 \left(1 - \frac{\alpha}{2} \right) \right) = 1 - \alpha,$$

přeuspořádáním dostaneme oboustranný intervalový odhad pro rozptyl σ^2 normálního rozdělení o spolehlivosti $q = 1 - \alpha$

$$\left[\frac{S^2(n-1)}{\chi_{n-1}^2 \left(1 - \frac{\alpha}{2} \right)}; \frac{S^2(n-1)}{\chi_{n-1}^2 \left(\frac{\alpha}{2} \right)} \right]. \quad (14.6)$$

Opět je možné odvodit jednostranné odhady rozptylu:

- levostranný

$$\left[\frac{S^2(n-1)}{\chi_{n-1}^2(1-\alpha)}; \infty \right),$$

- pravostranný

$$\left(0; \frac{S^2(n-1)}{\chi_n^2(\alpha)} \right].$$

Příklad 14.3. [1]^x Při kontrole automatu, který má plnit cukrem balíčky o váze 1 kg, byly při převážení pěti balíčků zjištěny odchylky $-3, 2, -2, 0, -1$ v gramech od nominální hodnoty. Vypočítejte bodový a intervalový odhad směrodatné odchylky balicího automatu.

Řešení. Předpokladem pro bodový a intervalový odhad rozptylu σ^2 je kolekce nezávislých náhodných veličin s rozdělením $N(\mu; \sigma^2)$. Bodový odhad rozptylu σ^2 vypočítáme jako

$$S^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right)$$

$$s^2 = \frac{1}{4} \{ [(-3)^2 + 2^2 + (-2)^2 + 0^2 + (-1)^2] - 5(-0,8)^2 \} = 3,7.$$

= **VAR.S(B2:B6)** (naměřená data mějme v buňkách B2:B6).

Bodovým odhadem směrodatné odchylky σ je $s = \sqrt{s^2} \doteq 1,923\,5$.

= **SMODCH.VÝBĚR.S(B2:B6)**

Pro určení intervalového odhadu předpokládejme, že odchylky měření mají normální rozdělení. K určení mezí intervalového odhadu potřebujeme v tom případě kvantily $\chi^2_{n-1}(0,975) \doteq 11,14$ a $\chi^2_{n-1}(0,025) \doteq 0,48$, které zjistíme pomocí software.

= **CHISQ.INV(0,975;4)**, = **CHISQ.INV(0,025;4)**

Intervalový odhad o spolehlivosti $1 - \alpha = 0,95$ pro rozptyl je

$$\left[\frac{s^2(n-1)}{\chi^2_{n-1}(0,975)}; \frac{s^2(n-1)}{\chi^2_{n-1}(0,025)} \right] = \left[\frac{3,7 \cdot 4}{11,143}; \frac{3,7 \cdot 4}{0,484} \right] = [1,33; 30,55].$$

Meze intervalového odhadu pro směrodatnou odchylku jsou odmocninou mezí odhadu pro rozptyl, tj. intervalový odhad směrodatné odchylky je $[1,15; 5,53]$. $\triangle$

14.4 Intervalový odhad střední hodnoty pomocí CLV

V případě, že podkladová náhodná veličina a z ní konstruovaný náhodný výběr nemají normální rozdělení, nemůžeme použít výše definované intervalové odhady, protože v takovém případě náhodné veličiny

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \quad \text{a} \quad (n-1) \frac{S^2}{\sigma^2} \quad (14.7)$$

nemají požadované Studentovo, resp. Pearsonovo rozdělení. Je-li však výběrový soubor dostatečně velkého rozsahu, můžeme pak využít centrální limitní větu, která jednoduše řečeno říká, že součet většího počtu náhodných veličin se chová přibližně jako normální rozdělení. Důsledkem je, že veličiny (14.7) mají Studentovo, resp. Pearsonovo rozdělení alespoň přibližně a intervalové odhady zkonstruované v předchozích odstavcích je možné nadále použít. Pro použití této aproximace se obvykle doporučuje rozsah náhodného výběru $n \geq 30$.

Alternativní možností vycházející rovněž z centrální limitní věty je použít následující úvahu. Mějme $X_1, \dots, X_n$ náhodný výběr z rozdělení s konečnou



střední hodnotou μ a konečným rozptylem σ^2 . Pak podle centrální limitní věty má pro $n \rightarrow \infty$ statistika

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

asymptoticky normální rozdělení $N(0;1)$. Podle definice kritické hodnoty kvantilu normovaného normálního rozdělení je

$$\mathbb{P}\left(-u\left(1 - \frac{\alpha}{2}\right) \leq \frac{\bar{X} - \mu}{S} \sqrt{n} \leq u\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha.$$

Přeuspořádáním dostaneme oboustranný intervalový odhad pro střední hodnotu μ o spolehlivosti $1 - \alpha$

$$\left[\bar{X} - u\left(1 - \frac{\alpha}{2}\right) \frac{S}{\sqrt{n}}; \bar{X} + u\left(1 - \frac{\alpha}{2}\right) \frac{S}{\sqrt{n}}\right]. \quad (14.8)$$

Jednostranné odhady mají tvar:

- levostranný intervalový odhad spolehlivosti pro střední hodnotu μ

$$\left[\bar{X} - u(1 - \alpha) \frac{S}{\sqrt{n}}; \infty\right);$$

- pravostranný intervalový odhad spolehlivosti pro střední hodnotu μ

$$\left(-\infty; \bar{X} + u(1 - \alpha) \frac{S}{\sqrt{n}}\right].$$

Příklad 14.4.  Byl proveden pokus, při němž jsme 600krát hodili kostkou a z toho 75krát padla šestka. Zajímá nás odhad pravděpodobnosti padnutí šestky na této kostce.

Řešení. Zavedme si náhodné veličiny $X_1, \dots, X_{600}$ s alternativním rozdělením $A(p)$, kde úspěch ($X = 1$) nastane, když padne 6, a neúspěch ($X = 0$) nastane při výsledcích hoďu 1–5. Zajímá nás $p = \mathbb{P}[X = 1]$. Připomeňme, že $p = \mathbb{E}X$. Bodový odhad p je tedy $\bar{x} = \frac{75}{600} = 0,125$. Výběrový rozptyl je

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{599} (75 \cdot (1 - 0,125)^2 + 525 \cdot (0 - 0,125)^2) \doteq 0,11.$$

Pro výpočet intervalového odhadu o spolehlivosti $1 - \alpha = 95 \%$ potřebujeme znát ještě hodnotu $u(0,975) \doteq 1,96$.

$$= \text{NORM.S.INV}(0,975)$$

$$\left[\bar{x} - u(0,975) \frac{s}{\sqrt{n}}; \bar{x} + u(0,975) \frac{s}{\sqrt{n}}\right] = [0,098; 0,151].$$

Tudíž skutečná pravděpodobnost padnutí šestky na této kostce leží s pravděpodobností 0,95 v intervalu $[0,098; 0,151]$. Pokud by kostka byla symetrická, pak by tato pravděpodobnost byla $\frac{1}{6} \doteq 0,166$. Tato pravděpodobnost však v námi zjištěném intervalovém odhadu neleží, tudíž s pravděpodobností alespoň 95 % tato kostka není spravedlivá. Toto je identický závěr s testem hypotézy, který provedeme v příkladu 15.6. $\triangle$

14.5 Úlohy

14.1 Stanovte nestranné bodové odhady střední hodnoty výšky studentů a rozptylu výšky na základě dostupných výběrových dat: 167, 169, 186, 197, 184, 185, 159, 174, 175, 168, 193, 174, 182, 172, 184, 187, 203, 177, 181, 163, 175, 171, 202, 199, 182, 184, 185, 173, 172, 174, 183, 171, 169, 164, 179, 186, 169, 201, 168, 195, 172, 187, 180, 174, 165.

14.2 Na základě údajů v úloze 14.1 stanovte 95%, 90% a 99% oboustranný, jednostranný a jednostranný interval spolehlivosti pro

- a) střední hodnotu μ ,
- b) rozptyl σ^2 ,
- c) směrodatnou odchylku σ .

14.3 Kolik měření by bylo potřeba provést, abyste se při odhadu střední hodnoty μ výšky studentů z příkladu 14.2 dopustili chyby převyšující ± 3 cm? Určete pro spolehlivost 90 %, 95 % i 99 %.

14.4 Jaká je spolehlivost odhadu, použijte-li data z příkladu 14.1 a spokojíte-li se s chybou nejvýše ± 4 cm?

14.5 Při výlovu rybníka bylo náhodně vybráno a zváženo 15 kaprů v kg: 3,0; 3,1; 2,5; 2,5; 4,2; 2,1; 3,25; 2,5; 4,8; 2,3; 4,1; 3,6; 3,0; 3,6; 4,0.

- a) S 99% spolehlivostí odhadněte střední hodnotu hmotnosti kapra.
- b) Jakou minimální garantovanou nosnost musí mít taška, aby unesla kapra s 95% spolehlivostí?
- c) Pro publikaci výsledků výlovu je potřeba uvést výsledky s předem určenou přesností. Kolik je potřeba zvážit kaprů, aby bylo možno odhadnout střední hodnotu hmotnosti kapra s 99% spolehlivostí a maximální přípustnou chybou 0,25 kg?
- d) S jakou spolehlivostí lze odhadnout střední hodnotu váhy kapra v rybníce při maximální přípustné chybě 0,25 kg pouze na základě stávajícího výběru čítajícího 15 kaprů?

14.6 V souvislosti s kontrolou dodržování předpisů o rychlosti jízdy přes obec byla změřena rychlost 36 náhodně vybraných aut. Porovnejte odlišnost

intervalových odhadů průměrné rychlosti jízdy aut pro $\alpha = 0,05$ a použitím CLV.

14.7 Byla zjišťována hmotnost prasat poražených ve stáří 182 dní v kg: 124,3; 119,4; 117,8; 131,6; 118,2; 113,9; 135,1; 133; 138,4; 125,2; 129,9; 127,9; 131,8; 137,2; 119,7; 122,7; 111,4; 135,2; 129,1; 117,2; 134,3; 133,9; 129,8; 126,4; 123,5; 126,6; 121; 127,2; 117,8; 113,5; 126,6; 121,6; 122,1; 124,6; 121,4; 120,8; 122,3; 134,5; 110,1; 127; 140,5; 127,7; 137,5; 130,4; 136,5; 136,6; 124,5. Se spolehlivostí 95 % odhadněte zisk zemědělce z prodeje 60 prasat ve stáří 182 dní, jestliže se platí 45 Kč za 1 kg živé váhy. Porovnejte obvyklý intervalový odhad s intervalovým odhadem získaným na základě CLV.

14.8 V lesním porostu bylo náhodně vybráno a změřeno 80 průměrů kmene smrku v [cm]:

20, 29, 39, 29, 21, 26, 36, 18, 58, 35, 45, 49, 38, 20, 17, 42, 56, 46, 28, 59, 29, 46, 36, 45, 17, 68, 26, 23, 37, 22, 57, 48, 44, 25, 50, 21, 16, 74, 27, 28, 38, 50, 52, 26, 15, 69, 20, 16, 31, 19, 21, 40, 29, 18, 37, 45, 80, 29, 50, 20, 19, 59, 24, 19, 21, 24, 19, 38, 20, 25, 18, 15, 23, 42, 66, 55, 29, 27, 15, 49.

Výběr smrků dobře reprezentuje celou populaci smrků, která vykazuje normální rozdělení. Odhadněte s 95% spolehlivostí, kolik dřevní hmoty lze získat z 200 kmenů, které jsou 4 m vysoké? (Zúžení kmene zanedbejte.) Porovnejte obvyklý intervalový odhad s intervalovým odhadem získaným na základě CLV.

Řešení úloh

14.1 $\bar{x} = 179,1$; $s^2 \doteq 122,419$.

14.2

- a) intervalové odhady střední hodnoty μ : odhad: oboustranný, jednostranný, jednostranný
 spolehlivost 90 %: [176,340; 181,882], [176,965; ∞), ($-\infty$; 181,257],
 spolehlivost 95 %: [175,787; 182,435], [176,340; ∞), ($-\infty$; 181,882],
 spolehlivost 99 %: [174,671; 183,552], [175,129; ∞), ($-\infty$; 183,093],
- b) intervalové odhady rozptylu σ^2 :
 odhad: oboustranný, jednostranný, jednostranný
 spolehlivost 90 %: [89,060; 180,829], [95,060; ∞), [0; 165,802],
 spolehlivost 95 %: [83,899; 195,341], [89,060; ∞), [0; 180,829],
 spolehlivost 99 %: [74,923; 228,397], [78,394; ∞), [0; 214,190],

- c) intervalové odhady směrodatné odchylky σ :
odhad: oboustranný, levostranný, pravostranný
spolehlivost 90 %: $[9,437; 13,447]$, $[9,77; \infty)$, $[0; 12,876]$,
spolehlivost 95 %: $[9,160; 13,976]$, $[9,437; \infty)$, $[0; 13,447]$,
spolehlivost 99 %: $[8,656; 15,112]$, $[8,854; \infty)$, $[0; 14,635]$.

14.3 Po dosazení do vzorce pro přípustnou chybu $\Delta = t_{m-1} \left(1 - \frac{\alpha}{2}\right) \cdot \frac{S}{\sqrt{n}}$ určíme n , pro spolehlivost $q = 90\%$ – 38 měření, pro spolehlivost $q = 95\%$ – 55 měření a pro spolehlivost $q = 99\%$ – 120 měření.

14.4 $1 - \alpha \doteq 0,979$, tj. asi 97,9 %.

14.5

- a) Bodový odhad střední hodnoty – 3,237 kg, intervalový odhad střední hodnoty – $[2,624; 3,849]$.
b) 3,599 kg.
c) Asi 90,057, tj. asi 91 kaprů.
d) 0,756.

14.6 Pomocí statistiky s t -rozdělením je intervalový odhad $[49,69; 52,66]$, pomocí CLV je intervalový odhad $[49,74; 52,61]$, odhady se tedy příliš neliší.

14.7 Zisk zemědělce se bude se spolehlivostí 95 % pohybovat v intervalu $[335\,069,60; 347\,134,20]$ Kč; odhad podle CLV je $[335\,228,20; 346\,975,60]$ Kč, tedy lze říct, že odhad pomocí CLV je prakticky totožný s původním odhadem.

14.8 Předpokládané množství dřevní hmoty se bude se spolehlivostí 95 % pohybovat v intervalu $[60,3; 91,05]$ m³ dřeva; odhad pomocí CLV se příliš neliší: $[60,52; 90,8]$ m³ dřevní hmoty.

Kapitola 15

Parametrické testy

15.1 Principy testování hypotéz

Ve vědeckém výzkumu i v aplikacích se problémy často formulují ve tvaru hypotéz. Statistická hypotéza je tvrzení, které se týká pravděpodobnostního rozdělení, případně závislosti hodnot parametrů náhodné veličiny.

Hypotézy lze dělit na dva druhy:

- 1) parametrické hypotézy – hypotéza je určena pomocí parametru určitého modelu,
- 2) neparametrické hypotézy – hypotézy o tvaru rozdělení, o závislosti proměnných apod.

Každá úloha testování hypotéz je formulována tak, že proti sobě stojí dvě hypotézy, a to *nulová hypotéza* (H_0) proti hypotéze alternativní (H_1 nebo také H_A). V této kapitole se budeme zabývat pouze parametrickými testy, tzn. budeme předpokládat znalost pravděpodobnostního rozdělení příslušné náhodné veličiny a testovat budeme parametr daného rozdělení.

Předpokládejme, že rozdělení náhodné veličiny závisí na parametru θ . O parametru θ se domníváme, že by mohl být roven danému číslu θ_0 . V tomto případě nulovou hypotézu zapisujeme ve tvaru

$$H_0: \theta = \theta_0.$$

Alternativní hypotéza H_1 může být ve tvaru

$$H_1: \theta \neq \theta_0, \quad \text{nebo} \quad H_1: \theta > \theta_0, \quad \text{nebo} \quad H_1: \theta < \theta_0.$$

V prvním případě se jedná o tzv. oboustrannou hypotézu, ve druhém a třetím případě jde o jednostrannou (přesněji pravostrannou, resp. levostrannou hypotézu).

Tabulka 15.1: Chyby při testování hypotéz

Rozhodnutí Skutečnost	Zamítáme H_0	Nezamítáme H_0
H_0 platí	α chyba I. druhu	$1 - \alpha$ správné rozhodnutí
H_0 neplatí	$1 - \beta$ správné rozhodnutí	β chyba II. druhu

Při svém rozhodnutí o platnosti H_1 či H_0 se můžeme dopustit jedné ze dvou chyb. Stane-li se, že zamítneme H_0 , ačkoli ve skutečnosti platí, uděláme tzv. *chybu prvního druhu* (falešná pozitivita). Stane-li se, že nezamítneme H_0 , přestože ve skutečnosti neplatí, uděláme tzv. *chybu druhého druhu* (falešná negativita). Při testování samozřejmě požadujeme, aby pravděpodobnosti obou chyb byly co možná nejmenší. Pravděpodobnost chyby prvního druhu typicky označujeme α a nazýváme ji *hladina významnosti testu*, pravděpodobnost chyby druhého druhu označujeme β . Pravděpodobnost $1 - \beta$ je označována také jako „síla testu“ nebo „rozhlišovací schopnost“ testu a představuje pravděpodobnost, že správně zamítneme nulovou hypotézu H_0 , pokud skutečně neplatí. Síla testu obvykle závisí na předem zvolené hladině významnosti testu α , a to tak, že s klesající hladinou významnosti síla testu klesá. Přehled o chybách a jejich pravděpodobnostech při testování ukazuje tabulka 15.1.



Při rozhodování o platnosti nulové či alternativní hypotézy se opíráme o tzv. testovou statistiku T . Testová statistika je předem daný funkční předpis, který závisí na náhodném výběru $X_1, X_2, \dots, X_n$ z podkladového rozdělení. Hodnoty statistiky T mohou ležet v jedné ze dvou disjunktních množin, a to buď v kritickém oboru W (představuje obor zamítnutí hypotézy H_0) nebo v oboru přijetí V (lépe obor nezamítnutí hypotézy H_0). Jak už bylo řečeno, můžeme se při testování dopustit jedné ze dvou chyb, přičemž se obvykle trvá jen na požadavku, aby pravděpodobnost chyby prvního druhu byla rovna nebo menší než α , kde α je



nějaké dané číslo z intervalu $(0; 1)$. V praxi se nejčastěji volí $\alpha = 0,05$ nebo $\alpha = 0,01$. Obor zamítnutí W a obor nezamítnutí V jsou odděleny tzv. *kritickými hodnotami testu*. Požadujeme tedy, aby platilo

$$\mathbb{P}(T \in W | H_0) \leq \alpha.$$

V současné době udává běžný statistický software (Statistica, R, S+, SAS, ale i MS Excel) tzv. *p-hodnotu*, neboli dosaženou hladinu testu. *p-hodnota* je nejmenší hladina testu, při které bychom ještě hypotézu H_0 zamítli. Tudíž zvolíme-li $\alpha = 0,05$ a *p-hodnota* vyjde menší než 0,05 (nebo rovna), pak zamítáme hypotézu H_0 na hladině $\alpha = 0,05$. Pokud *p-hodnota* vyjde větší než 0,05, pak hypotézu H_0 na hladině $\alpha = 0,05$ zamítnout nelze. „Nezamítnutí“ hypotézy H_0 však neznamená, že hypotézu potvrzujeme, přijímáme za platnou, tedy že tvrzení v ní obsažené je pravdivé, nýbrž „nezamítnutím“ nulové hypotézy pouze nevylučujeme její platnost. Říkáme také, že pozorovaná data nejsou s H_0 v rozporu.



Z tabulky 15.1 plyne, že zamítne-li test hypotézu H_0 , pak se můžeme dopustit pouze chyby prvního druhu, jejíž pravděpodobnost kontrolujeme hladinou významnosti α a víme, že je nízká. Zatímco nezamítne-li test H_0 , pak se můžeme dopustit pouze chyby druhého druhu, kterou však nemáme výše uvedeným postupem pod kontrolou. Proto je tvrzení „zamítneme H_0 “ silnější než tvrzení „nezamítneme H_0 “. To je vyjádřeno i lingvistickou asymetrií těchto vyjádření. Nezamítneme-li H_0 , pak to může být způsobeno různými důvody. Buďto H_0 opravdu platí, nebo prostě jen nemáme dostatečné množství dat, tedy dostatečnou informaci k tomu, abychom mohli H_0 zamítnout.

Testování hypotéz lze přirovnat k principu presumpce neviny: jestliže výběrový soubor neukáže na rozpor s nulovou hypotézou, pak nelze nulovou hypotézu zamítnout – podobně jako u presumpce neviny se požaduje, aby na obžalovaného bylo pohlíženo jako na nevinného do té doby, dokud se nepředloží přesvědčivé důkazy o jeho vině. Soudcem v případě je pak statistický test: statistický test rozhodne, zda data z výběrového souboru odpovídají nulové hypotéze stejně jako soudce rozhodne, zda svědci podali výpověď ve prospěch obhajoby. Požaduje se, aby $\mathbb{P}(T \in W | H_0) = \alpha$, tj. aby svědci (výběrový soubor) podali falešné svědectví v neprospěch obhajoby za situace, kdy je obžalovaný nevinný (tedy za podmínky platnosti H_0), pouze s pravděpodobností α . (Rozdíl s principem presumpce neviny je, že

soudce nemůže úroveň pravděpodobnosti α stanovit a ani ji pro konkrétní případ nezná.)

Nejčastěji je testování hypotéz používáno k porovnávání jednoho výběru s danou hodnotou, nebo dvou či více výběrů mezi sebou. Testujeme-li s pomocí pouze jediného náhodného výběru, mluvíme o jednovýběrovém testu, pokud s pomocí více výběrů, mluvíme o dvouvýběrovém testu či testu z více výběrů. V druhém případě je vždy potřeba rozlišovat, zda se jedná o náhodné výběry závislé nebo nezávislé. U nezávislých výběrů se předpokládá, že jednotky vybrané z jednoho souboru nezávisí na vybraných jednotkách z druhého souboru. Jsou-li vybrané jednotky z jednoho souboru pozorovány na stejných objektech jako jednotky z druhého souboru, případně jsou svázány jiným způsobem, pak jsou jednotky obou souborů na sobě závislé; v případě dvou souborů mluvíme o párovém testu.

15.2 (Studentův) Jednovýběrový t -test střední hodnoty

Nechť $X_1, \dots, X_n$, je náhodný výběr z $N(\mu; \sigma^2)$ s rozsahem $n > 1$. Parametr $\sigma^2 > 0$ není znám. Ve Studentově oboustranném jednovýběrovém t -testu je potřeba testovat hypotézu $H_0: \mu = \mu_0$, kde μ_0 je dané číslo, proti alternativě $H_1: \mu \neq \mu_0$. Hypotézu H_0 zamítneme, bude-li $\bar{X}$ hodně vzdáleno od čísla μ_0 . Z věty 13.3 víme, že za platnosti hypotézy H_0 má statistika

$$T = \frac{(\bar{X} - \mu_0)}{S/\sqrt{n}} \sim t_{n-1}$$

Studentovo rozdělení o $n - 1$ stupních volnosti. Podle definice kritické hodnoty Studentova rozdělení dostaneme, že

$$\mathbb{P}\left(|T| \geq t_{n-1}\left(1 - \frac{\alpha}{2}\right)\right) = \alpha.$$

Tedy hypotézu H_0 zamítneme na hladině α , jestliže platí

$$|t| \geq t_{n-1}\left(1 - \frac{\alpha}{2}\right).$$

V případě jednostranných alternativ $H_1: \mu > \mu_0$ (pravostranný test), resp. $H_1: \mu < \mu_0$ (levostranný test) hypotézu H_0 zamítneme, jestliže

$$t \geq t_{n-1}(1 - \alpha), \quad \text{resp.} \quad t \leq -t_{n-1}(1 - \alpha).$$

Příklad 15.1.  Vraťme se k příkladu 14.3 na straně 277. Má se rozhodnout, zda má automat systematickou výchylku. Tudíž je třeba testovat hypotézu $H_0 : \mu = 0$ proti alternativě $H_1 : \mu \neq 0$ na hladině $\alpha = 0,05$ (tj. že odchylky kolísají kolem nuly a nejsou systematicky posunuty ani do kladných, ani do záporných hodnot).

Řešení. Vzhledem k zadání a formulaci hypotéz využíváme oboustrannou variantu testu. Máme $n = 5$, $\mu_0 = 0$, $\bar{x} = -0,8$, $s \doteq 1,923\,5$,

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \doteq -0,93.$$

Kritickou hodnotu t -rozdělení s obecně $n - 1$ stupni volnosti (zde $5 - 1 = 4$) $t_4(0,975) \doteq 2,776$ určíme z tabulek nebo výpočtem v MS Excel.

 `= T.INV(0,975;4)`

Protože $|-0,93| \not\geq t_4(0,975) \doteq 2,776$, nezamítáme H_0 . Tudíž zjištěná data neodporují předpokladu, že automat nemá systematickou odchylku. Všimněte si, že $\mu_0 = 0$ (střední hodnota za platnosti hypotézy H_0) leží uvnitř intervalového odhadu o spolehlivosti 95 % (viz příklad 14.3). Neboli 0 je pravděpodobná hodnota skutečné střední hodnoty a tudíž nemůžeme zamítnout H_0 . Oba přístupy k testování hypotéz, jak klasický přístup, tak přes intervalový odhad, jsou ekvivalentní. $\triangle$

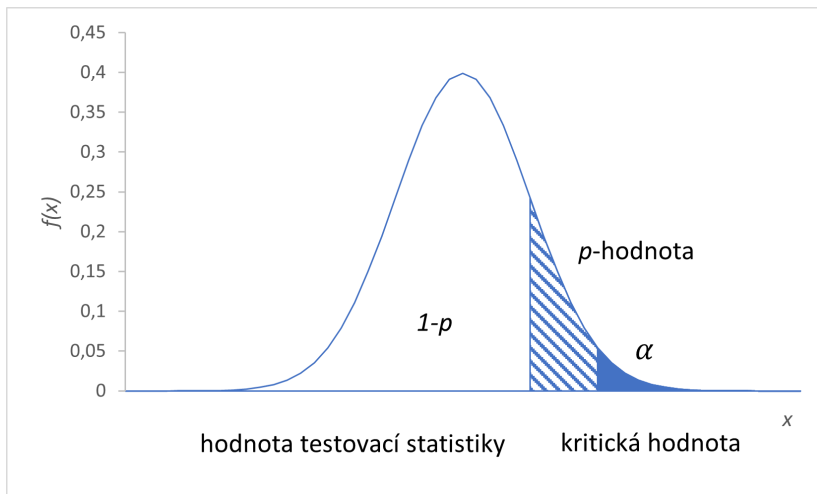
15.2.1 Výpočet oboustranné a jednostranné p -hodnoty t -testu

Již jsme zmínili, že p -hodnota nebo také dosažená hladina významnosti (signifikance) testu je číselná hodnota, pomocí níž můžeme alternativně rozhodovat při statistickém testování hypotéz. Testujeme-li hypotézu H_0 pomocí nějaké statistiky T , lze na p -hodnotu nahlížet dvěma ekvivalentními způsoby:

- 1) p -hodnota je nejmenší hladina významnosti, při které ještě zamítneme H_0 ,
- 2) p -hodnota je podmíněná pravděpodobnost, že při platnosti H_0 nabývá testová statistika T dosažené hodnoty t nebo hodnot v absolutní hodnotě větších (pro oboustrannou alternativní hypotézu)¹.

Při řešení úloh si většinou nejdříve stanovíme hladinu významnosti α , pak si vypočítáme p -hodnotu a tyto dvě hodnoty porovnáme. Vyjde-li p -hodnota menší než zvolená hladina významnosti ($p < \alpha$), zamítneme H_0 , v opačném případě ($p > \alpha$) ji nezamítneme. Čím menší je tedy p -hodnota, tím za méně věrohodnější lze označit H_0 .

¹V případě pravostranné alternativy hodnot vyšších, v případě levostranné alternativy hodnot menších než t .



Obrázek 15.1: Vztah mezi α a p -hodnotou pro pravostranný test (zde na příkladu hustoty standardizovaného normálního rozdělení U)

Podle druhu testování rozlišíme tři případy:

- V případě oboustranného testu a symetrického rozdělení testového kritéria představuje p -hodnota dvojnásobek obsahu plochy pod grafem hustoty pravděpodobnosti testové statistiky napravo od absolutní hodnoty testové statistiky a vypočítá se obecně jako $2 \cdot [1 - F(|t|)]$, v případě asymetrického rozdělení platí $p = 2 \cdot \min\{F(t), 1 - F(t)\}$.
- V případě pravostranného testu je p -hodnota dána obsahem plochy pod grafem hustoty pravděpodobnosti testové statistiky napravo od její hodnoty a vypočítá se jako rozdíl $1 - F(t)$, kde $F(t)$ je hodnota distribuční funkce pro dosaženou hodnotu t testové statistiky.
- V případě levostranného testu je p -hodnota dána obsahem plochy pod grafem hustoty pravděpodobnosti testové statistiky nalevo od její hodnoty a vypočítá se jako $F(t)$.

Ve všech třech případech hustotou, resp. distribuční funkcí F testové statistiky míníme hustotu, resp. distribuční funkci testové statistiky za podmínky, že platí H_0 . Konkrétně p -hodnoty například pro jednovýběrový t -test se vypočítají pomocí vzorce

$$p = 2 \cdot [1 - F(|t|)]$$

pro oboustrannou variantu testu a pomocí

$$p = F(t), \quad \text{resp.} \quad p = 1 - F(t)$$

pro levostrannou, resp. pravostrannou variantu testu.

15.3 Test rozptylu normálního rozdělení

Jednovýběrový test o rozptylu slouží k ověření hypotézy o rozptylu dat, tj. zda testovaná hodnota rozptylu je totožná s populačním ekvivalentem, tj. zda jsou při zvolené hladině významnosti hodnoty ve výběru stejně proměnlivé jako v populaci.

Nechť $X_1, \dots, X_n$ je náhodný výběr z $N(\mu; \sigma^2)$, kde $n > 1$. Je třeba testovat hypotézu

$$H_0: \sigma^2 = \sigma_0^2 \quad \text{proti alternativě} \quad H_1: \sigma^2 \neq \sigma_0^2,$$

kde σ_0^2 je dané kladné číslo. Hypotézu H_0 zamítneme, bude-li S^2 hodně vzdáleno od čísla σ_0^2 . Z věty 13.3 víme, že za platnosti hypotézy H_0 má statistika Q rozdělení χ^2 o $n - 1$ stupních volnosti:

$$Q = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi_{n-1}^2.$$

Podle definice kritické hodnoty kvantilů rozdělení χ^2 dostaneme, že

$$\mathbb{P} \left(\chi_{n-1}^2 \left(\frac{\alpha}{2} \right) \leq (n-1) \frac{S^2}{\sigma_0^2} \leq \chi_{n-1}^2 \left(1 - \frac{\alpha}{2} \right) \right) = 1 - \alpha,$$

tedy hypotézu H_0 zamítneme na hladině α , jestliže pro realizaci q statistiky Q platí:

$$q \leq \chi_{n-1}^2 \left(\frac{\alpha}{2} \right) \quad \text{nebo} \quad q \geq \chi_{n-1}^2 \left(1 - \frac{\alpha}{2} \right).$$

V případě jednostranné alternativy $H_1: \sigma^2 > \sigma_0^2$ (pravostranná varianta), resp. $H_1: \sigma^2 < \sigma_0^2$ (levostranná varianta) hypotézu H_0 zamítneme, jestliže

$$q \geq \chi_{n-1}^2(1 - \alpha), \quad \text{resp.} \quad q \leq \chi_{n-1}^2(\alpha).$$

p -hodnotu testu o rozptylu výběru z normálního rozdělení vypočítáme vzhledem k asymetrii χ^2 rozdělení

$$p = 2 \cdot \min\{F(q); 1 - F(q)\},$$

kde $\min$ značí minimum z vypočtených dvou hodnot a $F(q)$ je hodnota distribuční funkce pro vypočtenou statistiku q .

p -hodnoty jednostranné varianty testu (pravostranný, resp. levostranný) rozptylu normálního rozdělení se vypočítají podle vzorců

$$p = F(q), \quad p = 1 - F(q).$$

Příklad 15.2.  Zácvek laboranta na určitém optickém přístroji považujeme za ukončený, jestliže při měření určitého objektu dosahuje rozptylu nejvýše 0,0196. Byly naměřeny hodnoty: 6,82; 6,44; 6,38; 6,21; 6,38; 6,60; 6,32. Zjistěte, zda je nutné pokračovat v zácviku.

Řešení. V tomto případě je potřeba provést test s jednostrannou alternativou. Zajímá nás, zda je potřeba pokračovat v zácviku, tj. zda s vysokou mírou jistoty $\sigma^2 > 0,0196$. Tuto situaci volíme jako alternativní hypotézu. Nulová hypotéza H_0 je formulována ve tvaru rovnosti. Tudíž testujeme $H_0: \sigma^2 = 0,0196$ proti alternativě $H_1: \sigma^2 > 0,0196$. Spočteme výběrový rozptyl $s^2 \doteq 0,0406$

= VAR.S(B3:B9)

a statistiku $q \doteq \frac{6s^2}{0,0196} \doteq 12,44$. Míru jistoty stanovíme volbou hladiny (zde v obvyklé výši) $\alpha = 0,05$. Kritická hodnota testu je v tomto případě $\chi^2_6(0,05) \doteq 12,59$,

= CHISQ.INV(0,95;6)

a protože $q \not\geq \chi^2_6(0,05)$, nemůžeme zamítnout H_0 .

Všimněte si, že odhad rozptylu (0,0406) je větší než požadovaný rozptyl (0,0196), ale naměřená data nám neumožňují zamítnout hypotézu, že rozptyl je roven požadované hodnotě na hladině $\alpha = 0,05$. Neboli máme více než 5% pravděpodobnost, že naměřená data mohla vzniknout z normálního rozdělení s rozptylem 0,0196. Ke stejnému závěru dojdeme také při řešení užitím p -hodnoty: $p \doteq 0,05286 > 0,05$,

= 1-CHISQ.DIST(12,44;6)

tedy na hladině $\alpha = 0,05$ nezamítáme H_0 . Všimněte si však rovněž, že p -hodnota převyšuje hladinu významnosti testu α jen velmi nepatrně – při jen o málo slabší jistotě, např. pro hladinu $\alpha = 0,06$ (ve skutečnosti pro každou hladinu alespoň 0,05286) bychom již nulovou hypotézu zamítlí. $\triangle$

15.4 Párový t -test rovnosti středních hodnot

Mějme náhodný výběr $(Y_1, Z_1), (Y_2, Z_2), \dots, (Y_n, Z_n)$ z nějakého dvourozměrného rozdělení, jehož vektor středních hodnot je (μ_1, μ_2) . Chceme testovat hypotézu

$$H_0: \mu_2 - \mu_1 = \Delta \quad \text{proti alternativě} \quad H_1: \mu_2 - \mu_1 \neq \Delta$$

kde Δ je nějaké dané číslo, nejčastěji $\Delta = 0$. Položíme

$$X_1 = Z_1 - Y_1, \quad X_2 = Z_2 - Y_2, \dots, X_n = Z_n - Y_n.$$

Pořadí členů v odčítání náhodných veličin $Z - Y$ volíme podle interpretace: pokud jsou hodnoty Y zjišťovány před aplikací a hodnoty Z po aplikaci nějakého opatření (například), pak X představuje zvýšení měřené veličiny.

Veličiny $X_1, X_2, \dots, X_n$ jsou nezávislé. Předpokládejme, že $X_i \sim N(\mu; \sigma^2)$, $i = 1, 2, \dots, n$. Zřejmě $\mu = \mu_2 - \mu_1$. Jsou-li tyto předpoklady splněny, pak je úloha převedena na jednovýběrový t -test: z realizace $x_1, x_2, \dots, x_n$ vypočteme $\bar{x}$, s^2 a hodnotu testové statistiky t . Hypotézu H_0 zamítneme na hladině α , platí-li

$$|t| = \left| \frac{\bar{x} - \Delta}{s/\sqrt{n}} \right| \geq t_{n-1} \left(1 - \frac{\alpha}{2} \right).$$

Párový t -test se používá v situacích, kdy na každém z n objektů byla naměřena data dvou různých veličin. Jednotlivé objekty lze zpravidla pokládat za nezávislé, ale měření na témž objektu nikoli. Párový t -test použijeme, když například testujeme účinnost nějakého léku na n pacientech, přičemž Y_i jsou hodnoty naměřené před podáním léku a Z_i jsou hodnoty naměřené po podání léku, tedy může jít o výběry vzniklé například při tzv. pretestu a posttestu. Párový test by se využil také v případech, kdy na dvou částech jednoho objektu měříme stejnou veličinu. Typickým příkladem je zjišťování, zda se po správném seřízení geometrie vozidla sjíždí obě pneumatiky stejně.

p -hodnoty pro oboustranný párový t -test se vypočítají podle vzorce

$$p = 2 \cdot (1 - F(|t|)),$$

pro jednostranné varianty testu

$$p = 1 - F(|t|), \quad \text{resp. } p = F(|t|).$$

p -hodnotu jednostranného i oboustranného testu lze vypočítat také pomocí funkce `= T.TEST(matice1;matice2;chvosty;typ)` v MS Excel, kde pod „chvosty“ volíme buď 1 (jednostranné rozdělení) nebo 2 (dvojstranné rozdělení) a u „typu“ volíme možnost 1 (párový test).

Příklad 15.3.  Na hladině $\alpha = 0,05$ se má rozhodnout, zda lék na snížení krevního systolického tlaku je účinný či nikoli. Proto bylo vybráno 6 pacientů, jimž byl změřen systolický tlak před aplikací léku a hodinu po aplikaci léku. Větší z obou hodnot měření systolického tlaku každého pacienta je zaznamenána v tabulce 15.2.



Tabulka 15.2: Větší z hodnot naměřeného tlaku pacientů

Tlak / Pacient	1	2	3	4	5	6
Před podáním léku	180	160	150	165	170	175
Po podání léku	150	155	155	150	155	170
Rozdíl hodnot	30	5	-5	15	15	5

Řešení. Rozdíly měření budeme považovat za realizace nezávislých náhodných veličin s rozdělením $N(\mu; \sigma^2)$, kde σ^2 není známo. Pokud lék nemá vliv na tlak krve, platí hypotéza $H_0: \mu = 0$; je-li lék účinný, platí alternativní hypotéza $H_1: \mu > 0$. Máme tedy $n = 6$, $\Delta = 0$ a

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{6} \cdot (30 + 5 - 5 + 15 + 15 + 5) \doteq 10,833.$$

= PRŮMĚR(D4:D9)

$$\begin{aligned} s^2 &= \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right) \doteq \\ &\doteq \frac{1}{5} \cdot (30^2 + 5^2 + 5^2 + 15^2 + 15^2 + 5^2 - 6 \cdot (-10,833)^2) \doteq \\ &\doteq 144,167. \end{aligned}$$

= VAR.S(D4:D9)

$$s = \sqrt{s^2} \doteq \sqrt{144,167} \doteq 12,007.$$

= SMOUCH.VÝBĚR.S(D4:D9)

Vypočítáme hodnotu testové statistiky T :

$$t = \frac{\bar{x} - \Delta}{s/\sqrt{n}} \doteq 2,21.$$

V tabulkách zjistíme nebo v MS Excel vypočítáme kritickou hodnotu $t_5(0,05) \doteq 2,015$.

= T.INV(0,95;5)

Protože $t \geq t_5(0,05)$, na základě uvedených měření zamítáme hypotézu H_0 ve prospěch H_1 , s vysokou pravděpodobností je tedy lék účinný.

p -hodnotu jednostranného testu lze spočítat v MS Excel přímo pomocí funkce = T.TEST takto:

Dvouvýběrový párový t -test na střední hodnotu		
	Soubor 1	Soubor 2
Stř. hodnota	166.6666667	155.8333333
Rozptyl	116.6666667	54.16666667
Pozorování	6	6
Pears. korelace	0.167725574	
Hyp. rozdíl stř. hodnot	0	
Rozdíl	5	
t Stat	2.210066303	
P(T<=t) (1)	0.039048983	
t krit (1)	2.015048373	
P(T<=t) (2)	0.078097966	
t krit (2)	2.570581836	

Obrázek 15.2: Výřez z MS Excel nástroje Analýza dat pro dvouvýběrový párový t -test

`= T.TEST(matice1;matice2;1;1)` $\doteq 0,039\,049 < \alpha$. Pravděpodobnost, že při neúčinnosti léku bychom dostali hodnotu testové statistiky 2,21 (nebo více) je pouze 3,9 %, méně než 5% hladina významnosti testu, tedy opět na této hladině významnosti zamítáme H_0 .

K řešení příkladu lze využít nástroj Analýza dat v MS Excel: na kartě Data zvolíme Analýza dat a dále možnost Dvouvýběrový párový t -test na střední hodnotu. Nejdříve vybereme příslušné soubory dat, hypotetický rozdíl středních hodnot předpokládáme roven 0 (není nutné zadávat) a zvolíme hladinu významnosti alfa (je přednastaveno 0,05). Vybereme buňku, do které chceme umístit začátek výsledné tabulky, a klikneme OK.

Obrázek 15.2 ukazuje tabulku, která je výsledkem nástroje Analýza dat pro data z příkladu 15.3. V tabulce lze kromě testové statistiky a kritické hodnoty pro zvolenou hladinu významnosti nalézt také výše vypočítanou p -hodnotu pro jednostrannou i oboustrannou alternativu.

△

15.5 Dvouvýběrový t -test rovnosti průměrů při stejných rozptylech

Uvažujme situaci, že máme k dispozici dva soubory nepárových pozorování různého rozsahu. Předpokládejme, že soubory mají sobě rovné rozptyly. V takovém případě pro porovnání výběrových průměrů obou souborů použijeme dvouvýběrový t -test, někdy též označovaný jako dvouvýběrový Studentův test.

Nechť $X_1, X_2, \dots, X_n$ je náhodný výběr z normálního rozdělení $N(\mu_1; \sigma^2)$ a $Y_1, Y_2, \dots, Y_m$ náhodný výběr z normálního rozdělení $N(\mu_2; \sigma^2)$. Nechť tyto dva výběry jsou na sobě nezávislé a všimněme si též, že předpokládáme totožný (ne nutně známý) rozptyl, o kterém předpokládáme $\sigma^2 > 0$. Předpokládejme dále, že $n \geq 2$ a $m \geq 2$. Chceme testovat hypotézu

$$H_0: \mu_1 - \mu_2 = \Delta \quad \text{proti} \quad H_1: \mu_1 - \mu_2 \neq \Delta,$$

kde Δ je nějaké dané číslo (nejčastěji $\Delta = 0$). Označme $\bar{X}$, S_X^2 a $\bar{Y}$, S_Y^2 výběrové průměry a výběrové rozptyly těchto výběrů. Lze ukázat (viz [1]), že za platnosti H_0 má testová statistika T ve tvaru

$$T = \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{(n-1)S_X^2 + (m-1)S_Y^2}} \cdot \sqrt{\frac{nm(n+m-2)}{n+m}} \sim t_{n+m-2},$$

Studentovo rozdělení t o $n+m-2$ stupních volnosti. Hypotézu H_0 tedy zamítneme na hladině α , jestliže

$$|t| \geq t_{n+m-2} \left(1 - \frac{\alpha}{2}\right).$$

Dvouvýběrový t -test hypotéz používáme v případech, kdy se například na n pacientech zkouší působení léku A a na jiných m pacientech působení léku B. Účelem pokusu je zjistit, zda je působení obou léků stejné.

Často dochází k záměně párového a dvouvýběrového t -testu, což je hrubá chyba. Dvouvýběrový t -test můžeme použít pouze v případě, když máme zajištěnu nezávislost všech veličin $X_1, X_2, \dots, X_n, Y_1, Y_2, \dots, Y_m$. V případě záměny těchto testů dojdeme zpravidla k nesmyslným výsledkům.

p -hodnotu pro oboustrannou variantu testu vypočítáme podle vzorce

$$p = 2 \cdot (1 - F(|t|)),$$

pro jednostranné varianty testu

$$p = 1 - F(|t|), \quad \text{resp.} \quad p = F(|t|).$$

Pro přímý výpočet p -hodnoty jednostranného i oboustranného testu lze využít také funkci `=T.TEST(matice1;matice2;chvosty;typ)` v MS Excel, kde pod „chvosty“ volíme buď 1 (jednostranné rozdělení) nebo 2 (oboustranné rozdělení), u „typu“ volíme možnost 2 (dva výběry se shodným rozptylem).

Při řešení úloh na testování rovnosti průměrů pomocí dvouvýběrového t -testu lze využít nástroj Analýza dat v MS Excel: na kartě Data zvolíme Analýza dat a dále možnost Dvouvýběrový t -test s rovností rozptylů.

Nejdříve vybereme příslušné soubory dat, hypotetický rozdíl středních hodnot ponecháme na výchozí hodnotě 0 (není nutné zadávat) a stejně tak ponecháme hladinu významnosti alfa na výchozí hodnotě 0,05 (nechceme-li pracovat s jinou hladinou). Vybereme buňku, do které chceme umístit začátek výsledné tabulky, a klikneme OK.

Pro testy z odstavců 15.2 až 15.5 platí určité předpoklady. Nejdůležitějším z nich je nezávislost jednotlivých veličin, v případě dvouvýběrového testu i obou výběrů. Porušení má závažné důsledky a činí závěry založené na testech chybnými. V praxi k porušení tohoto předpokladu dojde v případě, kdy realizovaný výběr není náhodný, tedy nereprezentující zkoumanou populaci (např. systematicky pomíjí některé podskupiny populace).

Dalším předpokladem testů je normalita rozdělení. Vzhledem k platnosti centrální limitní věty (CLV) a zákonů velkých čísel není její porušení při větším rozsahu náhodného výběru závažné a všechny parametrické testy se při velkém rozsahu náhodného výběrů rutinně používají. V odstavci 15.8 je pro rozsáhlé výběry rovněž představen alternativní test založený na CLV, který normalitu ani nepředpokládá. Při závažném porušení normality a malém rozsahu náhodného výběru dáváme přednost použití některého neparametrického testu. Testy na normalitu náhodného výběru jsou uvedeny v odstavcích 20.2 a 20.4.

U dvouvýběrového *t*-testu je ještě dodatečný požadavek na rovnost rozptylů obou rozdělení. Vždy v tomto případě usilujeme o testování takových souborů dat, u nichž není rozdíl ve velikosti rozptylů příliš veliký. Shodu rozptylů můžeme otestovat pomocí testu v následujícím odstavci.

15.6 Test rovnosti dvou rozptylů (*F*-test)

Nechť $X_1, X_2, \dots, X_n$ je výběr z $N(\mu_1; \sigma_1^2)$ a $Y_1, Y_2, \dots, Y_m$ výběr z $N(\mu_2; \sigma_2^2)$. Nechť tyto dva výběry jsou na sobě nezávislé. Předpokládejme, že $n \geq 2$, $m \geq 2$, $\sigma_1^2 > 0$, $\sigma_2^2 > 0$. Testujeme hypotézu

$$H_0: \sigma_1^2 = \sigma_2^2 \quad \text{proti} \quad H_1: \sigma_1^2 \neq \sigma_2^2.$$

Protože S_X^2 je nestranný odhad parametru σ_1^2 a S_Y^2 parametru σ_2^2 , lze očekávat, že za platnosti hypotézy H_0 bude podíl

$$F = \frac{S_X^2}{S_Y^2}$$

blízký 1. Proti H_0 budou tedy svědčit buď hodnoty blízké nule nebo hodnoty velké.

Tabulka 15.3: Průměrné výnosy polí [t/ha] pro příklad 15.4

Hnojivo A	62	54	55	60	53	58
Hnojivo B	52	56	49	50	51	-

Statistika F se podle části b) věty 13.3 a definičního vzorce (8.19), jako podíl dvou nezávislých náhodných veličin s rozdělení χ^2 , řídí Fisherovým–Snedecorovým rozdělením F o $n - 1$ a $m - 1$ stupních volnosti. Hypotézu H_0 zamítneme, jestliže

$$f = \frac{s_X^2}{s_Y^2} \leq F_{n-1, m-1}\left(\frac{\alpha}{2}\right) \quad \text{nebo} \quad f = \frac{s_X^2}{s_Y^2} \geq F_{n-1, m-1}\left(1 - \frac{\alpha}{2}\right).$$

kde $F_{n-1, m-1}(\cdot)$ jsou hodnoty kvantilu Fisherova–Snedecorova rozdělení F o $n - 1$ a $m - 1$ stupních volnosti.

p -hodnoty pro oboustranný F -test se vypočítají podle vzorce

$$p = 2 \cdot \min\{F(f); 1 - F(f)\},$$

a pro jednostranný F -test podle vzorce

$$p = 1 - F(f) \quad \text{pro } f > 1, \quad p = F(f) \quad \text{pro } f < 1.$$

Pro přímý výpočet p -hodnoty oboustranného F -testu se v MS Excel využívá funkce `= F.TEST(matice1;matice2)`.

Příklad 15.4.  Zemědělci oseli jedenáct polí, z toho u šesti polí použili hnojivo A a u zbylých pěti polí použili hnojivo B. Po sklizni byly u každého pole stanoveny průměrné výnosy (v tunách na hektar) (viz tabulku 15.3). Je potřeba ověřit, zda jsou obě hnojiva stejně efektivní.

Řešení. Průměrné výnosy v první skupině budeme pokládat za výběr z rozdělení $N(\mu_1; \sigma_1^2)$, průměrné výnosy ve druhé skupině za výběr z $N(\mu_2; \sigma_2^2)$. Nejprve prověříme hypotézu o shodě rozptylů obou výběrů

$$H'_0: \sigma_1^2 = \sigma_2^2 \quad \text{proti} \quad H'_1: \sigma_1^2 \neq \sigma_2^2.$$

Vypočteme $s_X^2 = 12,8$, $s_Y^2 = 7,3$ a hodnotu testové statistiky

$$f = \frac{s_X^2}{s_Y^2} = \frac{12,8}{7,3} \doteq 1,753.$$

Hypotézu o rovnosti rozptylů bychom zamítlí, kdyby platilo buď

$$f \leq F_{(5,4)}(0,025) \doteq 0,135,$$

Dvouvýběrový F-test pro rozptyl		
	Soubor 1	Soubor 2
Stř. hodnota	57	51.6
Rozptyl	12.8	7.3
Pozorování	6	5
Rozdíl	5	4
F	1.753424658	
P(F<=f) (1)	0.303172533	
F krit (1)	6.256056502	

Obrázek 15.3: Tabulka F -testu v MS Excel pro příklad 15.4

= F.INV(0,025;5;4), což není pravda, nebo

$$f \geq F_{(5,4)}(0,975) \doteq 9,364,$$

= F.INV(0,975;5;4), což také není pravda². p -hodnotu oboustranné varianty F -testu vypočítáme jako

= F.TEST(A2:A7;B2:B6) $\doteq 0,606\,345$ (s daty v buňkách A2:A7 a B2:B6), nebo také

$$= \text{MIN}(1-\text{F.DIST}(1,753;5;4); \text{F.DIST}(1,753;5;4)) \doteq 0,606\,345.$$

Protože je p -hodnota testu větší než α , nezamítáme H_0 na zvolené hladině významnosti.

K řešení příkladu lze využít nástroj Analýza dat v MS Excel: na kartě Data zvolíme Analýza dat a dále možnost Dvouvýběrový F -test na rozptyl. Nejdříve vybereme příslušné soubory dat, hypotetický rozdíl středních hodnot předpokládáme roven 0 (není nutné zadávat) a zvolíme hladinu významnosti alfa (je přednastaveno 0,05). Vybereme buňku, do které chceme umístit začátek výsledné tabulky, a klikneme OK. Tabulka na obrázku 15.3 ukazuje výsledky řešení pro příklad 15.4. Tabulka však neobsahuje kritickou hodnotu ani p -hodnotu pro požadovaný oboustranný test; oboustrannou p -hodnotu nicméně můžeme dostat jako prostý dvojnásobek publikované jednostranné p -hodnoty.

Hypotézu o rovnosti rozptylů tedy nezamítneme, a tudíž můžeme přistoupit k samotnému dvouvýběrovému t -testu hypotézy

$$H_0: \mu_1 = \mu_2 \quad \text{proti} \quad H_1: \mu_1 \neq \mu_2$$

²Vzhledem k vlastnosti (13.2) symetrie kvantilů Fisherova–Snedecorova rozdělení je možné tuto úvahu zkrátit: jestliže výběry označíme tak, že první výběr je ten s vyšším rozptylem, stačí bez újmy na obecnosti ověřovat pouze druhou ze zmíněných nerovností.

za předpokladu shodných rozptylů. Máme $n = 6$, $m = 5$, $\Delta = 0$, $\bar{x} = 57$, $\bar{y} = 51,6$ a statistika T má hodnotu

$$t = \frac{\bar{x} - \bar{y} - \Delta}{\sqrt{(n-1)s_X^2 + (m-1)s_Y^2}} \cdot \sqrt{\frac{nm(n+m-2)}{n+m}} \doteq 2,771\,2.$$

Hodnotu testové statistiky T porovnáme s kritickou hodnotou

$$= \text{T.INV}(0,975;9) = t_9(0,975) \doteq 2,26,$$

a protože $|t| \geq t_9(0,975)$, zamítáme hypotézu H_0 o rovnosti efektivity hnojiva A a B. p -hodnota oboustranné varianty t -testu je

$$= 2 * (1 - \text{T.DIST}(\text{ABS}(2,7712);9;\text{PRAVDA})) \doteq 0,021\,71.$$

Pro přímý výpočet p -hodnoty jednostranné nebo oboustranné varianty t -testu lze také využít funkci

$$= \text{T.TEST}(\text{matice1};\text{matice2};2;2) \doteq 0,021\,71.$$

V případě potřeby p -hodnot pro jednostranné varianty testu³ bychom počítali takto:

$$= \text{T.DIST}(2,7712;9;\text{PRAVDA}) \doteq 0,989\,145 \text{ (levostranný test), resp.}$$

$$= 1 - \text{T.DIST}(2,7712;9;\text{PRAVDA}) \doteq 0,010\,855 \text{ (pravostranný test).}$$

K řešení příkladu lze využít nástroj Analýza dat v MS Excel: na kartě Data zvolíme Analýza dat a dále možnost Dvouvýběrový t -test na střední hodnotu. Nejdříve vybereme příslušné soubory dat, hypotetický rozdíl středních hodnot předpokládáme roven 0 (není nutné zadávat) a zvolíme hladinu významnosti alfa (je přednastaveno 0,05). Vybereme buňku, do které chceme umístit začátek výsledné tabulky, a klikneme OK. Tabulka na obrázku 15.4 ukazuje výsledky testování středních hodnot pro data z příkladu 15.4.

△

15.7 Testování rovnosti středních hodnot souborů s různými rozptyly

Pokud si nejsou rovny rozptyly obou výběrů pocházejících z normálního rozdělení na základě zjištění F -testu ($\sigma_1^2 \neq \sigma_2^2$), rovnost vůbec nepředpokládáme, případně jsou-li výběry různých rozsahů ($m \neq n$), využijeme pro porovnání průměrů mezi dvěma nezávislými výběry alternativní testy, např. Cochranův–Coxův test, anebo Welchův⁴ (také Aspinové–Welchův) test.

³Například pro ověření hypotézy, že hnojivo A je účinnější než B.

⁴Bernard Lewis Welch (1911–1989), britský statistik.

Dvouvýběrový t-test s rovností rozptylů		
	Soubor 1	Soubor 2
Stř. hodnota	57	51.6
Rozptyl	12.8	7.3
Pozorování	6	5
Společný rozptyl	10.35555556	
Hyp. rozdíl stř. hodnot	0	
Rozdíl	9	
t Stat	2.77122216	
P(T<=t) (1)	0.010855041	
t krit (1)	1.833112933	
P(T<=t) (2)	0.021710083	
t krit (2)	2.262157163	

Obrázek 15.4: Tabulka t -testu v MS Excel pro příklad 15.4

Předpokládejme, že $X_1, X_2, \dots, X_n$ je výběr z $N(\mu_1; \sigma_1^2)$ a $Y_1, Y_2, \dots, Y_m$ je výběr z $N(\mu_2; \sigma_2^2)$ nezávislý na prvním výběru. Zjistíme-li při testování rovnosti rozptylů obou výběrů, že $\sigma_1^2 \neq \sigma_2^2$, můžeme střední hodnoty porovnat přibližným testem. Nejprve se vypočte

$$S_X^2 = \frac{1}{m-1} \left(\sum_{i=1}^m X_i^2 - m\bar{X}^2 \right), \quad S_Y^2 = \frac{1}{n-1} \left(\sum_{j=1}^n Y_j^2 - n\bar{Y}^2 \right),$$

$$S = \sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}, \quad v_x = \frac{S_X^2}{m}, \quad v_y = \frac{S_Y^2}{n}.$$

a dále spočteme statistiku

$$T = \frac{\bar{X} - \bar{Y} - \Delta_0}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}}.$$

Testujeme $H_0: \mu_1 - \mu_2 = 0$ proti $H_1: \mu_1 - \mu_2 \neq 0$. Za platnosti H_0 má T přibližně Studentovo rozdělení. Dle Cochranova–Coxova pravidla zamítneme H_0 v případě, že platí

$$|t| \geq \frac{v_x t_{m-1} \left(1 - \frac{\alpha}{2}\right) + v_y t_{n-1} \left(1 - \frac{\alpha}{2}\right)}{v_x + v_y}. \quad (15.1)$$

Tento test má hladinu významnosti přibližně α . V MS Excel se pro ověření této hypotézy test aproximuje počet stupňů volnosti jako

$$\tilde{\mu} = \frac{\left(\frac{S_X^2}{m} + \frac{S_Y^2}{n}\right)^2}{\frac{(S_X^2/m)^2}{m-1} + \frac{(S_Y^2/n)^2}{n-1}}.$$

Tabulka 15.4: Průměrná bodová hodnocení studentů v příkladu 15.5

PM	90	85	88	89	94	91	79	83	87	88	91	90
Bez PM	67	90	71	95	88	83	72	66	75	86	93	84

p -hodnoty Welchova testu se vypočítají již standardně pomocí vzorců

$$p = 2 \cdot (1 - F(|t|))$$

pro oboustranný test a

$$p = F(t), \quad p = 1 - F(t)$$

pro jednostranné verze testu, kde F je distribuční funkce Studentova rozdělení o $\tilde{\mu}$ stupních volnosti.

Funkce **= T.TEST** v MS Excel počítá p -hodnotu Welchova testu s aproximovaným počtem stupňů volnosti $\tilde{\mu}$. Nástroj Analýza dat i ruční výpočet pomocí funkce **= T.DIST** vrátí méně přesnou p -hodnotu, neboť usekávají případné neceločíselné stupně volnosti na celé číslo. Rozdíl je ale velmi malý a obvykle nemá vliv na rozhodnutí o platnosti hypotézy H_0 .

Příklad 15.5.  Tabulka 15.4 ukazuje bodová hodnocení dvou skupin různých studentů z testu z matematiky. První skupina při přípravě na test využívala pomocný materiál (PM) od učitele, druhá skupina nikoliv. Otestujte na hladině významnosti $\alpha = 0,05$, zda pomocný materiál učitele přispěl k lepším průměrným výsledkům v testu.

Řešení. Čísla v obou výběrech pocházejí od různých studentů, výběry jsou tedy nezávislé. Předpokládáme, že výběry mají normální rozdělení. Nejprve provedeme shodu rozptylů obou výběrů. Stanovíme pomocné hypotézy $H'_0: \sigma_1 = \sigma_2$ (tj. rozptyly bodových výsledků jsou u obou skupin stejné) proti $H'_1: \sigma_1 \neq \sigma_2$ (rozptyly bodových výsledků obou skupin jsou různé). V MS Excel dostáváme hodnotu testové statistiky $f \doteq 0,156$ a kritické hodnoty asi 0,288 a 3,474, tedy hodnota f testové statistiky leží v kritickém oboru a H'_0 zamítáme. Také p -hodnota

$$= \text{F.TEST}(A2:A13;B2:B13) \doteq 0,004625 \leq 0,05$$

potvrzuje s vysokou mírou jistoty, že rozptyly obou výběrů si nejsou rovny.

Pro testování průměrů formulujeme nulovou hypotézu $H_0: \mu_1 = \mu_2$ (bodové výsledky jsou u obou skupin stejné) proti pravostranné alternativě $H_1: \mu_1 > \mu_2$ (bodové výsledky skupiny, která používala pomocný materiál, jsou v průměru lepší než výsledky skupiny bez pomůcek). Testová statistika

má po zaokrouhlení hodnotu $t \doteq 2,236$. Kritická hodnota pro Cochranův–Coxův pravostranný⁵ test je po zaokrouhlení 1,796; nástroj Analýza dat vrací kritickou hodnotu 1,761 (přesnější hodnota s neceločíselnými stupni volnosti je 1,758, tu však není snadné v Excelu spočítat). Ve všech případech H_0 zamítáme ve prospěch alternativy H_1 . Také p -hodnota Welchova testu

$$= \text{T.TEST}(A2:A13;B2:B13;1;3) \doteq 0,02085 \leq \alpha = 0,05$$

potvrzuje, že pomocný materiál učitele pomohl studentům k dosažení lepších výsledků. $\triangle$

15.8 Test střední hodnoty pomocí CLV

V případě, že náhodné veličiny výrazně nesplňují normalitu, nemůžeme použít předchozí testy, neboť testové statistiky T , Q a F nemají patřičná rozdělení. Je-li však náhodných veličin větší počet, můžeme využít centrální limitní větu (dále používáme zkratku CLV), podle které se součet většího počtu náhodných veličin chová jako normální rozdělení. Pro použití aproximace pomocí CLV se obvykle doporučuje rozsah náhodného výběru $n \geq 20$.

Mějme $X_1, \dots, X_n$ náhodný výběr z rozdělení s konečnou střední hodnotou μ a konečným rozptylem σ^2 . Je třeba testovat hypotézu $H_0: \mu = \mu_0$, kde μ_0 je dané číslo, proti alternativě $H_1: \mu \neq \mu_0$. Hypotézu H_0 zamítneme, bude-li $\bar{X}$ hodně vzdáleno od čísla μ_0 . Podle CLV má statistika

$$Z = \frac{\bar{X} - \mu_0}{\sigma_0 / \sqrt{n}}$$

za platnosti H_0 asymptoticky normované normální rozdělení. Podle definice kvantilu normovaného normálního rozdělení je asymptoticky

$$\mathbb{P}\left(|Z| \leq u\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha.$$

Tedy hypotézu H_0 zamítneme na hladině α , jestliže platí

$$|z| \geq u\left(1 - \frac{\alpha}{2}\right).$$

V případě jednostranné alternativy $H_1: \mu > \mu_0$, resp. $H_1: \mu < \mu_0$ hypotézu H_0 zamítneme, jestliže

$$z \geq u(1 - \alpha), \quad \text{resp.} \quad z \leq u(\alpha).$$

⁵Ve vzorci (15.1) použijeme $(1 - \alpha)$ -kvantily Studentova rozdělení.

V případě, že rozptyl σ_0^2 není znám, použijeme místo něj ve výpočtu statistiky Z jeho nestranný odhad S^2 . Takto modifikovaná statistika má rovněž asymptoticky normované normální rozdělení.

p -hodnota pro oboustrannou variantu testu se vypočítá podle vzorce

$$p = 2 \cdot (1 - \Phi(|z|)),$$

p -hodnoty pro jednostranné varianty testu se vypočítají podle vzorců

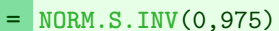
$$p = \Phi(z), \quad \text{resp.} \quad p = 1 - \Phi(z).$$

Příklad 15.6.  Vraťme se k příkladu 14.4. Má se rozhodnout o tom, zda kostka je z hlediska pravděpodobnosti padnutí šestky spravedlivá. Tudíž je třeba testovat hypotézu $H_0: \mu = \frac{1}{6}$ proti alternativě $H_1: \mu \neq \frac{1}{6}$ na hladině $\alpha = 0,05$.

Řešení. Máme $n = 600$, $\mu_0 = \frac{1}{6}$, $\bar{x} = 0,125$, $s^2 = 0,109$. Rozptyl náhodné veličiny $X_i \sim A(\mu_0)$ za platnosti hypotézy H_0 je $\sigma_0^2 = \mu_0(1 - \mu_0) = \frac{5}{36}$. Spočteme

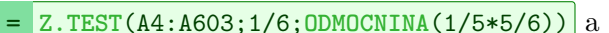
$$z = \frac{\bar{x} - \mu_0}{\sigma_0 / \sqrt{n}} \doteq -2,74. \quad (15.2)$$

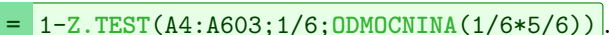
Vypočteme kritickou hodnotu $u(0,975) \doteq 1,96$.

 = NORM.S.INV(0,975)

Protože $|z| \geq u(0,975)$, zamítáme H_0 . Tudíž zjištěná data odporují předpokladu, že kostka je spravedlivá. Všimněte si, že u tohoto testu jsme použili σ_0^2 , zatímco v příkladu 14.4 jsme použili s^2 . Pokud bychom ve vzorci (15.2) také nahradili parametr σ_0^2 statistikou s^2 , byl by test ekvivalentní přístupu přes intervalový odhad. Obě možnosti jsou však z praktického hlediska rovnocenné.

Pro výpočet p -hodnoty testu lze využít funkci  = Z.TEST v MS Excel, která má tři parametry: oblast dat, testovanou střední hodnotu (x) a známou směrodatnou odchylku základního souboru ($sigma$). V našem případě vycházíme z oblasti A4:A603, která obsahuje 75 jedniček a 525 nul. p -hodnota pro oboustranný test je rovna dvojnásobku minima z hodnot

 = Z.TEST(A4:A603;1/6;ODMOCNINA(1/5*5/6)) a

 = 1-Z.TEST(A4:A603;1/6;ODMOCNINA(1/6*5/6)).

$p \doteq 0,00617$, tedy zamítáme H_0 . △

V současné době se v praxi již upouští od provádění výše uvedených testů založených na centrální limitní větě. Místo toho je centrální limitní věta teoretickým podkladem a zdůvodněním, proč je v případě rozsáhlých

výběrů nadále možné rutinně používat (exaktní) parametrické testy (t -testy, χ^2 -testy a F -testy), i když podkladové rozdělení není normální – všechny jsou asymptoticky ekvivalentní.

15.9 Úlohy

15.1 U náhodně vybraného vzorku osmi cigaret stejné značky byl zjištěn průměrný obsah nikotinu 4,2 mg a směrodatná odchylka 1,3 mg. Rozhodněte, zda výrobce neporušuje své reklamní tvrzení, že průměrný obsah nikotinu u tohoto typu cigaret nepřekračuje 3,5 mg. Zvolte hladinu významnosti 0,01 a předpokládejte normální rozdělení obsahu nikotinu v cigaretách.

15.2 Při kontrole automatu, který má plnit cukrem balíčky o váze 1 kg byly při převážení pěti balíčků zjištěny odchylky $-3, 2, -2, 0, -1$ v gramech od počáteční hodnoty (hladina významnosti 0,01). Zjistěte, zda automat vykazuje systematickou odchylku od požadované hodnoty. (Jednotlivé odchylky budeme považovat za realizace nezávislých náhodných veličin s rozdělením $N(\mu; \sigma^2)$, kde σ^2 je neznámý parametr.)⁶

15.3 Automat je nastaven na dávkování vitamínového premixu na úroveň 100 g na 50 kg krmné směsi. Technická kontrola vybrala náhodně padesát vzorků a zjistila přesnou hmotnost premixu. Výsledky jsou uvedeny v tabulce 15.5. Rozhodněte, zda se hmotnost dávky premixu významně ($\alpha = 0,05$, $\alpha = 0,01$) liší od požadovaného nastavení.

Tabulka 15.5: Data k příkladu 15.3

Dávka [g]	96	98	100	102	104	Celkem
Počet vzorků	7	29	9	3	2	50

15.4 U jedné skupiny osob byla sledována teplota těla a provedeno její měření před ošetřením a po něm. Výsledky měření ukazuje tabulka 15.6. Na hladině významnosti 0,01 (pro větší sílu testu) porovnejte rozdíly teplot před ošetřením a po něm.



Tabulka 15.6: Výsledky měření teploty t před ošetřením a po něm

$t [^{\circ}\text{C}]$	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
Před	36,2	36,4	35,9	37,1	38,2	39,1	37	38,4	36,8	37,1
Po	37,3	36,5	37,2	38,2	39,2	39,5	38,4	38,5	36,8	36,9

⁶Úloha včetně dat převzata z [4], s. 98.

15.5 Dvěma voltmetry byla současně naměřena vždy různá napětí se směrodatnou odchylkou $\sigma = 10 \text{ mV}$ (viz tabulku 15.7). Posuďte na 2% hladině významnosti hypotézu, že střední hodnoty údajů obou voltmetrů se sobě rovnají.



Tabulka 15.7: Výsledky měření dvěma voltmetry [mV]

1. měření	3,02	3,51	4	4,57	4,25	3,6	3	2,9	2,7
2. měření	3,005	3,65	4,002	4,512	3,91	3,04	3,2	3,3	2,62

15.6 Zjišťujeme, zda zvýšení teploty snižuje výkon dělníků v dílně. Pro tento účel bylo testováno patnáct dělníků a měřen jejich výkon za směnu (počet výrobků za směnu, tabulka 15.8). Na hladině významnosti $\alpha = 0,05$ otestujte, zda je výkon dělníků při odlišné teplotě stejný či ne.

Tabulka 15.8: Počty výrobků jednotlivých dělníků v závislosti na teplotě

Dělník č.	19 °C	23 °C	Dělník č.	19 °C	23 °C
1	52	50	9	67	61
2	58	51	10	68	57
3	59	60	11	70	64
4	65	51	12	63	64
5	71	58	13	51	50
6	55	52	14	56	51
7	59	60	15	62	53
8	63	59			

15.7 Při zkoušce ze statistiky bylo zjišťováno, zda studenti mají stejné výsledky v písemné a v ústní části zkoušky. Bylo testováno deset studentů, kteří dosáhli bodové zisky, jak uvádí tabulka 15.9.

Tabulka 15.9: Počty bodů jednotlivých studentů v písemné a ústní části zkoušky

Počet bodů/Student č.	1	2	3	4	5	6	7	8	9	10
Písemná část	28	16	11	24	26	30	14	16	26	35
Ústní část	25	10	0	30	28	25	8	12	27	30

15.8 Osmi rostlinám tabáku byl odebrán druhý list. Jedna náhodně vybraná polovina listu byla ošetřena přípravkem A, druhá přípravkem B. Potom byly listy potřeny suspenzí agresora a byl sledován počet skvrn na každé

polovině. Na hladině významnosti $\alpha = 0,05$ otestujte, zda se počty skvrn v tabulce 15.10 vzniklých přípravkem A a B významně liší.

Tabulka 15.10: Počty skvrn na listech

Počet skvrn / Rostlina č.	1	2	3	4	5	6	7	8
Přípravek A	9	17	31	18	7	8	20	10
Přípravek B	10	11	18	14	6	7	17	5

15.9 Zjistěte, zda se od sebe významně liší hmotnosti hlávek dvou různých odrůd zelí v kg (viz tabulku 15.11). Hypotézu testujte na hladině významnosti 0,05.

Tabulka 15.11: Výsledky vážení hlávek zelí odrůd z_1 a z_2 v kg

z_1	5,50	6,20	4,00	4,80	5,30	4,40	3,90	6,40	5,30
z_2	5,80	6,50	5,30	5,60	4,80	7,30	-	-	-

15.10 Byla porovnávána účinnost dvou protikoročních látek. První látka byla aplikována v 10 případech, druhá v 11 případech. Po stanovené době byl zjištěn stupeň poškození s těmito výsledky: $\bar{x}_1 = 82,4 \mu\text{m}$, $s_1^2 = 12,1 \mu\text{m}^2$, $\bar{x}_2 = 80,0 \mu\text{m}$, $s_2^2 = 10,5 \mu\text{m}^2$. Porovnejte průměrný účinek obou látek.

15.11 Bylo potřeba zjistit, zda příměs do oceli zvyšuje její pevnost. Výsledky měření pevnosti oceli bez příměsi a s příměsí v MPa ukazuje tabulka 15.12. Na hladině významnosti $\alpha = 0,05$ otestujte hypotézu, podle které nemá přidání speciální příměsi do oceli vliv na její pevnost.

Tabulka 15.12: Výsledky měření pevnosti oceli s příměsí a bez příměsi v MPa

Bez příměsi	5,6	5,9	5,9	5,7	5,8	5,7	6	5,5	5,7	5,5
S příměsí	6,2	5,7	5,6	6	6,3	5,8	5,7	6	6	5,8

15.12 Ve výrobně směsí betonu bylo vyrobeno a vyzkoušeno jedenáct výrobků ze dvou různých betonových směsí. Na hladině významnosti $\alpha = 0,05$ otestujte rovnost kvality pevnosti v MPa u obou směsí.

Tabulka 15.13: Výsledky měření pevnosti betonu s příměsí a bez příměsí v MPa

Směs 1	18	19	19	21	22	20	19	21	22	18	19
Směs 2	19	20	21	20	18	22	21	19	21	20	21

15.13 S využitím sebraných dat o pomerančových nektarech firem v tabulce 15.14 otestujte na hladině významnosti $\alpha = 0,05$, zda pomerančové nektary firmy A

- se průkazně liší v obsahu pomerančové šťávy od výrobků firmy B,
- obsahují průměrně méně pomerančové šťávy než výrobky firmy B.



Tabulka 15.14: Obsah pomerančového nektaru firem A a B

A	58,8	50,2	53,5	49,4	51,7	47,9	50,3	53,8
B	51,1	49,7	53,8	52,3	50,5	49,1	52,8	57,1

15.14 Důležitým ukazatelem jakosti odrůdy lnu je hmotnost vlákna v gramech. U dvou odrůd lnu bylo provedeno měření hmotnosti vlákna (tabulka 15.12). Otestujte na hladině významnosti $\alpha = 0,05$, zda existují významné rozdíly v hmotnosti vláken u obou odrůd lnu.

Tabulka 15.15: Výsledky měření pevnosti betonu s příměsí a bez příměsí v MPa

Odrůda I	47,5	57,7	47,1	38,6	45,2	49,8	43,4
	50,8	41,5	38,8	47,1	51,2	52,6	46,1
Odrůda II	49,2	44,1	38,1	40,9	46,4	57,1	36,8
	42,3	50,5	52,5	39,5	40,9	41,2	50,1

15.15 Dvacet šest stejně starých selat bylo rozděleno do dvou skupin. V první skupině jich bylo čtrnáct a byla krmena podle plánu A, ve druhé skupině jich bylo 12 a měla krmný plán B. Po šesti měsících byly zaznamenány jejich hmotnostní přírůstky (viz tabulku 15.16). Otestujte rovnost rozptylů obou výběrů. Testováním rozhodněte, zda se průměrné hmotnostní přírůstky selat v obou skupinách významně liší.

Tabulka 15.16: Přírůstky hmotnosti selat v kg

Plán A	62	54	55	60	53	58	60	57	63	70	58	64	61	55
Plán B	52	56	49	50	51	47	60	53	57	60	62	70		

Řešení úloh

15.1 Ze zadání je $n = 8$, $\bar{x} = 4,2$, $s_x = 1,3$. Testujeme $H_0: \mu = 3,5$ proti $H_1: \mu > 3,5$, neboť alternativou je, že obsah nikotinu přesahuje danou mez.

Jednovýběrový t -test o střední hodnotě: $t \doteq 1,4246$, $p = 0,01$, kritické hodnoty asi $\pm 2,9979$, rozhodnutí: nezamítáme H_0 o tom, že průměrná hodnota nikotinu v cigaretách nepřesahuje 3,5 mg. Výrobce tedy neporušuje tvrzení v reklamě.

15.2 Ze zadání je $n = 5$, $\mu = 0$. Vypočítáme $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = -0,8$ a $s^2 = \frac{1}{n-1} (\sum_{i=1}^n x_i^2 - n\bar{x}^2) \doteq 3,7$, tedy $s \doteq 1,9235$. Testujeme $H_0: \mu = 0$ proti $H_1: \mu \neq 0$. Jednovýběrový t -test o střední hodnotě: Hodnota testové statistiky $|t| \doteq 0,93$, $p = 0,01$, kritické hodnoty asi $\pm 3,747$, rozhodnutí: nezamítáme H_0 o tom, že průměrná odchylka při plnění balíčků je 0 mg. Automat tedy podle zjištěných dat nevykazuje systematickou chybu.

15.3 $\bar{x} \doteq 98,56$, $s_X^2 \doteq 3,435$, hodnota testové statistiky $t \doteq 5,722$, kritické hodnoty: 2,009, 2,678. Rozhodnutí o H_0 : zamítáme H_0 pro hladinu významnosti $\alpha = 0,05$ i $\alpha = 0,01$. Hmotnost dávky premixu se v 50 kg směsi významně liší od požadované normy.

15.4 Testujeme H_0 : ošetření nemá vliv na změnu teploty proti H_1 : ošetření má vliv na změnu teploty, tj. teplota se zvýší, nebo sníží. Párový oboustranný t -test na střední hodnotu: hodnota testové statistiky asi $-3,28$, kritická hodnota asi $-3,25$, H_0 se zamítá na hladině významnosti 0,01 (rovněž p -hodnota oboustranné varianty testu asi $0,0095 < 0,01$), závěr: ošetření má vliv při zvolené hladině významnosti 0,01 na teplotu člověka.

15.5 Párový oboustranný t -test na střední hodnotu: hodnota testové statistiky asi 0,364, kritická hodnota asi 2,896, H_0 se zamítá na hladině významnosti 0,02 (rovněž p -hodnota oboustranné varianty testu asi $0,725 > 0,02$), závěr: H_0 se nezamítá na hladině významnosti 0,02. Není průkazné, že se střední hodnoty měření liší, neboli oba voltmetry dávají stejné výsledky.

15.6 Párový t -test, $t \doteq 4,089$, H_0 zamítáme, rozdíly ve výkonu dělníků jsou statisticky významné.

15.7 Párový t -test, $t \doteq 2,017$, nezamítáme H_0 , že jsou výkony studentů u obou částí zkoušky vyrovnané.

15.8 Párový t -test, H_0 : není rozdíl v účinku přípravků A a B. proti H_1 : Je rozdíl v účinku přípravků A a B. $t \doteq 2,625$, kritická hodnota 2,365, zamítáme H_0 , tedy z výběru lze usuzovat na rozdíl v účinku přípravků A a B.

15.9 Dvouvýběrový (nepárový) t -test, testujeme H_0 : Rozdíl ve váze hlávek zelí je neprůkazný (tj. není průkazný). proti H_1 : Rozdíl ve váze hlávek zelí není neprůkazný (tj. je průkazný).

15.10 F -test na shodu rozptylů: $H_0: \sigma_1^2 = \sigma_2^2$, $H_1: \sigma_1^2 \neq \sigma_2^2$, $f \doteq 0,152 \not\geq F_{9,10}(0,975) \doteq 3,799$, $p \doteq 0,411$, na hladině $\alpha = 0,05$ nelze zamítnout H_0 , tedy rozptyly nejsou prokazatelně rozdílné;

dvouvýběrový t -test: $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$, $|t| = |1,637| \not\geq t_{19}(0,975) \doteq 2,093$, H_0 nelze zamítnout na hladině významnosti $\alpha = 0,05$, nebylo tedy prokázáno, že by se účinnost obou látek významně lišila, $p = \mathbb{P}(|T| > 1,637) \doteq 0,118$, tedy H_0 by bylo možné zamítnout až na hladině významnosti 11,6 % a vyšší.

15.11 F -test na shodu rozptylů: $H_0: \sigma_1^2 = \sigma_2^2$, $H_1: \sigma_1^2 \neq \sigma_2^2$, $\sigma_1^2 \doteq 0,029$, $\sigma_2^2 \doteq 0,052$, $f \doteq 0,556 \not\geq F_{9,9}(0,975) \doteq 4,025$, $p \doteq 0,478$, na hladině $\alpha = 0,05$ nelze zamítnout H_0 , rozptyly se rovnají, jinak také rozptyly nejsou prokazatelně rozdílné;

dvouvýběrový t -test: $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$, $\bar{x}_1 \doteq 5,73$, $\bar{x}_2 \doteq 5,91$, $|t| = |-1,998| \not\geq t_{18}(0,975) \doteq 2,1$, H_0 nelze zamítnout na hladině významnosti $\alpha = 0,05$, nebylo tedy prokázáno, že by se příměsí významně zvyšovala pevnost oceli, $p = \mathbb{P}(|T| > 1,998) \doteq 0,061$, tedy H_0 není možné zamítnout až do hladiny asi 6,1 % a vyšší.

15.12 F -test na shodu rozptylů: $H_0: \sigma_1^2 = \sigma_2^2$, $H_1: \sigma_1^2 \neq \sigma_2^2$, $\sigma_1^2 \doteq 2,164$, $\sigma_2^2 \doteq 1,364$, $f \doteq 1,587 \not\geq F_{10,10}(0,975) \doteq 3,717$, $p \doteq 0,478$, na hladině $\alpha = 0,05$ nelze zamítnout H_0 , rozptyly nejsou prokazatelně rozdílné;

dvouvýběrový t -test: $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$, $\bar{x}_1 \doteq 19,818$, $\bar{x}_2 \doteq 20,182$, $|t| = |-0,642| \not\geq t_{20}(0,975) \doteq 2,086$, H_0 nelze zamítnout na hladině významnosti $\alpha = 0,05$, nebylo tedy prokázáno, že by se pevnost, resp. kvalita obou směsí významně lišila, $p = \mathbb{P}(|T| > 0,642) \doteq 0,528$, tedy H_0 není možné zamítnout na jakékoliv smysluplné hladině významnosti (až do hladiny 52,8 % a vyšší).

15.13 F -test na shodu rozptylů: $H_0: \sigma_A^2 = \sigma_B^2$, $H_1: \sigma_A^2 \neq \sigma_B^2$, $n_A = n_B = 8$, $\sigma_A^2 \doteq 11,642$, $\sigma_B^2 \doteq 6,674$, $f \doteq 1,744 \not\geq F_{7,7}(0,975) \doteq 4,995$, $p \doteq 0,48$, na hladině významnosti $\alpha = 0,05$ nelze zamítnout H_0 , rozptyly nejsou prokazatelně rozdílné;

dvouvýběrový t -test: $H_0: \mu_A = \mu_B$, $H_1: \mu_A \neq \mu_B$, $\bar{x}_A \doteq 51,95$, $\bar{x}_B \doteq 52,05$, $|t| = |-0,066| \not\geq t_{14}(0,975) \doteq -2,145$, H_0 nelze zamítnout na hladině významnosti $\alpha = 0,05$, nebylo tedy prokázáno, že by se oba nektary pro-

kazatelně lišily v obsahu pomerančové šťávy, resp. kvalita obou směsí významně lišila, $p = \mathbb{P}(|T| > 0,066) \doteq 0,948$, tedy H_0 není možné zamítnout na jakékoliv smysluplné hladině významnosti (až do hladiny 94,8 % a vyšší).

15.14 Dvouvýběrový t -test bez rovnosti rozptylů obou výběrů, H_0 : Hmotnosti vláken lnu se u obou odrůd rovnají. proti H_1 : Hmotnosti vláken lnu se u obou odrůd liší., $t \doteq 0,92$, kritická hodnota 2,05, nezamítáme H_0 o rovnosti hmotností vláken lnu u obou odrůd lnu.

15.15 $s_X^2 = 21,14$, $s_Y^2 = 43,53$, $\bar{x} \doteq 59,285$, $\bar{y} \doteq 55,583$, $t \doteq 1,679$, nezamítáme H_0 o rovnosti hmotnostních přírůstků selat u diety A a B.

Kapitola 16

Neparametrické testy

Parametrické testy jsou založeny na několika předpokladech. Jedním z nich je předpoklad, že výběr pochází z daného rozdělení, které je známo až na některé parametry. Pokud je dané rozdělení normální, používáme parametrické testy z kapitoly 15, přičemž porušení normality při dostatečně velkém výběru nemění závěry testů – v tomto případě se totiž můžeme opřít o centrální limitní větu a zákony velkých čísel. Často se však setkáváme s výběry malých rozsahů, které pocházejí z výrazně „nenormálních“ základních souborů. V takových případech nelze parametrické testy z kapitoly 15 použít, protože testové statistiky nemají patřičná rozdělení (Studentovo, Pearsonovo či Fisherovo–Snedecorovo). Při práci s nimi pak využíváme tzv. neparametrické testy. Tyto testy mají velmi obecné předpoklady a jsou matematicky nenáročné. Cenou za obecnost je však jejich menší síla, kterou oproti parametrickým testům mají.

Před uvedením některých neparametrických testů zavedeme pojem pořadí. Mějme daná různá reálná čísla $x_1, x_2, \dots, x_n$. Pořadím R_i čísla x_i nazýváme počet těch čísel $x_1, x_2, \dots, x_n$, která jsou menší nebo rovna číslu x_i . Mějme například čísla 5, 8, 9, 3, 2, 1. Číslo 5 má pořadí 4, protože čísla 5, 3, 2, 1 jsou menší nebo rovna 5. Příslušná pořadí jsou shrnuta v tabulce 16.1.

Může se stát, že čísla $x_1, x_2, \dots, x_n$ nejsou různá, tzn. některá z nich jsou si rovna a vytvářejí tzv. shody. V tomto případě se pak číslům, která tvoří shodu, přiřazuje průměrné pořadí odpovídající takové skupince. Například

Tabulka 16.1: Tabulka vybraných čísel s přiřazeným pořadím

Čísla x_i	5	8	9	3	2	1
Pořadí R_i	4	5	6	3	2	1

čísłům 5, 5, 5, 9, 9, 2, 1 se přiřadí pořadí $R_1, R_2, \dots, R_7$, uvedené v tabulce 16.2.

Tabulka 16.2: Tabulka čísel vytvářejících shody v pořadí

Očíslování hodnot x_i	1	2	3	4	5	6	7
Vzestupně uspořádané hodnoty x_i	1	2	5	5	5	9	9
Pořadí R_i	1	2	4	4	4	6,5	6,5

Pro určování pořadí dat v souboru o libovolném rozsahu lze v MS Excel využít funkci `= RANK.AVG`, která vrátí pořadí dané hodnoty v číselném seznamu. První argument je číslo, kterému chceme udělit pořadí, druhý argument je oblast dat, v rámci kterých pořadí hledáme, a třetí argument nabízí volbu 1 nebo 0, odpovídající řazení vzestupně nebo sestupně. Pokud má více hodnot stejné pořadí, bude vráceno průměrné pořadí.

16.1 Znaménkový test

Nechť $X_1, X_2, \dots, X_n$ je náhodný výběr z rozdělení se spojitou distribuční funkcí a $\tilde{\mu}$ je medián tohoto rozdělení, pak platí

$$\mathbb{P}(X < \tilde{\mu}) = \mathbb{P}(X > \tilde{\mu}) = \frac{1}{2}.$$

Chceme testovat hypotézu

$$H_0: \tilde{\mu} = \tilde{\mu}_0 \quad \text{proti alternativě} \quad H_1: \tilde{\mu} \neq \tilde{\mu}_0,$$

kde $\tilde{\mu}_0$ je dané číslo (nejčastěji rovno nule). Utvoříme nejprve rozdíly

$$X_1 - \tilde{\mu}_0, X_2 - \tilde{\mu}_0, \dots, X_n - \tilde{\mu}_0.$$

Náhodná veličina Y pak bude označovat počet těch rozdílů, které mají kladné znaménko. Za předpokladu platnosti hypotézy H_0 má náhodná veličina Y binomické rozdělení s parametry n a $\frac{1}{2}$. Znaménkový test se proto také označuje jako znaménkový binomický test. Při oboustranném testu tvoří kritický obor jednak příliš malé hodnoty Y (tj. hodnoty ležící blízko nule), jednak příliš velké hodnoty Y (tj. hodnoty blízko n). V případě malého rozsahu výběru (tj. pro malá n) jsou vypočteny kritické hodnoty testu k_1, k_2 tak, že

$$\mathbb{P}(Y \leq k_1) \leq \frac{\alpha}{2}, \quad \mathbb{P}(Y \geq k_2) \leq \frac{\alpha}{2}.$$

Tabulka 16.3: Množství nadojeného mléka [l] a výpočet pořadí diferencí pro příklad 16.1

Kráva č.	1	2	3	4	5	6	7	8	9	10
1. laktace	14	12	15	12	12	14	11	12	15	10
2. laktace	15	16	19	15	14	11	11	15	14	13
Diference	-1	-4	-4	-3	-2	3	0	-3	1	-3
Znaménko difference	-	-	-	-	-	+		-	+	-

Kritické hodnoty k_1, k_2 je možné nalézt v tabulkách; pro $\alpha = 0,05$ a pro $\alpha = 0,01$ je uvádíme v tabulce 21.29 na straně 445. Hypotézu H_0 tedy zamítáme, jestliže zjistíme, že $Y \leq k_1$ nebo $Y \geq k_2$. Při velkém rozsahu náhodného výběru (stačí $n \geq 20$) vypočítáme statistiku

$$Z = \frac{2Y - n}{\sqrt{n}}.$$

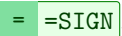
Veličina Z má za platnosti H_0 podle CLV asymptoticky normální rozdělení $N(0; 1)$, tudíž hypotézu H_0 zamítneme, jestliže

$$|z| \geq u\left(1 - \frac{\alpha}{2}\right).$$

Znaménkový test je možné provést též jako párový test. Na rozdíl od párového t -testu nemusíme k provedení znaménkového testu znát přesné hodnoty $X_i, Y_i, i = 1, 2, \dots, n$, ale stačí vědět, zda je rozdíl $X_i - Y_i$ kladný nebo záporný. Z tohoto důvodu je znaménkový test použitelný i v případech, kdy jsou k dispozici pouze kvalitativní srovnání, například lék A působí lépe než lék B. U znaménkového testu se může stát, že některé rozdíly budou rovny nule. Například u kvalitativního srovnání není subjekt schopen rozhodnout o vlivu tím či oním směrem. V tomto případě se doporučuje nulové hodnoty vynechat a za n vzít jen počet nenulových hodnot. Nejčastěji se znaménkový test používá k orientačnímu hodnocení předběžných pokusů.

Příklad 16.1.  U deseti krav téhož plemene bylo zjišťováno v první a druhé laktaci množství mleziva třetí den po porodu (viz první tři řádky tabulky 16.3). Zjistěte, zda existuje statisticky významný rozdíl mezi naměřenými daty.

Řešení. Poslední řádek tabulky 16.3 obsahuje znaménka diferencí¹; nulovou diferencii u 7. krávy vynecháváme, tudíž $n = 9$. Počet kladných znamének

¹V Excelu je možné je také zjistit prostřednictvím funkce  `=SIGN`.

(hodnota testové statistiky Y) je $y = 2$, dolní kritická hodnota pro $\alpha = 0,05$ a $n = 9$ je $k_1 = 1$ (viz tabulku 21.29 na straně 445). Horní kritická hodnota má díky symetrii hodnotu $k_2 = n - k_1 = 8$, proto se pro zjednodušení v tabulce kritických hodnot neuvádí. Protože $2 \not\leq 1$ i $2 \not\geq 8$, nulovou hypotézu H_0 nezamítáme, tj. rozdíl v množství nadojeného mléka mezi oběma laktacemi není tímto testem statisticky prokázán. $\triangle$

16.2 Jednovýběrový Wilcoxonův test

Jednovýběrový Wilcoxonův² test je neparametrickou obdobou jednovýběrového Studentova t -testu a párového Studentova t -testu. Jeho provedení je o něco náročnější než provedení znaménkového testu, zato je však citlivější.

Při užití Wilcoxonova testu jako neparametrické verze jednovýběrového t -testu předpokládejme, že $X_1, X_2, \dots, X_n$ je náhodný výběr ze spojitého rozdělení s distribuční funkcí $F(x)$ a hustotou symetrickou kolem nuly, tj.

$$F(x) = 1 - F(-x), \quad -\infty < x < \infty.$$

V tomto případě je nula mediánem daného rozdělení. Testujeme hypotézu

$$H_0: \tilde{\mu} = 0 \quad \text{proti alternativě} \quad H_1: \tilde{\mu} \neq 0,$$

případně proti jednostranným alternativám

$$H_1: \tilde{\mu} < 0, \quad \text{resp.} \quad H_1: \tilde{\mu} > 0.$$

Seřadme $X_1, X_2, \dots, X_n$ do rostoucí posloupnosti podle velikosti jejich absolutní hodnoty, tj.

$$|X|^{(1)} < |X|^{(2)} < \dots < |X|^{(n)}.$$

Při tomto uspořádání označíme R_i^+ pořadí $|X_i|$ a zavedeme veličiny

$$S^+ = \sum_{X_i \geq 0} R_i^+, \quad S^- = \sum_{X_i < 0} R_i^+$$

vyjadřující součet pořadí nezáporných hodnot X_i , resp. záporných hodnot. Pokud jsme určili veličiny S^+ a S^- správně, musí platit $S^+ + S^- = \frac{n(n+1)}{2}$, neboť sčítáme čísla od 1 do n .

Pro vyhodnocení platnosti nulové hypotézy proti oboustranné alternativě použijeme statistiku $W = \min(S^+, S^-)$. Pokud je tato statistika menší nebo rovna tabelované kritické hodnotě, nulovou hypotézu zamítneme.

²Frank Wilcoxon (1892–1965), irský a americký chemik a statistik.

Tabulka 16.4: Výsledky měření únavy kovů v tisících kmitů do okamžiku lomu

X_i	$Z_i = X_i - 8\,000$	R_i	a_i
3 322	-4 678	2	0
14 713	6 713	5	1
763	-7 237	6	0
46 296	38 296	8	1
2 845	-5 155	3	0
9 411	1 411	1	1
1 532	-6 468	4	0
24 023	16 023	7	1

V případě levostranné alternativy je testovací statistika $W = S^+$, v případě pravostranné alternativy je testovací statistika $W = S^-$. Kritické hodnoty jsou uvedeny v tabulkách. Pro $n \geq 30$ použijeme testovou statistiku Z , která bude mít asymptoticky normální rozdělení $N(0; 1)$ a tvar

$$Z = \frac{S^+ - \frac{1}{4}n(n+1)}{\sqrt{24n(n+1)(2n+1)}}.$$

V případě, že

$$|z| \geq u\left(1 - \frac{\alpha}{2}\right),$$

zamítneme hypotézu na hladině, která je asymptoticky rovna α . Dodejme ještě závěrem, že Wilcoxonův test je založen na předpokladu symetrie; jestliže je však rozdělení výrazně asymetrické, může test zamítnout nulovou hypotézu i v případě, kdy je medián roven nule.

Příklad 16.2. Při měření únavy kovů bylo u osmi zkoušek při napětí 560 MPa dosaženo následujících hodnot počtu kmitů do okamžiku lomu (v tisících): 3 322, 14 713, 763, 46 296, 2 845, 9 411, 1 532, 24 023. Na hladině významnosti $\alpha = 0,05$ testujte hypotézu $H_0: \tilde{\mu} = 8 \cdot 10^6$ proti $H_1: \tilde{\mu} > 8 \cdot 10^6$.

Řešení. Jednovýběrový Wilcoxonův test byl výše navržen pro test nulového mediánu. Takový má za platnosti nulové hypotézy veličina $Z_i = X_i - \tilde{\mu}$, jejíž hodnoty a také pořadí R_i jsou uvedeny v tabulce 16.4. Hodnoty a_i přiřazují rozdílu $X_i - \tilde{\mu}$ hodnotu 1, resp. 0, jde-li o číslo kladné, resp. záporné. Testujeme proti pravostranné alternativě $H_1: \tilde{\mu} - 8 \cdot 10^6 > 0$, v jejíž prospěch budou svědčit nízké hodnoty S^- . Tudíž v pravostranném testu použijeme

Tabulka 16.5: Výsledky délky řešení kontrolních úloh v příkladu 16.3

Před cvičením (x_i)	87	61	98	90	93	74	83	72	81	75	83
Po cvičení (y_i)	50	45	79	90	88	65	52	79	84	61	52
Rozdíly	37	16	19	0	5	9	31	-7	-3	14	31
Abs. hodnoty	37	16	19	0	5	9	31	7	3	14	31
Pořadí	11	7	8	1	3	5	9,5	4	2	6	9,5

jako testovou statistiku S^- , kterou porovnáme s kritickou hodnotou jednostranného testu. Z tabulky 16.4 se vypočítá hodnota $s^- = 2 + 6 + 3 + 4 = 15$, testová statistika pro pravostrannou verzi testu je $w = s^- = 15$. Kritická hodnota $w_8(0,05) = 5$. Protože $15 \not\leq 5$, nezamítáme H_0 o počtu kmitů. $\triangle$

Při užití Wilcoxonova jednovýběrového testu jako neparametrické verze párového t -testu předpokládáme, že $(X_1, Y_1), \dots, (X_n, Y_n)$ je náhodný výběr ze spojitého dvourozměrného rozdělení. Testujeme, že mediány obou náhodných veličin jsou si rovny proti oboustranné alternativě, resp. jednostranným alternativám: $H_0: \tilde{\mu}_X - \tilde{\mu}_Y = 0$ proti $H_1: \tilde{\mu}_X - \tilde{\mu}_Y \neq 0$, resp. $H_1: \tilde{\mu}_X - \tilde{\mu}_Y < 0$, resp. $H_1: \tilde{\mu}_X - \tilde{\mu}_Y > 0$. Určíme rozdíly $Z_i = X_i - Y_i, i = 1, 2, \dots, n$ a poté postupujeme stejně jako u neparametrické verze pro jednovýběrový t -test.

Příklad 16.3.  Speciální cvičení na pamětné počítání bylo testováno na jedenácti žácích. V tabulce 16.5 jsou uvedeny časy v sekundách, za které žáci vyřešili kontrolní úlohy před cvičením a po cvičení. Můžeme tvrdit, že tato cvičení zlepšují schopnost žáků při řešení úloh na hladině významnosti $\alpha = 0,05$?

Řešení. Testujeme hypotézu $H_0: \tilde{\mu}_X - \tilde{\mu}_Y = 0$ (cvičení nemá vliv na schopnost řešení úloh) proti $H_1: \tilde{\mu}_X - \tilde{\mu}_Y > 0$ (cvičení zlepšuje schopnost žáků při řešení úloh). Rozdíly spočtené v tabulce 16.5 nejsou normálně rozdělené (čtenář si nakreslí histogram, případně ověří normalitu pomocí testu z odstavce 20.2), tudíž použijeme neparametrický Wilcoxonův test. Spočteme statistiky

$$s^+ = 1 + 3 + 5 + 6 + 7 + 8 + 9 + 10 + 11 = 60, \quad s^- = 2 + 4 = 6,$$

$$w = \min(s^+, s^-) = 6$$

Kritická hodnota jednovýběrového Wilcoxonova testu pravostranné varianty je $w_{11}(0,05) = 13$ (viz tabulku kritických hodnot na straně 447), z čehož plyne, že zamítáme hypotézu H_0 na hladině 0,05.

Tato úloha se dá řešit i znaménkovým testem, avšak při něm nevyužijeme všech informací, které známe, a to může vést k mylným závěrům. Počet kladných hodnot je $y = 9$. Kritické hodnoty znaménkového testu jsou $k_1 = 1$, $k_2 = 10$. Z $k_1 < y < k_2$ plyne, že nezamítáme hypotézu H_0 . V tomto případě dostaneme rozdílné výsledky obou testů. Znaménkový test nemá dostatek informací pro zamítnutí hypotézy H_0 , protože využívá pouze počtu záporných hodnot, zatímco u Wilcoxonova testu využijeme navíc znalosti toho, že záporné hodnoty jsou poměrně malé. Řekneme, že Wilcoxonův test je silnější než znaménkový test. $\triangle$

16.3 Dvouvýběrový Wilcoxonův test

Tento test se používá nejčastěji místo dvouvýběrového t -testu. Opět dochází k zobecnění předpokladu, který je kladen na distribuční funkce daných náhodných výběrů. Nechť $X_1, X_2, \dots, X_m$ a $Y_1, Y_2, \dots, Y_n$ jsou dva nezávislé výběry ze dvou spojitých rozdělení. Chceme testovat hypotézu, že distribuční funkce obou rozdělení jsou totožné. Oba výběry $X_1, X_2, \dots, X_m$, $Y_1, Y_2, \dots, Y_n$ uspořádáme společně, jako jeden výběr, vzestupně podle velikosti. Zjistíme součet pořadí hodnot $X_1, X_2, \dots, X_m$ ve spojených výběrech. Součet označíme T_1 . Dále zjistíme součet pořadí hodnot $Y_1, Y_2, \dots, Y_n$ a označíme ho T_2 . Vypočteme

$$U_1 = mn + \frac{m(m+1)}{2} - T_1, \quad U_2 = mn + \frac{n(n+1)}{2} - T_2.$$

Vzhledem k tomu, že $T_1 + T_2 = \frac{(m+n+1)(m+n)}{2}$, platí $U_1 + U_2 = mn$. Pokud $\min(u_1, u_2)$ je menší nebo rovno kritické hodnotě uvedené v tabulce, zamítneme hypotézu, že distribuční funkce obou rozdělení jsou stejné. Při velkém rozsahu obou výběrů opět přejdeme ke statistice, která má za platnosti hypotézy asymptoticky normální rozdělení $N(0; 1)$ a tvar

$$Z = \frac{U_1 - \frac{mn}{2}}{\sqrt{\frac{mn}{12}(m+n+1)}}.$$

Pokud $|z| \geq u(1 - \frac{\alpha}{2})$, zamítneme hypotézu na hladině asymptoticky rovné α .

Příklad 16.4.  V tabulce 16.6 jsou uvedeny délky larev chrobáků žijících v osivech zimní rýže a prosa.³ Porovnejte rozdělení těchto délek na hladině významnosti $\alpha = 0,05$.

³Úloha včetně dat převzata ze sbírky [17], s. 185.

Tabulka 16.6: Délky larev v rýži a prosu a pořadí délek pro příklad 16.4

Délka larvy v rýži	7	10	14	15	12	16	12
Délka larvy v prosu	11	12	16	13	18	15	-
Pořadí délek larev v rýži	1	2	8	9,5	5	11,5	5
Pořadí délek larev v prosu	3	5	11,5	7	13	9,5	-

Řešení. Součet pořadí v rýži je roven $t_1 = 42$ a v prosu je roven $t_2 = 49$. $n_1 = 7$ a $n_2 = 6$. Spočteme $u_1 = 42 + \frac{56}{2} - 42 = 28$, $u_2 = 42 + \frac{42}{2} - 49 = 14$. Kritická hodnota dvouvýběrového Wilcoxonova testu na hladině $\alpha = 0,05$ pro rozsahy výběrů 7 a 6 je rovna $w_{7,6}(0,05) = 6$ (viz tabulku na straně 448). Nezamítáme hypotézy o shodnosti rozdělení obou výběrů, protože $\min(u_1, u_2) \not\leq w_{7,6}(0,05)$. $\triangle$



16.4 Úlohy

16.1 Prověřte účinnost nové tréninkové metody na základě rozdílů maximálních výsledků skokanského družstva před zavedením nové metody a po třech měsících jejího používání. Rozdíly výkonů jednotlivých skokanů ukazuje tabulka 16.7.

Tabulka 16.7: Rozdíly výkonů v cm

Skokan č.	1	2	3	4	5	6	7	8
Rozdíl výkonu	+5	-10	+15	-1	+8	+21	+7	-6
Skokan č.	9	10	11	12	13	14	15	16
Rozdíl výkonu	+4	+12	-3	+7	+9	-2	+3	+7

16.2 U dvaceti náhodně vybraných zásilek byly sebrány údaje o počtu dnů mezi dodáním objednané zásilky do místa bydliště a jejím vyzvednutím zákazníkem: 1, 2, 3, 0, 5, 7, 10, 4, 5, 0, 7, 8, 5, 4, 1, 12, 3, 5, 10, 1. Na hladině významnosti 5 % otestujte nulovou hypotézu, že medián doby převzetí zásilky je 5 dní.

16.3 Šestnácti pacientům byly odebrány vzorky moči. Na první polovinu každého odběru byl aplikován Furantoin (F) a na druhou Penicilin (P). Po 24 hodinách byl sledován počet bakterií v 1. a 2. polovině každého odběru. Zjistilo se, že z deseti případů byl počet bakterií menší u látky F než u látky P (tj. $F < P$), u čtyř případů byl počet bakterií větší u látky F než u P (tj. $F > P$) a u dvou případů nebylo možno rozhodnout (tj. $F = P$). Zjistěte, zda se Furantoin a Penicilin liší ve své účinnosti. Znaménkovým testem otestujte hypotézu o stejném vlivu látky F a P.

16.4 U osmnácti zákazníků restaurace byla změřena doba čekání na objednání (d_o) a doba čekání na příchod vrchního (d_z), aby mu mohl zákazník zaplatit. Výsledky měření v minutách ukazuje tabulka 16.8. Na hladině významnosti 0,05 otestujte hypotézu, že jsou obě doby čekání stejné.

Tabulka 16.8: Čekací doby v minutách na objednání a na zaplacení v restauraci

d_o	7, 0, 5, 2, 4, 5, 3, 7, 9, 3, 1, 4, 3, 5, 5, 4, 6, 3
d_z	5, 4, 6, 3, 8, 5, 2, 6, 3, 7, 2, 8, 2, 7, 1, 8, 2, 2

16.5 Od osmi pacientů byly odebrány vzorky moči a rozděleny každý na polovinu. První polovina byla ošetřena látkou A, druhá látkou B. Poté byly zjišťovány rozdíly v počtu kolonií bakterií v obou řadách A a B. Výsledky ukazuje tabulka 16.9. Zhodnoťte výsledky testu vzorků po úpravě látkami A a B.

Tabulka 16.9: Počty kolonií bakterií ve vzorcích u látky A a B

A	2, 0, 4, 5, 2, 1, 2, 3
B	1, 2, 6, 3, 5, 3, 0, 4

16.6 Na 5% hladině významnosti porovnejte dva způsoby měření kompresního tlaku (v 10^5 Pa) ve spalovacím prostoru traktorového motoru. U šesti motorů bylo provedeno měření kompresního tlaku oběma způsoby a byly stanoveny difference získaných výsledků (viz tabulku 16.10).

Tabulka 16.10: Výsledky měření kompresního tlaku v 10^5 Pa

Motor č.	1	2	3	4	5	6
Diference d_i	-0,1	-0,2	0,2	0,1	-0,2	-0,1

16.7 Testujte rovnost rozdělení dvou nezávislých výběrů V_1 a V_2 (viz tabulku 16.11) představujících bodové výsledky žáků v testu z matematiky.

Tabulka 16.11: Bodové výsledky žáků v testu z matematiky

V_1	17, 15, 13, 2, 5, 11, 30, 24, 19, 23, 21, 24, 29, 25, 15, 22, 6, 8
V_2	5, 14, 26, 30, 18, 15, 2, 26, 29, 17, 12, 8, 12, 7, 21, 4, 12, 22

16.8 S použitím neparametrického testu ověřte, zda dva rozdílné způsoby posklizňového dosoušení mají vliv na klíčivost sladovnického ječmene. U každého způsobu bylo odebráno deset vzorků se stovkou semen s výsledky uvedenými v tabulce 16.12.⁴

Tabulka 16.12: Počty vyklíčených semen při sušení A a sušení B

Vzorek č.	1	2	3	4	5	6	7	8	9	10
Sušení A	89	88	92	94	90	81	93	90	87	83
Sušení B	88	77	83	76	79	82	80	72	78	72

16.9 Na čtyřech kombajnech se zkoušel nový způsob setí, zbývajících šest kombajnů selo standardním způsobem. Po sklizni se sledovala výnosnost polí v metrických centech na 1 ha, jak ukazuje tabulka 16.13. Zjistěte, zda nový

⁴Úloha včetně dat převzata z [19], s. 59.

způsob setí obilí má vliv na průměrné hektarové výnosy polí než původní způsob setí.

Tabulka 16.13: Výnosnost polí v metrických centech na 1 ha

Nový způsob setí	51	52	49	55		
Původní způsob setí	45	54	48	44	53	50

Řešení úloh

16.1 Znaménkový test, H_0 : Nová metoda výcviku je stejně účinná jako původní. proti H_1 : Nová metoda výcviku je účinnější než původní. $m_+ = 11$, $m_- = 5$, $n = 16$, testové kritérium: $m = \min(11; 5) = 5$, kritická hodnota pro $\alpha = 0,05$ a $n = 16$ párů je 4, $5 > 4$, nezamítáme H_0 .

16.2 $m_+ = 6$, $m_- = 10$, $n = 16$, testové kritérium: $m = \min(6; 10) = 6$, kritická hodnota pro $\alpha = 0,05$ a $n = 16$ párů je 3, $m > 3$, tedy $6 > 3 \Rightarrow$ nezamítáme hypotézu o hodnotě mediánu 5, neboli nebyl prokázán statisticky významný rozdíl ($p < 0,05$) mezi hodnotami a mediánem.

16.3 $m_+ = 10$, $m_- = 4$, $n = 14$, testové kritérium: $m = \min(10; 4) = 4$, kritická hodnota pro $\alpha = 0,05$ a $n = 14$ párů je 2, $m > 2$, tedy $4 > 2 \Rightarrow$ nezamítáme hypotézu o shodném působení obou látek, neboli nebyl prokázán statisticky významný rozdíl ($p < 0,05$) v účinnosti Furantoinu a Penicilínu na růst bakterií ve vzorcích moči pacientů.

16.4 $m_+ = 8$, $m_- = 9$, $n = 17$, testové kritérium: $m = \min(8, 9) = 8$, kritická hodnota pro $\alpha = 0,05$ a $n = 17$ párů je 4, $m > 4$, $8 > 4$, tedy nezamítáme hypotézu o rovnosti dob čekání, neboli nebyl prokázán statisticky významný rozdíl ($p < 0,05$) mezi oběma čekacími dobami.

16.5 $n = 8$, žádné nulové rozdíly, součet pořadí odpovídajících kladným rozdílům $w^+ = 11,5$, součet pořadí odpovídajících záporným rozdílům $w^- = 24,5$, testové kritérium je $w = \min(11,5; 24,5) = 11,5$, kritická hodnota $w_8(0,05) = 3$, $11,5 > 3$, tedy nezamítáme nulovou hypotézu o rovnosti účinku látek A a B, neboli nebyl prokázán významný rozdíl vlivu látek A a B na odebraný vzorek.

16.6 Výsledkům měření přiřadíme pořadí dle tabulky 16.14 a použijeme Wilcoxonův jednovýběrový test.

Tabulka 16.14: Výsledky měření kompresního tlaku v 10^5 Pa

Motor č.	1	2	3	4	5	6
Diference d_i	-0,1	-0,2	0,2	0,1	-0,2	-0,1
Pořadí d_i^+			5	2		
Pořadí d_i^-	2	5			5	2

$w^+ = 5 + 2 = 7$, $w^- = 2 + 5 + 5 + 2 = 14$, kritická hodnota Wilcoxonova testu pro $\alpha = 0,05$ a $n = 7$ je 2, testové kritérium $w = \min(7, 14) = 7$, $7 > 2$, nezamítáme nulovou hypotézu, že mezi oběma metodami měření kompresního tlaku není významný rozdíl.

16.7 Wilcoxonův test pro dva nezávislé výběry: $n_1 = n_2 = 18$, $T_1 = 352$, $T_2 = 314$, $U_1 = 143$, $U_2 = 181$, testová statistika $W = \min(143; 181) = 143$, kritická hodnota asi 200, $143 < 200 \Rightarrow$ zamítáme H_0 o shodnosti rozdělení (distribučních funkcí) obou výběrů, tedy výsledky žáků mají jiné rozdělení a nelze je tedy považovat za sobě rovná.

16.8 Wilcoxonův dvouvýběrový test: všem hodnotám určíme pořadí (viz tabulku 16.15), sečteme pořadí v první a druhé skupině a vypočteme hodnotu testového kritéria.

Tabulka 16.15: Počty vyklíčených semen při sušení A a sušení B

Vzorek č.	1	2	3	4	5	6	7	8	9	10
Sušení A	89	88	92	94	90	81	93	90	87	83
Pořadí x	15	13,5	18	20	16,5	8	19	16,5	12	10,5
Sušení B	88	77	83	76	79	82	80	72	78	72
Pořadí y	13,5	4	10,5	3	6	9	7	2	5	1

Součet pořadí $T_x = 149$, $T_y = 61$, $U_x = 6$, $U_y = 94$, testové kritérium pro $\alpha = 0,05$, $n_1 = n_2 = 10$ je 23, $6 < 23$, zamítáme nulovou hypotézu, tj. existuje významný rozdíl v klíčivosti semen dle způsobu dosoušení.

16.9 Všem seřazeným hodnotám přiřadíme pořadí (viz tabulku 16.16), $T_1 = 27$, $T_2 = 28$, $U_1 = 7$, $U_2 = 17$, $m = 4$, $n = 6$, $\alpha = 0,05$, $\min(7, 17) = 7$, $7 > 2$, nulovou hypotézu nezamítáme, tedy nový způsob setí nemá větší vliv na průměrné hektarové výnosy než původní způsob setí.

Tabulka 16.16: Výnosnost polí v metrických centech na 1 ha

Počet semen	44	45	48	49	50	51	52	53	54	55
Pořadí	1	2	3	4	5	6	7	8	9	10

Kapitola 17

Porovnání více výběrů

17.1 Analýza rozptylu jednoduchého třídění

Tento test je zobecněním dvouvýběrového t -testu, který rozšíříme na případ ($I \geq 3$) výběrů. Uvažujme tedy I nezávislých výběrů,

$Y_{11}, \dots, Y_{1n_1}$ je výběr z $N(\mu_1; \sigma^2)$,

$Y_{21}, \dots, Y_{2n_2}$ je výběr z $N(\mu_2; \sigma^2)$,

atd. až

$Y_{I1}, \dots, Y_{In_I}$ je výběr z $N(\mu_I; \sigma^2)$.

Chceme testovat hypotézu

$$H_0: \mu_1 = \dots = \mu_I$$

proti alternativě

H_1 : Existují alespoň dvě střední hodnoty, které si nejsou rovny.

Někdy se uvedená situace zapisuje modelem:

$$Y_{ij} = \mu + \alpha_i + e_{ij},$$

kde $\mu + \alpha_i = \mu_i$ a $e_{ij} \sim N(0; \sigma^2)$ je chyba vyplývající z nepřesnosti měření nebo ze systematické odchylky od průměru. Hypotézu H_0 přepíšeme na jednodušší model, který je splněn, pokud platí hypotéza H_0 :

$$Y_{ij} = \mu + e_{ij}.$$

Test provedeme následovně. Nejprve si označme průměry jednotlivých výběrů

$$\bar{Y}_i = \frac{Y_{i1} + \cdots + Y_{in_i}}{n_i},$$

pro $i = 1, \dots, I$ a průměr všech hodnot

$$\bar{Y} = \frac{\sum_i \sum_j Y_{ij}}{n},$$

kde $n = n_1 + \cdots + n_I$. Nyní spočtěme celkový součet čtverců S_T (tj. celková kvadratická chyba modelu za platnosti H_0 , tedy v případě že $\mu_1 = \mu_2 = \cdots = \mu_I = \mu$). Za odhad μ se bere $\bar{Y}$.

$$S_T = \sum_i \sum_j (Y_{ij} - \bar{Y})^2 = \sum_i \sum_j Y_{ij}^2 - n\bar{Y}^2.$$

Reziduální součet čtverců S_e je celková kvadratická chyba modelu za předpokladu, že hypotéza H_0 neplatí, tedy v případě že $\mu_1 \neq \cdots \neq \mu_I$. Za odhad μ_i se bere $\bar{Y}_i$.

$$S_e = \sum_i \sum_j (Y_{ij} - \bar{Y}_i)^2 = \sum_i \sum_j Y_{ij}^2 - \sum_i n_i \bar{Y}_i^2.$$

Veličina $S_A = S_T - S_e$ se interpretuje jako součet čtverců připadající na rozdíly v ošetřeních. Tato veličina je vždy kladná, protože chyba obecnějšího modelu S_e je vždy menší než chyba jednoduššího modelu S_T . Je-li S_A malé, pak jsou si oba modely podobné, a tudíž nebudeme zamítat hypotézu H_0 . Je-li S_A velké, pak obecnější model vysvětluje velkou část celkové chyby S_T a tudíž zamítneme H_0 .

Za platnosti hypotézy H_0 má statistika

$$F_A = \frac{S_A/(I-1)}{S_e/(n-I)} \sim F_{I-1, n-I}$$

F rozdělení o $I-1$ a $n-I$ stupních volnosti. Tedy hypotézu H_0 zamítneme na hladině α v případě, že

$$f_A \geq F_{I-1, n-I}(1-\alpha).$$

Výsledky celého testu se stručně zapisují do tabulky (viz tabulku 17.1).

Tabulka 17.1: Tabulka výsledků analýzy rozptylu

Variabilita	Součet čtverců S	Počet stupňů volnosti df	Podíl S/df	F
Ošetření	S_A	$df_A = I - 1$	S_A/df_A	F_A
Reziduální	S_e	$df_e = n - I$	S_e/df_e	-
Celková	S_T	$df_T = n - 1$	-	-

Předpoklady tohoto testu jsou obdobné předpokladům dvouvýběrového t -testu (viz stranu 297). Nejdůležitější je opět nezávislost jednotlivých výběrů, normalita může být porušena, pokud rozsahy výběrů umožňují použití CLV. Není-li tomu tak, je vhodnější provést neparametrickou obdobu tohoto testu, která se nazývá Kruskalův-Wallisův test. Posledním předpokladem je rovnost rozptylů všech výběrů. V případě zamítnutí hypotézy o shodě rozptylů všech výběrů (homoskedasticity) můžeme použít Kruskalův-Wallisův test nebo obecnější váženou podobu testu ANOVA, který nám umožní modelovat nestejné rozptyly. K testování shody rozptylů můžeme použít například Bartlettův test nebo Leveneův test (viz např. [2]).

Veličina $S^2 = \frac{S_e}{(n-I)}$ se nazývá reziduální rozptyl a je nestranným odhadem rozptylu σ^2 .

Mohlo by se zdát, že výše uvedený test by se dal provádět sadou dvouvýběrových t -testů, provedených na každou dvojici výběrů. Ovšem takových testů bychom museli udělat $\frac{I(I-1)}{2}$. Kdyby každý z nich byl proveden na hladině α , byla by výsledná hladina výrazně větší než α . Pokud bychom hladinu každého testu snížili na $\frac{2\alpha}{I(I-1)}$, byla by celková hladina naopak podstatně menší než α . Ukazuje se, že tento postup nevede k dobrým výsledkům.

V případě, že hypotézu H_0 zamítneme, je často třeba rozhodnout, pro které dvojice indexů platí $\mu_i \neq \mu_j$. Tento problém řeší Tukeyova¹ metoda mnohonásobného porovnání.

Protože $\bar{Y}_i$ je odhadem pro μ_i , vytvoří se nejprve tabulka rozdílů $\bar{Y}_i - \bar{Y}_j$ (viz tabulku 17.2).

¹John Wilder Tukey (1915–2000), americký matematik a statistik.

Tabulka 17.2: Rozdíly průměrů

i	j			
	2	3	...	I
1	$\bar{Y}_1 - \bar{Y}_2$	$\bar{Y}_1 - \bar{Y}_3$	...	$\bar{Y}_1 - \bar{Y}_I$
2		$\bar{Y}_2 - \bar{Y}_3$	...	$\bar{Y}_2 - \bar{Y}_I$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$I - 1$				$\bar{Y}_{I-1} - \bar{Y}_I$

Statistika

$$\frac{|\bar{Y}_i - \bar{Y}_j|}{S \sqrt{\frac{1}{2} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}} \sim q_{I, n-I}$$

má rozdělení, které se nazývá studentizované rozpětí. Kritická hodnota $q_{I, f}(\alpha)$ studentizovaného rozpětí je takové číslo, pro něž platí

$$\mathbb{P}[Q \geq q_{I, f}(\alpha)] = \alpha.$$

Tyto kritické hodnoty jsou tabelovány. Tudíž platí-li

$$|\bar{y}_i - \bar{y}_j| \geq sq_{I, n-I}(\alpha) \sqrt{\frac{1}{2} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)},$$

zamítáme hypotézu o rovnosti $\mu_i = \mu_j$. Provedeme-li tento postup pro všechny dvojice, pak hladina testu je menší nebo rovna α . Rovnost nastává v případě, že všechny výběry mají stejný rozsah.

Abychom se lépe orientovali ve výsledcích Tukeyovy metody, připisuje se do tabulky 17.2 ke každému rozdílu hvězdička, pokud je rozdíl významný (signifikantní) na hladině 0,05. Dvě hvězdičky pro významnost na hladině 0,01 a tři pro 0,001.

17.1.1 Levenův test

Tzv. Levenův² test se používá k posouzení rovnosti rozptylů pro proměnnou vypočítanou pro dva nebo více výběry. Z předpokladu nezávislosti jednotlivých výběrů testujeme hypotézu

$$H_0: \sigma_1^2 = \sigma_2^2 = \dots$$

²Howard Levene (1914–2003), americký statistik a genetik.

Tabulka 17.3: Tabulka s výnosy pšenice na pěti polích

Rok	Číslo pole/Výnosy v tunách				
	1	2	3	4	5
2024	105	98	95	101	105
2023	89	121	100	102	104
2022	120	110	99	98	99
2021	111	98	98	105	104
2020	100	130	100	95	92
2019	98	110	120	110	100
2018	97	120	104	94	92
2017	125		90	112	98
2016	103		99		100
2015	90				102

<i>F</i>	<i>Hodnota P</i>	<i>F krit</i>
2.281754	0.077820589	2.612306

Obrázek 17.1: Výřez z MS Excel části ANOVA v testu Anova: jeden faktor nástroje Analýza dat – pro Levenův test

proti alternativě

H_1 : Rozptyly aspoň jedné dvojice náhodných výběrů se nerovnají.

K výchozímu souboru dat Y_{ij} nalezneme $Z_{ij} = |Y_{ij} - \bar{Y}_i|$, tedy absolutní hodnoty rozdílu dané hodnoty a průměru výběru, do kterého daná hodnota patří, kde $\bar{Y}_i$ je aritmetický průměr hodnot z daného i -tého výběru. Na takto vytvořenou tabulku s novými daty aplikujeme jednofaktorovou verzi testu ANOVA.

K testování shodnosti rozptylů lze využít také Bartlettův test. Podobu testových statistik pro oba testy lze nalézt například v [2].

Příklad 17.1.  Tabulka 17.3 obsahuje data o výnosech pšenice v tunách na pěti polích za posledních 10 let. Ne všichni zemědělci však poskytli všechna data potřebná data. Ověřte na hladině významnosti 0,05, zda mezi výnosy existují významné rozdíly.

Řešení. Užitím Leveneova testu nejdříve testujeme rovnost rozptylů všech výběrů. Vypočítáme průměry výnosů jednotlivých polí za všechny roky.

Tabulka 17.4: Tabulka pro Leveneův test (data jsou absolutní hodnoty rozdílů zaokrouhlená na tři desetinná místa)

Rok	Číslo pole/Výnosy v tunách				
	1	2	3	4	5
2024	1,200	14,429	5,556	1,125	5,400
2023	14,800	8,571	0,556	0,125	4,400
2022	16,200	2,429	1,556	4,125	0,600
2021	7,200	14,429	2,556	2,875	4,400
2020	3,800	17,571	0,556	7,125	7,600
2019	5,800	2,429	19,444	7,875	0,400
2018	6,800	7,571	3,444	8,125	7,600
2017	21,2		10,556	9,875	1,600
2016	0,800		1,556		0,400
2015	13,800				2,400

<i>F</i>	<i>Hodnota P</i>	<i>F krit</i>
2.459707	0.061311426	2.612305612

Obrázek 17.2: Výřez z MS Excel části ANOVA u testu Anova: jeden faktor nástroje Analýza dat – pro test shody středních hodnot výnosů

Vytvoříme novou tabulku, která bude obsahovat absolutní hodnoty rozdílů původních hodnot od průměru daného výběru (viz tabulku 17.4). Na data z tabulky 17.4 aplikujeme test Anova: jeden faktor z nástroje Analýza dat v MS Excel. Z části ANOVA zjistíme hodnotu testové statistiky W , kritickou hodnotu pro zvolenou hladinu významnosti α a p -hodnotu testu (viz obrázek 17.1). Protože testová statistika $F \doteq 2,282$ je menší než kritická hodnota asi 2,61, nezamítáme H_0 o rovnosti rozptylů výnosů pšenice na všech polích. Také $p \doteq 0,078$ ukazuje na nezamítnutí H_0 .

Při splnění rovnosti rozptylů jednotlivých výběrů můžeme testovat rovnost průměrů jednotlivých výběrů, a to opět pomocí testu Anova: jeden faktor z nástroje Analýza dat v MS Excel. V tomto případě však již využijeme zadaná data. Z testování průměrů vyplývá, že nezamítáme H_0 o rovnosti průměrných výnosů na vybraných polích (viz obrázek 17.2). $\triangle$

Příklad 17.2. Sleduje se účinnost tří protikoročních látek. První byla použita ve 20 případech, druhá ve 25 případech a třetí ve 22 případech. Po stanovené době byl zjištěn stupeň poškození s těmito výsledky: $\bar{y}_1 = 82,4$,

$s_1^2 = 12$, $\bar{y}_2 = 80$, $s_2^2 = 10$, $\bar{y}_3 = 85,8$, $s_3^2 = 12$. Bohužel konkrétní měření se nedochovala. Vypočítejte statistickou rozdílnost účinnosti na hladině významnosti 0,01.

Řešení. Postupně vypočteme: $n = 67$, $\bar{y} = \frac{n_1\bar{y}_1+n_2\bar{y}_2+n_3\bar{y}_3}{n} \doteq 82,64$. Podle definice výběrového rozptylu $s_i^2 = \frac{1}{n-1} \sum_j y_{ij}^2 - \frac{n}{n-1} \bar{y}_i^2$ vypočteme $\sum_j y_{1j}^2 = 136\,023$, $\sum_j y_{2j}^2 = 160\,240$, $\sum_j y_{3j}^2 = 162\,208$. Odtud již můžeme sestavit tabulku analýzy rozptylu:

Tabulka 17.5: Tabulka analýzy rozptylu

Variabilita	S	df	S/df	F
Mezi látkami	183,5	2	91,75	8,15**
Reziduální	720	64	11,25	-
Celková	903,5	66	-	-

Kritická hodnota $F_{2,64}(0,01) \doteq 4,98$, tudíž zamítáme hypotézu o rovnosti účinnosti všech protikorozních látek.

Nyní je třeba odhalit, které rovnosti jsou porušeny. Vytvořme tabulku rozdílů průměrů (viz tabulku 17.6).

Tabulka 17.6: Tabulka rozdílů průměrů

Rozdíly	2. látka	3. látka
1. látka	2,4	-3,4**
2. látka		-5,8**

V tabulkách nalezneme kritickou hodnotu studentizovaného rozpětí: $q_{3,64}(0,01) \doteq 4,28$ a vypočteme hodnoty zamítnutí pro každou dvojici zvlášť (viz tabulku 17.7).

Tabulka 17.7: Hodnoty zamítnutí pro jednotlivé dvojice

Hodnoty zamítnutí	2. látka	3. látka
1. látka	2,93	3,01
2. látka		2,85

Vidíme, že byla prokázána rozdílnost třetí protikorozní látky s ostatními na hladině významnosti 0,01, naproti tomu nebyla prokázána rozdílnost první a druhé látky na hladině 0,01. $\triangle$

17.2 Kruskalův–Wallisův test

Kruskalův³–Wallisův⁴ test je neparametrickou obdobou analýzy rozptylu jednoduchého třídění a je zobecněním Wilcoxonova dvouvýběrového testu, který rozšíříme na případ I výběrů ($I \geq 3$). Uvažujme tedy I nezávislých výběrů, které jsou postupně rozsahu $n_1, n_2, \dots, n_I$. Označme $n = n_1 + n_2 + \dots + n_I$. Předpokládejme, že každý tento výběr pochází z nějakého rozdělení se spojitou distribuční funkcí, tedy nelze se spoléhat na fakt, že jsou všechny výběry z normálního rozdělení. Chceme testovat hypotézu, že všechny výběry pocházejí z téhož rozdělení. Tento test je citlivý zejména na vzájemné posunutí jednotlivých rozdělení.

Podobně jako u Wilcoxonova dvouvýběrového testu seřadíme všech n prvků z I výběrů do rostoucí posloupnosti a určíme pořadí každého prvku. Označme T_i součet pořadí těch prvků, které patří do i -tého výběru ($i = 1, 2, \dots, I$). Vzhledem k tomu, že celkový počet prvků ze všech výběrů je n , musí platit $T_1 + T_2 + \dots + T_I = \frac{n(n+1)}{2}$. Tohoto vztahu můžeme využít ke kontrole správnosti výpočtu T_i . V případě platnosti hypotézy má pak statistika

$$Q = \frac{12}{n(n+1)} \sum_{i=1}^I \frac{T_i^2}{n_i} - 3 \cdot (n+1)$$

pro $n \rightarrow \infty$ asymptoticky χ^2 rozdělení o $I-1$ stupních volnosti. Jestliže $q \geq \chi_{I-1}^2(1-\alpha)$, zamítneme hypotézu na hladině, která je asymptoticky rovna α . Pokud hypotézu zamítneme, tvrdíme, že všechny výběry nepocházejí z téhož rozdělení. V tomto případě nás pak zajímá, které výběry se od sebe vzájemně liší.

V případě, že rozsahy všech výběrů jsou stejné, použijeme za tímto účelem Neményiovu metodu mnohonásobného srovnávání. Je-li číslo $|T_i - T_j|$ větší nebo rovno kritické hodnotě (kritické hodnoty, s nimiž tato metoda pracuje, jsou tabelovány), zamítá se hypotéza, že i -tý a j -tý výběr pocházejí z téhož rozdělení. Tento postup se aplikuje na všech $\frac{I(I-1)}{2}$ čísel $|T_i - T_j|$.

V případě, že rozsahy všech výběrů nejsou stejné, označíme $t_i = \frac{T_i}{n_i}$, $i = 1, \dots, I$, a prohlásíme, že se distribuční funkce i -tého a j -tého výběru

³William Henry Kruskal (1919–2005), americký matematik a statistik.

⁴Wilson Allen Wallis (1912–1998), americký ekonom a statistik.

Tabulka 17.8: Výsledky chodu přístrojů bez poruch (v hodinách)

A	48	16	75	29	96	67	89	22
B	46	94	20	87	66	25	14	75
C	58	17	65	34	26	63	74	106

Tabulka 17.9: Uspořádaná data ze všech výběrů

	Pořadí v celém souboru dat								Celkem
A	11	2	18,5	8	23	16	21	5	104,5
B	10	22	4	20	15	6	1	18,5	96,5
C	12	3	14	9	7	13	17	24	99

od sebe významně liší, pokud

$$|t_i - t_j| > \sqrt{\frac{1}{12} \left(\frac{1}{n_i} + \frac{1}{n_j} \right) n(n+1) \chi_{I-1}^2(1-\alpha)}.$$

Příklad 17.3.  Byla sledována doba (v hodinách) bezporuchového chodu osmi přístrojů tří různých značek A, B a C. Výsledky jsou shrnuty v tabulce 17.8. Porovnejte kvalitu všech tří značek ve smyslu doby bezporuchového chodu.

Řešení. Lze předpokládat, že jednotlivé výběry jsou z exponenciálního rozdělení, tudíž pro porovnání kvality provedení přístrojů u jednotlivých značek nemůžeme použít analýzu rozptylu. Proto použijeme Kruskalův–Wallisův test. Data uspořádáme v jednom společném výběru, výsledky jsou znázorněny v tabulce 17.9. Vypočteme $q = 0,083\,75$. Vypočítáme kritickou hodnotu $\chi_2^2(0,05) \doteq 5,99$.

 `= CHISQ.INV(0,95;2)`

Protože $q < \chi_2^2(0,95)$, nezamítáme hypotézu H_0 . Data neprokázala významný rozdíl mezi kvalitami provedení všech značek. △

17.3 Analýza rozptylu dvojného třídění

Předpokládejme, že zkoumáme vliv dvou faktorů A a B na závislou proměnnou veličinu Y . Například zjišťujeme, zda úroveň finanční gramotnosti měřená body v didaktickém testu (náhodná veličina Y) je ovlivněna typem

Tabulka 17.10: Výsledky didaktického testu z finanční gramotnosti

B	Faktor A				
	1	2	3	...	J
1	$Y_{111}, \dots, Y_{11c}$	$Y_{121}, \dots, Y_{12c}$	$Y_{131}, \dots, Y_{13c}$	...	$Y_{11I}, \dots, Y_{1Jc}$
2	$Y_{211}, \dots, Y_{21c}$	$Y_{221}, \dots, Y_{22c}$	$Y_{231}, \dots, Y_{23c}$	...	$Y_{211}, \dots, Y_{2Jc}$
...	...	...	...	...	...
I	$Y_{I11}, \dots, Y_{I1c}$	$Y_{I21}, \dots, Y_{I2c}$	$Y_{I31}, \dots, Y_{I3c}$	...	$Y_{IJ1}, \dots, Y_{IJc}$

výukové metody (faktor A) a vzděláním rodičů (faktor B). Předpokládáme, že faktor A má J úrovní (tj. počet použitých typů výukových metod) a faktor B má I úrovní (tj. nejvyšší dosažené vzdělání alespoň jednoho z rodičů: základní, vyučen, střední s maturitou, bakalářské, magisterské, ...). Přitom máme n_{ij} měření takových, že při použití i -tého typu výukové metody a při j -tém dosažení nejvyššího vzdělání rodičů žák dosáhl určitého počtu bodů. Výsledky (tzn. počet bodů z didaktického testu) těchto n_{ij} pokusů označíme $Y_{ij1}, Y_{ij2}, \dots, Y_{ijn_{ij}}$. Pokud bychom se omezili pouze na případy, kdy počet pozorování $n_{ij} = c \geq 1$ (jde o tzv. vyvážené třídění), pak obdržíme tabulku 17.10.

V této publikaci se zaměříme pouze na dvojné třídění bez opakování, tzn. že pro každou podtřídu budeme mít jedno měření (v každé buňce v tabulce 17.10 bude pouze jedno číslo). Dále budeme předpokládat, že sloupcový faktor A a řádkový faktor B se neovlivňují (např. to znamená, že použitá výuková metoda působí na výsledek žáka bez ohledu na vzdělání rodičů, faktory se skládají aditivně). Tomuto modelu se říká model dvojného třídění bez interakce a opakování.

Uvažujme model:

$$Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \text{ kde } i = 1, \dots, I, j = 1, \dots, J, \quad (17.1)$$

kde μ , α_i pro $i = 1, \dots, I$ a β_j pro $j = 1, \dots, J$ jsou neznámé parametry a $e_{ij} \sim (0; \sigma^2)$ je chyba vyplývající z nepřesnosti měření nebo ze systematické odchylky od průměru. To znamená, že naměřené veličiny Y_{ij} závisí jak na sloupci, tak na řádku, ve kterém se vyskytují. Navíc v každém řádku máme stejný počet prvků. Představme si například situaci, kdy J pacientům měříme tlak v I okamžicích, například ráno, v poledne a večer. Každý pacient má jinou průměrnou hodnotu tlaku $\mu + \beta_j$. Výchyly během dne jsou určeny parametry α_i .

Parametry modelu (17.1) nejsou určeny jednoznačně. Konkrétně je možné všechny hodnoty α_i posunout o nějakou hodnotu $\tilde{\alpha} \in \mathbb{R}$ a podobně

hodnoty β_j o hodnotu $\tilde{\beta} \in \mathbb{R}$ a součet $\tilde{\alpha} + \tilde{\beta}$ naopak odečíst od hodnoty μ , přesto půjde stále o tentýž model. Abychom tomuto předešli, klademe na parametry α_i a β_j dvě dodatečné podmínky:

$$\sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0.$$

Nyní chceme testovat hypotézu $H_0: \alpha_1 = \dots = \alpha_I = 0$ (tj. že nezáleží na řádkovém třídění), kterou přepíšeme na jednodušší model odpovídající jednoduchému třídění:

$$Y_{ij} = \mu + \beta_j + e_{ij}.$$

Test provedeme následovně. Nejprve si označme průměry jednotlivých výběrů

$$\begin{aligned} \bar{Y}_{i.} &= \frac{Y_{i1} + \dots + Y_{iJ}}{J} \text{ pro } i = 1, \dots, I, \\ \bar{Y}_{.j} &= \frac{Y_{1j} + \dots + Y_{Ij}}{I} \text{ pro } j = 1, \dots, J \end{aligned}$$

a průměr všech hodnot

$$\bar{Y} = \frac{\sum_i \sum_j Y_{ij}}{n},$$

kde $n = IJ$. Nyní spočteme celkový součet čtverců S_T , tj. celková kvadratická chyba v případě, že $\alpha_1 = \dots = \alpha_I = \beta_1 = \dots = \beta_J = 0$. Za odhad μ se bere $\bar{Y}$.

$$S_T = \sum_i \sum_j (Y_{ij} - \bar{Y})^2 = \sum_i \sum_j Y_{ij}^2 - n\bar{Y}^2.$$

Součet čtverců chyb v řádcích označíme

$$S_A = J \sum_i \bar{Y}_{i.}^2 - n\bar{Y}^2.$$

Součet čtverců chyb ve sloupcích označíme

$$S_B = I \sum_j \bar{Y}_{.j}^2 - n\bar{Y}^2.$$

Reziduální součet čtverců $S_e = S_T - S_A - S_B$ je celková kvadratická chyba modelu 17.1.

Stupně volnosti jednotlivých součtů čtverců jsou:

$$df_T = n - 1, \quad df_A = I - 1, \quad df_B = J - 1, \quad df_e = n - I - J + 1.$$

Hypotézu H_0 zamítneme na hladině α v případě, že

$$F_A = \frac{S_A/df_A}{S_e/df_e} \geq F_{df_A, df_e}(1 - \alpha).$$

Podobně postupujeme při testování hypotézy $H'_0: \beta_1 = \dots = \beta_J = 0$ (tj. že nezáleží na sloupcovém třídění), kterou přepíšeme na jednodušší model odpovídající jednoduchému třídění:

$$Y_{ij} = \mu + \alpha_i + e_{ij}.$$

Hypotézu H'_0 zamítneme na hladině α v případě, že

$$F_B = \frac{S_B/df_B}{S_e/df_e} \geq F_{df_B, df_e}(1 - \alpha).$$

Výsledky celého testu se stručně zapíší do tabulky (viz tabulku 17.11).

Tabulka 17.11: Tabulka analýzy rozptylu dvojného třídění

Variabilita	Součet čtverců S	Počet stupňů volnosti df	Podíl S/df	F
Řádková	S_A	$df_A = I - 1$	S_A/df_A	F_A
Sloupcová	S_B	$df_B = J - 1$	S_B/df_B	F_B
Reziduální	S_e	$df_e = n - I - J + 1$	S_e/df_e	-
Celková	S_T	$df_T = n - 1$	-	-

Reziduální rozptyl $S^2 = \frac{S_e}{df_e}$ je nestranným odhadem rozptylu σ^2 .

V případě, že hypotézu H_0 zamítneme, je často třeba rozhodnout, pro které dvojice indexů neplatí rovnost. Tento problém řeší, stejně jako u jednoduchého třídění, Tukeyova metoda mnohonásobného porovnání.

Rovnost $\alpha_i = \alpha_l$ zamítneme, platí-li

$$|\bar{Y}_{i.} - \bar{Y}_{l.}| \geq sq_{I, n-I-J+1}(\alpha) \sqrt{\frac{1}{J}}.$$

Rovnost $\beta_j = \beta_l$ zamítneme, platí-li

$$|\bar{y}_{.j} - \bar{y}_{.l}| \geq sq_{J, n-I-J+1}(\alpha) \sqrt{\frac{1}{I}}.$$

Jak již bylo v úvodu kapitoly řečeno, model (17.1) se dá dále zobecňovat. Například pro každé i, j můžeme mít C dat.

$$Y_{ijc} = \mu + \alpha_i + \beta_j + e_{ijc},$$

kde $i = 1, \dots, I, j = 1, \dots, J, c = 1, \dots, C$.

Pokud by se faktory působící na proměnnou vzájemně ovlivňovaly, pak použijeme model (17.2), ve kterém je zahrnuto sledování interakcí mezi řádky a sloupci, jde o model dvojného třídění s interakcemi:

$$Y_{ijc} = \mu + \alpha_i + \beta_j + \lambda_{ij} + e_{ijc}. \quad (17.2)$$

Dále je také možné sledovat závislost veličin na třech typech parametrů – tzv. trojném třídění. Tyto modely jsou řešeny například v [1], [2]. Pokud provádíme analýzu dvojného třídění s jedním pozorováním v každé podtřídě, jde o tzv. Friedmanův test, tedy neparametrickou obdobou výše popsaného dvojného třídění.

Příklad 17.4.  Byl sledován vliv tří preparátů na srážlivost krve. Kromě jiných ukazatelů byl zjišťován tzv. trombinový čas. U každé osoby byl stanoven nejprve kontrolní údaj K , který udává trombinovaný čas před zahájením pokusu. Pak byly aplikovány preparáty A, B, C , a to každý dostatečně dlouho po odeznění účinku těch předchozích. Údaje o 10 sledovaných osobách jsou uvedeny v tabulce 17.12.⁵

Tabulka 17.12: Výsledky měření trombinového času

Osoba	Preparát				$\bar{Y}_i$
	K	A	B	C	
1	11,3	11,2	11,4	11,0	11,225
2	11,9	12,1	11,8	9,5	11,325
3	11,8	13,2	12,0	11,1	12,025
4	12,1	12,8	12,0	12,5	12,35
5	11,2	13,5	11,5	8,4	11,15
6	11,3	12,5	11,5	9,0	11,075
7	10,8	10,7	10,9	9,7	10,525
8	12,0	13,8	11,6	12,2	12,4
9	11,5	12,9	11,3	10,3	11,5
10	11,7	11,9	11,3	8,2	10,775
$\bar{Y}_{.j}$	11,56	12,46	11,53	10,19	11,435

⁵Úloha včetně dat v tabulce 17.12 převzata z [1], s. 233.

Rozhodněte, zda mají všechny testované osoby po aplikaci preparátů A, B, C na hladině významnosti 0,05 (0,01) stejný vliv na trombinový čas.

Řešení. Odtud již můžeme vyjádřit tabulku analýzy rozptylu (viz tabulku 17.13).

Tabulka 17.13: Tabulka analýzy rozptylu

Variabilita	S	df	S/df	F
Řádková (osoby)	14,606	9	1,62	2,58*
Sloupcová (preparáty)	26,253	3	8,75	13,9**
Reziduální	16,992	27	0,63	–
Celková	57,851	39	–	–

Vypočítáme kritickou hodnotu $F_{9,27}(0,05) \doteq 2,25$.

= F.INV(0,95;9;27)

Tudíž zamítáme hypotézu, že všechny osoby mají stejný trombinový čas na hladině 0,05.

Vypočítáme kritickou hodnotu $F_{3,27}(0,01) \doteq 4,6$. Tudíž zamítáme hypotézu, že preparáty nemají vliv na trombinový čas na hladině 0,01.

V MS Excel lze při porovnání průměrů více výběrů postupovat pomocí numerických výpočtů podle vzorců (viz dostupný soubor MS Excel) nebo k testování zvolit Anova: dva faktory bez opakování nástroje Analýza dat. V tabulce ANOVA se v přední části zobrazují stejné informace, jako jsou v tabulce 17.13 (viz obrázek 17.3).

ANOVA				
Zdroj variability	SS	Rozdíl	MS	F
Řádky	14.606	9	1.622888889	2.578742938
Sloupce	26.253	3	8.751	13.90519068
Chyba	16.992	27	0.629333333	
Celkem	57.851	39		

Obrázek 17.3: Výřez z MS Excel pro dvoufaktorovou ANOVA nástroje Analýza dat

Ve stejné tabulce se v její pravé části nacházejí informace o hodnotách testových F -statistik, o kritických hodnotách a o p -hodnotách zvlášť pro řádky a sloupce (viz obrázek 17.4).

<i>F</i>	<i>Hodnota P</i>	<i>F krit</i>
2.578742938	0.027464187	2.250131477
13.90519068	1.13131E-05	2.960351318

Obrázek 17.4: Výřez z MS Excel pro dvoufaktorovou ANOVA nástroje Analýza dat

Hypotéza o shodnosti trombinového času v závislosti na osobách nás zajímat nebude. Její neplatnost jsme předpokládali již na začátku, proto jsme také zvolili dvouvýběrové třídění. Nyní je třeba odhalit, které rovnosti mezi preparáty jsou porušeny. Vytvořme tabulku rozdílů průměrů (viz tabulku 17.14).

Tabulka 17.14: Rozdíly průměrů trombinového času

Rozdíly	<i>A</i>	<i>B</i>	<i>C</i>
K	−0,9	0,03	1,37**
A		0,93	2,27**
B			1,34**

V tabulkách nalezneme kritickou hodnotu studentizovaného rozpětí: $q_{4,27}(0,01) \doteq 4,85$, $q_{4,27}(0,05) \doteq 3,875$ a vypočteme hodnoty zamítnutí: $s \cdot q_{4,27}(0,01) \sqrt{\frac{1}{10}} \doteq 0,79$, $s \cdot q_{4,27}(0,01) \sqrt{\frac{1}{10}} = 1,22$, $s \cdot q_{4,27}(0,05) \sqrt{\frac{1}{10}} \doteq 0,97$.

Vidíme, že preparát *C* významně snižuje trombinový čas jak ve vztahu k počátečnímu stavu *K*, tak k preparátům *A* a *B*. $\triangle$

17.4 Friedmanův test

Friedmanův⁶ test je neparametrickou obdobou analýzy rozptylu dvojného třídění. Máme IJ nezávislých pozorování, které uspořádáme do tabulky (viz tabulku 17.15).

⁶Milton Friedman (1912–2006), americký ekonom a statistik.

Tabulka 17.15: Hodnoty pozorovaných veličin

Sloupce	1	2	3	...	J
Řádky					
1	Y_{11}	Y_{12}	Y_{13}	...	Y_{1J}
2	Y_{21}	Y_{22}	Y_{23}	...	Y_{2J}
.	...	...	...	...	
.	...	...	...	...	
.	...	...	...	...	
I	Y_{I1}	Y_{I2}	Y_{I3}	...	Y_{IJ}

Náhodné veličiny Y_{ij} mají spojitou distribuční funkci F_{ij} , $i = 1, 2, \dots, I$, $j = 1, 2, \dots, J$ a jsou vzájemně nezávislé.

Budeme testovat hypotézu

$$H_0: F_{i1} = F_{i2} = \dots = F_{iJ}.$$

Neboli testujeme hypotézu, zda F_{ij} závisí na sloupcovém indexu j , přičemž předpokládáme, že mohou záviset na řádkovém indexu i .

Uspořádáme pozorování v každém řádku podle velikosti a označíme příslušná pořadí $R_{i1}, R_{i2}, \dots, R_{iJ}$, $i = 1, 2, \dots, I$ (viz tabulku 17.16).

Tabulka 17.16: Pořadí a součty pozorovaných veličin

Řádky	Sloupce					Řádkové součty
	1	2	3	...	J	
1	R_{11}	R_{12}	R_{13}	...	R_{1J}	$\frac{J(J+1)}{2}$
2	R_{21}	R_{22}	R_{23}	...	R_{2J}	$\frac{J(J+1)}{2}$
.	...	...	...	...		
.	...	...	...	...		
.	...	...	...	...		
I	R_{I1}	R_{I2}	...		R_{IJ}	$\frac{J(J+1)}{2}$
Sloupcové součty	$R_{.1}$	$R_{.2}$	...		$R_{.J}$	$R_{..} = \frac{IJ(J+1)}{2}$

Zde $R_{.j} = \sum_i R_{ij}$, $R_{..} = \sum_i \sum_j R_{ij}$.

Statistika Friedmanova testu je dána vzorcem

$$Q = \frac{12}{IJ(J+1)} \sum_{j=1}^J R_{.j}^2 - 3I(J+1). \quad (17.3)$$

Hypotézu H_0 zamítneme, jestliže Q překročí kritickou hodnotu uvedenou v tabulkách. Při větších hodnotách I se za kritickou hodnotu bere $\chi_{J-1}^2(1-\alpha)$.

Zamítneme-li hypotézu, zajímá nás, které sloupce se od sebe liší. Za tímto účelem vytvoříme tabulku hodnot $|R_{.j} - R_{.m}|$ pro všechna $j < m$.

Je-li některá z hodnot $|R_{.j} - R_{.m}|$ větší nebo rovna kritické hodnotě, která je tabelovaná, zamítne se na odpovídající hladině významnosti hypotéza, že $F_{ij} = F_{im}$.

Příklad 17.5. Příklad 17.4 řešte Friedmanovým testem.

Řešení. Vytvoříme pořadí R_{ij} (viz tabulku 17.17).

Tabulka 17.17: Pořadí

Osoba	Preparát			
	K	A	B	C
1	3	2	4	1
2	3	4	2	1
3	2	4	3	1
4	2	4	1	3
5	2	4	3	1
6	2	4	3	1
7	3	2	4	1
8	2	4	1	3
9	3	4	2	1
10	3	4	2	1
Součet	25	36	25	14

Podle vzorce 17.3 vypočteme

$$q = \frac{12}{10 \cdot 4 \cdot 5} (25^2 + 36^2 + 25^2 + 14^2) - 3 \cdot 10 \cdot 5 = 14,52^{**}.$$

V tabulkách nalezneme, že kritická hodnota na hladině 0,01 pro Friedmanův test je rovna 10,53. Protože $q > 10,53$, zamítáme hypotézu, že trombinový čas nezávisí na preparátech K , A , B , C na hladině významnosti 0,01.

Abychom zjistili, které z preparátů K , A , B , C se od sebe liší, vypočteme hodnoty $|R_{.j} - R_{.m}|$ (viz tabulku 17.18).

Tabulka 17.18: Absolutní hodnoty z rozdílů

	A	B	C
K	11	0	11
A		11	22**
B			11

Kritická hodnota pro mnohonásobné porovnávání u Friedmanova testu na hladině 0,01 je 18,0, tudíž významný se ukazuje pouze rozdíl mezi preparáty A a C .

Porovnáme-li výsledky s příkladem 17.4, vidíme, že Friedmanův test je slabší než analýza rozptylu dvojného třídění. Na druhou stranu Friedmanův test můžeme použít i v případě nenormálních rozdělení nebo v případě, kdy známe pouze pořadí výsledků. $\triangle$

17.5 Úlohy

17.1 Porovnejte úspěšnost pěti náhodně vybraných absolventů gymnázií (G), středních průmyslových škol (SPŠ) a odborných učilišť (OU) u přijímací zkoušky (viz tabulku 17.19). Testujte na hladině významnosti $\alpha = 0,05$. Pokud existují rozdíly mezi středními hodnotami, určete, mezi kterými soubory tomu tak je.

Tabulka 17.19: Bodová úspěšnost studentů různých škol u přijímací zkoušky

Student	G	SPŠ	OU
s_1	55	54	47
s_2	54	50	53
s_3	58	51	49
s_4	61	51	50
s_5	52	49	46

17.2 Zjistěte, zda je mezi některými ze čtyř prasnic A až D průkazný rozdíl v plodnosti. Pokud ano, zjistěte, mezi kterými významný rozdíl opravdu je. Tabulka 17.20 ukazuje počty narozených selat u jednotlivých prasnic.

Tabulka 17.20: Počty narozených selat prasnicím A – D v letech 2000–2009

Rok / Prasnice	A	B	C	D
2000	10	8	18	10
2001	12	3	13	12
2002	8	6	10	10
2003	13	7	12	8
2004	7	8	14	10
2005	10	7	12	5
2006	11	7	12	4
2007	8	6	10	-
2008	8	-	11	-
2009	-	-	14	-

17.3 Na náhodně vybraných pozemcích šesti zemědělců byly v průběhu let měřeny výnosy obilí (v t/ha) (viz tabulku 17.21). Zjistěte, mezi kterými zemědělci jsou ve výnosu obilí významné rozdíly.

Tabulka 17.21: Výnosy obilí jednotlivých zemědělců z let 2000–2009

Rok / Zemědělec	A	B	C	D	E	F
2000	4	8	6	7	5	4
2001	6	3	2	6	7	5
2002	8	6	3	8	6	4
2003	5	7	5	8	5	5
2004	7	8	6	4	5	5
2005	3	7	7	5	7	2
2006	2	7	1	4	2	3
2007	5	6	2		6	1
2008	6				5	2
2009					4	

17.4 Bylo sledováno, zda čtyři různé odrůdy brambor (A, B, C, D) mají stejné výnosnosti (v tunách). Každá odrůda byla pěstována na celkem pěti pozemcích. Zjištěné výnosnosti jsou uvedeny v tabulce 17.22.

Tabulka 17.22: Výnosnost odrůd brambor na pěti různých polích (v tunách)

Výnosnost odrůdy v tunách / Pole č.	1	2	3	4	5
A	22,1	24,0	23,6	20,8	21,5
B	25,9	28,0	26,1	26,7	25,4
C	26,6	27,2	26,1	26,9	27,1
D	28,3	27,9	29,1	27,2	28,6

Analýzou rozptylu jednoduchého třídění zjistíte, zda existují ve výnosech obilí významné rozdíly.

17.5 Skupina neurotických dětí byla rozdělena do skupin A, B, C, D podle příslušných sledovaných syndromů. K objektivnímu doplnění psychiatrické diagnózy byla provedena analýza jejich krve, která sledovala určité fyziologické hodnoty jejího složení. Předpokládejme, že analýza přinesla výsledky uvedené v tabulce 17.23.

Tabulka 17.23: Tabulka skupin

Pacient č.	Skupina A	Skupina B	Skupina C	Skupina D
1	13	13	12	10
2	12	14	10	11
3	11	15	9	12
4	15	14	9	-
5	14	-	10	-
$\bar{x}_i$	13	14	10	11

Zjistěte, zda se sledované hodnoty v příslušných skupinách výrazně liší tak, aby uvedená analýza mohla být využita k diagnostikování sledovaných syndromů.

17.6 V dílně byla zjišťována délka praxe a počet výrobků vyrobených jednotlivými dělníky během jednoho týdne. Výsledky jsou shrnuty v tabulce 17.24. Na hladině významnosti 5% a 1% ověřte, zda má délka praxe statisticky významný vliv na produktivitu práce. Záznamy o produktivitě práce pro různou délku praxe tvoří tři výběry ze tří základních souborů s rozděleními $N(\mu_1; \sigma^2)$, $N(\mu_2, \sigma^2)$ a $N(\mu_3, \sigma^2)$.

Tabulka 17.24: Tabulka produktivity

Praxe	Produktivita	n_i	Součet x_i	Průměr $\bar{x}_i$	$\sum_j x_{ij}^2$
< 10 let	76, 67, 80, 70	4	293	73,25	21 565
10–25 let	82, 102, 98, 91	4	373	93,25	35 013
> 25 let	81, 77, 75, 85, 69	5	387	77,4	30 101

17.7 Na pěti vybraných místech s omezenou rychlostí jízdy na 50 km/h byla měřena rychlost projíždějících automobilů (viz tabulku 17.25). Na hladině významnosti $\alpha = 0,05$ otestujte rovnost průměrů všech měření, jestliže výběry nepocházejí z normálního rozdělení.

Tabulka 17.25: Rychlosti automobilů v km/h

M_1	M_2	M_3	M_4	M_5
55,4	68,4	52,1	62,3	67,5
61,2	57,9	49,5	63,5	54,3
65,8	58	47,5	63,4	56,3
59,3	47,8	58	56,2	55,5
62,5	63,8	37,8	51	57,2
58,4	52,3	57,2	41,8	65,5
48,1	62,5	50,5	44,5	60,2

17.8 Entomolog studuje vertikální distribuci určitého druhu mouchy v listnatém lese a získal pět souborů jedinců z každé ze tří vegetačních vrstev (bylinné patro, keřové patro, stromové patro). Na hladině významnosti $\alpha = 0,05$ testujte hypotézy: H_0 : Vertikální distribuce druhu je stejná ve všech třech vegetačních patrech. H_1 : Vertikální distribuce druhu není stejná ve všech třech vegetačních patrech.

Tabulka 17.26: Tabulka s počty jedinců na m^3 vegetačního patra (v závorce je pořadí)

Jedinec č.	Vertikální distribuce na m^3		
	Bylinné patro	Keřové patro	Stromové patro
1	14,0 (15)	8,4 (11)	6,9 (8)
2	12,1 (14)	5,1 (2)	7,3 (9)
3	9,6 (12)	5,5 (4)	5,8 (5)
4	8,2 (10)	6,6 (7)	4,1 (1)
5	10,2 (13)	6,3 (6)	5,4 (3)

17.9 V tabulce 17.27 jsou uvedeny naměřené hodnoty měrného odporu plužního tělesa na středně těžké půdě při různých pracovních rychlostech: $v_1 = 4$ km/h, $v_2 = 6$ km/h, $v_3 = 8$ km/h.

Tabulka 17.27: Tabulka měrných odporů plužního tělesa v 10^5 Pa

Rychlost	Měrný odpor			
	Měření č. 1	Měření č. 2	Měření č. 3	Měření č. 4
v_1	0,71	0,66	0,63	0,65
v_2	0,59	0,64	0,67	0,69
v_3	0,70	0,72	0,75	0,73

Posuďte, zda vliv pracovní rychlosti na měrný odpor je statisticky významný či nikoliv.

17.10 V tabulce 17.28 jsou uvedeny hodnoty tří různých výběrů. Rozhodněte, zda všechny tři výběry mohou pocházet ze téhož základního souboru se stejným rozdělením.

Tabulka 17.28: Tabulka hodnot tří různých výběrů

Číslo výběru	Hodnoty
1	2, 3, 10, 12, 18, 20,5, 26, 28
2	1, 4,5, 10, 10, 15, 16, 19, 23, 26, 30
3	4,5, 6, 7, 8, 13, 14, 17, 20,5, 22, 24, 26, 29

17.11 V roce 2006 se provedl test majonéz od různých výrobců. Hodnotitelé pro pořadovou zkoušku dostali v kádinkách najednou pět vzorků a řadili je od nejchutnějšího po nejméně chutný. Výsledky hodnocení ukazuje tabulka 17.29. Jsou jednotlivé vzorky statisticky významně odlišné ($\alpha = 0,05$)?

Tabulka 17.29: Tabulka výsledků hodnocení majonéz různých výrobců (průměrné hodnoty na stobodové stupnici)

Hodnocená vlastnost / Vzorek	1	2	3	4	5
Příjemnost barvy	74	43	76	46	58
Intenzita barvy	62	37	77	37	44
Příjemnost vůně	71	56	54	48	43
Intenzita vůně	48	34	55	29	33

17.12 Společnost provozuje čtyři prodejny jízdních kol. Na hladině významnosti 5% testujte, zda se rozdíly v prodeji kol v jednotlivých prodejnách statisticky významně liší. K dispozici jsou počty prodaných kol v jednotlivých prodejnách za posledních pět měsíců (viz tabulku 17.30).

Tabulka 17.30: Tabulka s počty prodaných kol v jednotlivých prodejnách

Měsíc / Prodejna	A	B	C	D
I	23	25	22	18
II	18	15	20	16
III	16	16	17	20
IV	20	19	15	24
V	21	22	18	25

Řešení úloh

17.1 Test o rovnosti rozptylů: ANOVA: jeden faktor: testová F -statistika 1,391 < kritická hodnota asi 3,885, tedy nezamítáme rovnost rozptylů (rovněž p -hodnota F -testu asi 0,286 > 0,05).

Na základě rovnosti rozptylů testujeme střední hodnoty: H_0 : Mezi výsledky testu není průkazný rozdíl. proti alternativní hypotéze H_1 : Mezi výsledky testu je průkazný rozdíl. test ANOVA: jeden faktor: testová statistika F asi 8,298 > kritická hodnota asi 3,885 (rovněž p -hodnota 0,005 46 < 0,05), tedy mezi středními hodnotami všech tří souborů je průkazný rozdíl. Tukeyův test: q kritická hodnota testu je asi 3,773, významný rozdíl je pouze mezi střední hodnotu výsledků z gymnázia a odborných učilišť.

17.2 Test o rovnosti rozptylů: ANOVA: jeden faktor, F -statistika asi 1,175 < F -kritická hodnota asi 2,922 (rovněž p -hodnota asi 0,336 > 0,05), rozptyly jsou si na zvolené hladině významnosti rovny.

Testujeme hypotézu H_0 : Mezi zkoumanými prasnicemi není průkazný rozdíl v počtu selat. Proti alternativní hypotéze H_1 : Mezi zkoumanými prasnicemi je průkazný rozdíl v počtu selat. Test středních hodnot: test ANOVA: 1 faktor, testová statistika F asi 11,443 > F kritická hodnota asi 2,922 (rovněž p -hodnota asi 0,000 036 < 0,05).

Tukeyův test: q kritická hodnota testu je asi 3,093, podstatné rozdíly v porodnosti jsou mezi prasnicemi AB , AC , BC a CD .

17.3 Kritická hodnota Tukeyova testu je na hladině významnosti 6 a se 45 stupni volnosti asi 4,208, významný rozdíl v průměrných výnosech obilí je pouze mezi zemědělci B a F (testová statistika asi -5,119).

Tabulka 17.31: Tabulka s dílčími výsledky řešení

17.4	x_i	x_i^2	$\sum x_{ij}^2$	$\bar{x}_i$
	112	12 544	2 516,3	22,4
	132,1	17 450,41	3 494,1	26,4
	133,9	17 929,21	3 586,6	26,8
	141,1	19 909,21	3 983,9	28,2
	$\sum x_{i0} = 519,1$	$\sum x_i^2 = 67 832,83$	$\sum \sum x_{ij}^2 = 13 581$	

Počet skupin (řádků) $m = 4$, počet sloupců $n = 5$.

Součty čtverců odchylek: $S_l = \frac{1}{n} \sum x_i^2 - c = 93,325 5$, $S_r = S - S_l = 14,304$, $S = \sum \sum x_{ij}^2 - c = 107,629 5$, $c = \frac{\sum x_{i0}^2}{m \cdot n} = 13 473,240 5$.

Tabulka 17.32: Výpočtová tabulka analýzy rozptylu

Variabilita	SČO	Stupně volnosti	Rozptyl
Mezi třídami	93,3255	$m - 1$ (3)	31,107
Uvnitř tříd (reziduální)	14,304	$m \cdot (n - 1)$ (16)	0,894
Celkem	107,6295	$mn - 1$ (19)	

F -test: $F = 34,795$, $F_{0,05}(3; 16) = 3,24$, $F > F_{0,05}(3; 16)$, proto zamítáme hypotézu o významných rozdílech ve výnosech jednotlivých odrůd.

17.5 $H_0: \mu_A = \mu_B = \mu_C = \mu_D$. $S_A = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2 = 44$, $S_R = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{x}_{ij} - \bar{x}_i)^2 = 20$. Výběrové aritmetické průměry pocházejí ze základních souborů se shodnými průměry, tj. složení krve u všech základních souborů různých syndromů má průměrné fyziologické hodnoty shodné.

17.6 Testujeme $H_0: \mu_1 = \mu_2 = \mu_3$. proti $H_1: \text{Ne všechna } \mu_i \text{ si jsou rovna}$. Z tabulky 17.24 je $n = 13$, $\sum x_i = 1\,053$, $\sum_i \sum_j x_{ij}^2 = 86\,679$. Testovací kritérium $F = \frac{n-k}{k-1} \cdot \frac{S_A}{S_e}$ má rozdělení $F(k-1; n-k) = F(2; 10)$. Kritická hodnota $F_{0,05}(2; 10) \doteq 2,103$, kritické oblasti jsou $(4,103; \infty)$, $F_{0,01}(2; 10) \doteq 47,559$, kritická oblast je $(47,559; \infty)$. Hodnota testovacího kritéria: $S_T = 1\,386$, $S_A = 905,3$, $S_e \doteq 905,3$. Hodnota testovacího kritéria se vypočítá z informací tabulky analýz rozptylu (viz tabulku 17.33).

Tabulka 17.33: Tabulka analýzy rozptylu

Zdroj variability	Součet čtverců	Počet stupňů volnosti	Rozptyl	F
Délka praxe	905,3	2	452,65	$F = \frac{452,65}{48,07} \doteq 9,416$
Reziduální	480,7	10	48,07	
Celkový	1 386	12	-	

Na hladině významnosti 0,05 je $F = 9,416 > 4,103$, tedy H_0 zamítáme, na hladině významnosti 0,01 je $F = 9,416 > 47,559$, tedy H_0 také zamítáme. Šetření prokázalo statisticky významný vliv délky praxe na produktivitu práce, odlišuje se produktivita práce s délkou praxe 10 až 25 let.

17.7 Kruskalův-Wallisův test: $T \doteq 6,995 < \chi^2(4) \doteq 9,49$, tj. nezamítáme H_0 o rovnosti středních hodnot, tedy automobily projížděly místy M_1 – M_5 srovnatelnými průměrnými rychlostmi na hladině 0,05.

17.8 $n_1 = n_2 = n_3 = 5$, $N = 15$, Kruskalův-Wallisův test: $R_1 = 64$, $R_2 = 30$, $R_3 = 26$, $T \doteq 8\,720 > \chi^2(15) \doteq 5,78$, zamítáme H_0 , tj. vertikální distribuce druhu není stejná ve všech třech patrech. Liší se bylinné patro.

17.9 Naměřené hodnoty sloučíme do jednoho souboru a vyhodnotíme pořadí jednotlivých hodnot (viz tabulku 17.34).

Tabulka 17.34: Tabulka odporů plužního tělesa

v_1	0,71	0,66	0,63	0,65	
Pořadí	9	5	2	4	$T_1 = 20$
v_2	0,59	0,64	0,67	0,69	
Pořadí	1	3	6	7	$T_2 = 17$
v_3	0,70	0,72	0,75	0,73	
Pořadí	8	10	12	11	$T_3 = 41$

$k = 3$, $n_1 = n_2 = n_3 = n_4 = 3$, $n = 12$, Kruskalův-Wallisův test: $T \doteq 6\,577 > \chi^2(2) \doteq 5,991$, zamítáme H_0 o stejném měrném odporu při různých rychlostech.

17.10 $R_1 = 119,5$, $\bar{R}_1 = 14,94$, $R_2 = 154,5$, $\bar{R}_2 = 15,45$, $R_3 = 191$, $\bar{R}_3 = 15,92$, $T = 0,07$, $p = 4$ skupiny stejných pořadí, první obsahuje pořadí 4 a 5, nahradíme je průměrným pořadím 4,5, tedy $T_1 = 2^3 - 2 = 6$. Pro druhou a čtvrtou skupinu je $T_2 = T_4 = 3^3 - 3 = 24$. Pro třetí skupinu je $T_3 = T_1 = 6$. Korelační faktor je roven $1 - \frac{60}{26\,970} = 0,998$. Protože $T = 0,07/0,998 \doteq 0,07 < \chi^2(2) = 5,99$, nezamítáme H_0 , že všechny tři výběry pocházejí z téhož rozdělení.

17.11 Friedmanův test, hodnoty v tabulce 17.29 nahradíme pořadovými čísly v rámci každého řádku a najdeme součty za jednotlivé vzorky (viz tabulku 17.35).

Tabulka 17.35: Tabulka zpracování výsledků hodnocení majonéz různých výrobců

Hodnocená vlastnost / Vzorek	1	2	3	4	5	Celkem
Příjemnost barvy	4	3	3	3	4	17
Intenzita barvy	2	2	4	2	3	13
Příjemnost vůně	3	4	1	4	2	14
Intenzita vůně	1	1	2	1	1	6

Protože $T = 10,785 \geq 7,81$, zamítáme H_0 o stejné úrovni kvality testovaných vzorků majonéz.

17.12 Friedmanův skupinový test, hodnoty v tabulce 17.30 nahradíme pořadími v rámci každého řádku a najdeme součty pro jednotlivé prodejny (viz tabulku 17.36).

Tabulka 17.36: Tabulka s počty prodaných kol v jednotlivých prodejnách

Měsíc / Prodejna	A	B	C	D
I	3	4	2	1
II	3	1	4	2
III	1,5	1,5	3	4
IV	3	2	1	4
V	2	3	1	4
R_1	12,5	11,5	11	15

$T = 1,14$, počet souborů (podle měsíců) je 5, aproximujeme rozdělení statistiky T rozdělením χ^2 o 3 stupních volnosti ($v = k - 1 = 3$), kritická hodnota je kvantil $\chi^2_{0,05}(3) \doteq 7,81$. $T < 7,81$, nezamítáme H_0 o shodě středních hodnot, tedy mezi počtem prodaných kol v jednotlivých prodejnách nebyl prokázán žádný statisticky významný rozdíl.

Kapitola 18

Korelační analýza

18.1 Pearsonův výběrový korelační koeficient

Jestliže mezi studovanými statistickými znaky existuje lineární závislost, využijeme pro ohodnocení těsnosti vztahu výběrový korelační koeficient. Připomeňme, že v odstavci 9.2 jsme pro dvojici náhodných veličin X, Y s nenulovými rozptyly zavedli pojem koeficientu korelace předpisem

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var } X} \sqrt{\text{var } Y}} = \frac{\mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y)}{\sqrt{\text{var } X} \sqrt{\text{var } Y}} \quad (18.1)$$

Jestliže neznámé rozptyly a neznámou kovarianci nahradíme jejich (nestrannými) odhady

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2, \quad S_Y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2, \\ S_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

dostaneme výběrovou verzi korelačního koeficientu, kterou formálně zavedeme následující definicí.

Definice 18.1 (Pearsonův výběrový korelační koeficient)

Nechť $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ je náhodný výběr z nějakého dvourozměrného rozdělení. Pearsonův výběrový korelační koeficient je definován jako

$$R_P = \frac{S_{XY}}{\sqrt{S_X^2 \cdot S_Y^2}}. \quad (18.2)$$

Po drobné úpravě dostaneme vzorec vhodný pro výpočet:

$$R_P = \frac{\sum X_i Y_i - n \bar{X} \bar{Y}}{\sqrt{(\sum X_i^2 - n \bar{X}^2)(\sum Y_i^2 - n \bar{Y}^2)}}.$$

Z Cauchyho–Schwarzovy–Buňakovského nerovnosti přímo dostaneme, že $-1 \leq R_P \leq 1$. Pearsonův korelační koeficient se používá v případě, kdy chceme posoudit, zda mezi veličinami X a Y existuje lineární závislost. Jeho hodnota r_P je bezrozměrné číslo z intervalu $[-1; 1]$. Výběrový korelační koeficient není nestranný odhad ρ , i když výběrový rozptyl S^2 a výběrová kovariance S_{XY} ano.

Předpokládejme nyní, že $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ je náhodný výběr z nějakého dvourozměrného normálního rozdělení a $\text{var } X > 0$, $\text{var } Y > 0$, $|\rho| < 1$. Testujme nyní hypotézu $H_0: \rho = 0$ proti alternativě $H_1: \rho \neq 0$. Za platnosti hypotézy H_0 a za výše uvedených předpokladů má statistika

$$T = \frac{R_P}{\sqrt{1 - R_P^2}} \sqrt{n - 2} \sim t_{n-2}$$

Studentovo t rozdělení o $n - 2$ stupních volnosti. Tudíž hypotézu H_0 zamítneme na hladině α , v případě, že

$$|t| \geq t_{n-2} \left(1 - \frac{\alpha}{2}\right).$$

U tohoto testu je normalita náhodného výběru podstatný předpoklad. Nejsme-li si jisti tímto předpokladem, použijeme pro test nezávislosti raději Spearmanův korelační koeficient.

Při posuzování významu lineární závislosti dvou numerických proměnných se můžeme řídit obecně uznávanými pravidly:

- Je-li $r_P = 0$, jsou studované proměnné nekorelované.
- Je-li $|r_P| = 1$, jsou studované proměnné lineárně závislé: pro $\rho = 1$ jde o přímou lineární závislost, pro $\rho = -1$ nepřímou lineární závislost.
- Je-li $|r_P| < 0,3$, mluvíme o nízkém stupni korelační závislosti.
- Je-li $0,3 \leq |r_P| < 0,5$, mluvíme o mírném stupni korelační závislosti.
- Je-li $0,5 \leq |r_P| < 0,7$, mluvíme o středním stupni korelační závislosti.
- Je-li $0,7 \leq |r_P| < 0,9$, mluvíme o vysokém stupni korelační závislosti.
- Je-li $0,9 \leq |r_P| < 1$, mluvíme o velmi vysokém stupni korelační závislosti.

Příklad 18.1.  U deseti dvojčat byla zjištěna následující váha (v gramech) (viz tabulku 18.1). Ověřte, zda jsou váhy dvojčat korelované.

Tabulka 18.1: Váha dvojčat v gramech

Starší	Mladší	Starší	Mladší
2 440	2 700	2 690	2 350
3 500	3 080	2 750	3 500
2 820	2 200	2 750	2 500
2 540	2 700	2 650	2 420
2 650	2 550	2 200	2 520

	Sloupec 1	Sloupec 2
Sloupec 1	1	
Sloupec 2	0.338520934	1

Obrázek 18.1: Výřez z MS Excel v části Korelace nástroje Analýza dat

Řešení. Postupně vypočteme $s_X^2 \doteq 111\,965,6$ `= VAR.S(B4:B13)`, podobně $s_Y^2 = 145\,240$ `= VAR.S(C4:C13)`, $n = 10$, $s_{XY} = 43\,169$, $r_P^2 \doteq 0,191$. Hodnota korelačního koeficientu je $0,3385$ `= CORREL(B4:B13;C4:C13)`. Vypočteme hodnotu testové statistiky $t \doteq 0,550$. Volíme $\alpha = 0,05$ a vypočítáme kritickou hodnotu $t_8(0,975) \doteq 2,306$ `= T.INV(0,975;8)`. Jelikož $|t| \not\geq t_8(0,975)$, nezamítáme hypotézu o nekorelovanosti vah dvojčat na hladině významnosti $\alpha = 0,05$. Hodnota korelačního koeficientu není dostatečně vysoká na to, abychom dokázali zamítnout nekorelovanost porodních hmotností dvojčat.



V MS Excel lze pro výpočet hodnoty korelačního koeficientu využít také možnost Korelace v nástroji Analýza dat. Obrázek 18.1 ukazuje tabulku s výsledkem. △

18.2 Spearmanův korelační koeficient

Spearmanův¹ korelační koeficient r_s je neparametrický odhad korelačního koeficientu a využijeme jej tehdy, jestliže zjištěná závislost mezi statistickými znaky (spojitými náhodnými veličinami) není čistě lineární, ale je to

¹Charles Edward Spearman (1863–1945), britský psycholog a statistik.

Tabulka 18.2: Pořadí žáků podle nervové lability a podle matematiky

R_t (labilita)	1	2	3	4	5	6	7	8	9	10
Q_t (matematika)	9	3	8	5	4	2	10	1	7	6
$R_t - Q_t$	-8	-1	-5	-1	1	4	-3	7	2	4

nějaká monotonní funkce. Protože se u Spearmanova koeficientu vyjadřuje závislost studovaných znaků pomocí pořadí, označuje se tento koeficient také jako Spearmanův koeficient pořadové korelace.

Považujme $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ za náhodný výběr z nějakého dvourozměrného rozdělení. K sestrojení Spearmanova korelačního koeficientu nám postačí pouze znalost pořadí $X_1, X_2, \dots, X_n$ a pořadí $Y_1, Y_2, \dots, Y_n$. Jsou-li pořadí hodně podobná, svědčí to o závislosti mezi X_i a Y_i . Spearmanův korelační koeficient je – na rozdíl od Pearsonova korelačního koeficientu – odolný vůči odlehlým hodnotám.

Definice 18.2 (Spearmanův korelační koeficient)

Nechť $R_1, R_2, \dots, R_n$ označují pořadí $X_1, X_2, \dots, X_n$ a necht $Q_1, Q_2, \dots, Q_n$ označují pořadí $Y_1, Y_2, \dots, Y_n$. Spearmanův korelační koeficient je pak definován vztahem

$$R_S = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^n (R_i - Q_i)^2.$$

Hodnota koeficientu r_S se stejně jako u Pearsonova korelačního koeficientu pohybuje v mezích intervalu $[-1; 1]$, hodnota blízká 1 značí silnou přímou závislost, hodnota blízká -1 značí silnou nepřímou závislost.

Testujeme-li hypotézu $H_0: \rho = 0$ proti alternativě $H_1: \rho \neq 0$, pak jsou kritické hodnoty pro r_S tabelovány pro $n \leq 30$. Při $n > 30$ zamítneme hypotézu H_0 v případě, že

$$|r_S| \geq \frac{u(1 - \frac{\alpha}{2})}{\sqrt{n-1}},$$

kde $u(1 - \frac{\alpha}{2})$ je kritická hodnota rozdělení $N(0; 1)$.

Příklad 18.2.  (viz [1]) Bylo sledováno deset žáků. Na základě psychologického vyšetření byli tito žáci seřazeni podle nervové lability (čím byl žák více labilní, tím dostal vyšší pořadí R_i). Kromě toho sledovaní žáci dostali pořadí Q_i na základě svých výsledků v matematice (nejlepší žák v matematice dostal 1). Výsledky jsou uvedeny v tabulce 18.2. Ověřte závislost mezi nervovou labilitou a výsledky v matematice.

Řešení. Nejdříve určíme rozdíly pořadí. Dostáváme

$$r_S = 1 - \frac{6}{10 \cdot 99} \cdot (8^2 + 1^2 + 5^2 + 1^2 + 1^2 + 4^2 + 3^2 + 7^2 + 2^2 + 4^2) \doteq -0,127.$$

Jsou-li pořadí z tabulky 18.2 zapsána v buňkách A3:B12, potom Spearmanův korelační koeficient můžeme spočítat také přímo pomocí funkce

`= CORREL(A3:A13;B3:B13)`.

Kritická hodnota odpovídající hladině $\alpha = 0,05$ činí přibližně 0,636 4. Protože ji $|r_S|$ nepřekračuje, nemůžeme zamítnout hypotézu, že nervová labilita a výsledky v matematice jsou nezávislé. $\triangle$

18.3 Korelace versus kauzalita

Korelace a kauzalita jsou dva klíčové pojmy, které se často používají v kontextu statistiky a vědeckého zkoumání. Korelace se týká vztahu mezi dvěma proměnnými, jak jsme si ukázali v kapitolách výše. Změna v hodnotách jedné proměnné způsobuje změny v hodnotách druhé proměnné. Korelace mezi proměnnými však automaticky neznamená, že změna jedné proměnné je příčinou změny hodnot druhé proměnné. Pokud tomu tak je, mluvíme o příčinné souvislosti neboli kauzalitě. To znamená, že jedna událost je důsledkem výskytu druhé události, tj. mezi oběma událostmi existuje příčinný vztah. Korelace sice neimplikuje kauzalitu, ale na druhé straně její absence ji vylučuje. To znamená, pokud nejsou proměnné korelované, nemůžeme hovořit o kauzalitě.

V případě, že mezi dvěma veličinami je silná korelace (souvislost, nebo také mluvíme o asociativním vztahu), ale změna ani jedné proměnné není příčinou změny hodnot té druhé, může tato souvislost být náhodná nebo do vztahu těchto dvou proměnných zasahuje proměnná třetí. Uvedeme si několik příkladů.

Příklad 18.3 (pozitivní korelace, ale bez kauzality). V letních měsících, kdy je teplo, se zvyšuje prodej zmrzliny a zároveň se zvyšuje počet lidí, kteří se koupou, což může vést k vyššímu počtu utonutí. Tyto dvě proměnné (prodej zmrzliny a počet utonutí) jsou pozitivně korelované, protože obě rostou ve stejném období. Nicméně, prodej zmrzliny nezpůsobuje utonutí. Společným faktorem, který ovlivňuje obě proměnné, je teplé počasí.

Příklad 18.4 (nesprávná kauzalita). Často se můžeme setkat s tvrzením, že násilí ve videohrách či televizi (jev *A*) způsobuje, že děti se stávají více násilnými (jev *B*). Což je kauzální vztah. Avšak můžeme také tvrdit, že děti s přirozeným sklony k násilnému chování mají tendenci vyhledávat hry

a pořady s násilnou tematikou. Proto nemůžeme přesně určit, zda A způsobuje B , nebo B způsobuje A . V takovém případě se jedná pouze o korelaci, nikoliv o kauzalitu.

Příklad 18.5 (korelace způsobená třetí proměnnou – spotřeba zmrzliny a výše tržeb). Data mohou ukazovat pozitivní korelaci mezi vyššími tržbami zmrzlinového stánku a zvýšeným výskytem úpalu u lidí. Na první pohled bychom mohli usoudit, že vyšší tržby způsobují úpal (anebo obráceně). Avšak ve skutečnosti je zde třetí proměnná, která ovlivňuje oba tyto faktory a lépe vysvětluje pozorovaný jev – počasí. V teplém počasí se lidé více snaží ochladit a vyhledávají zmrzlinové stánky, což vede k vyšším tržbám. Zároveň je v teplém počasí větší riziko úpalu. Tedy počasí je klíčovou proměnnou, která má vliv na oba tyto jevy.

Příklad 18.6 (falešná korelace, náhodná – počet filmů, ve kterých se objevil Nicolas Cage, a počet utonutí v bazénech). Data mohou ukazovat na to, že v letech, kdy Nicolas Cage hrál ve více filmech, se také zvýšil počet utonutí v bazénech. Tato korelace je čistě náhodná a neexistuje žádný logický důvod, proč by počet filmů s Nicolasem Cagem měl ovlivňovat počet utonutí v bazénech. Tento příklad ukazuje, jak mohou být některé korelace zcela náhodné a bez jakéhokoli příčinného vztahu. Tyto příklady mohou být použity k ilustraci toho, jak důležité je pečlivě analyzovat data a nevyvozovat závěry pouze na základě korelace.

Příklad 18.7 (kauzalita bez třetí proměnné – cvičení a fyzická kondice). Když člověk pravidelně cvičí, zlepšuje se jeho fyzická kondice. V tomto případě je cvičení příčinou a zlepšení fyzické kondice je důsledkem. Neexistuje žádná třetí proměnná, která by tento vztah vysvětlovala. Přímo vidíme, že pravidelné cvičení vede ke zlepšení fyzické kondice.

18.4 Úlohy

18.1 Tabulka 18.3 ukazuje počty pracovníků ve výzkumu a vývoji (VaV) v tisících a celkovou výši vyplacených mezd v mld. Kč v letech 2010–2020. Ověřte závislost mezi počtem pracovníků a výší mzdových nákladů. Sestrojte bodový graf a vypočítejte Pearsonův a Spearmanův korelační koeficient.

Tabulka 18.3: Počty pracovníků ve VaV v ČR v tisících a výše mzdových nákladů v mld. Kč v letech 2010–2020

Rok	2010	2011	2012	2013	2014	2015
Počet pracovníků	52	57	60	62	64	67
Mzdy	26	25	30	29	45	46
Rok	2016	2017	2018	2019	2020	
Počet pracovníků	65	70	75	79	81	
Mzdy	35	35	71	70	69	

18.2 Tabulka 18.4 ukazuje průměrný počet zmetků za směnu u jednotlivých pracovníků a délku zaměstnání ve firmě. Určete Pearsonův a Spearmanův korelační koeficient. Má délka zaměstnání jednotlivých pracovníků vliv na množství zmetků při výrobě? Na hladině významnosti 0,05 testujte hypotézu o nezávislosti počtu zmetků na délce doby zaměstnání ve firmě.

Tabulka 18.4: Počty zmetků (z) z výroby a délka doby (d) zaměstnání jednotlivých pracovníků ve firmě

d	1	2	3	10	5	25	15	3	10	0	4	40	4	11	8	30	2	5
z	4	6	1	2	1	4	0	7	2	5	2	2	3	1	5	2	0	1

18.3 Tabulka 18.5 ukazuje průměrný počet zmetků za směnu, který připadá na daný stroj o určitém stáří. Výpočtem Pearsonova a Spearmanova korelačního koeficientu rozhodněte o míře korelovanosti těchto veličin. Vytvořte bodový graf závislosti obou veličin, zobrazte regresní přímkou a její rovnici. Nekorelovanost otestujte na hladině významnosti 0,05.

Tabulka 18.5: Počty zmetků (z) z výroby a stáří výrobního stroje (s)

s	10	2	7	12	20	25	11	3	0	5	16	22
z	4	6	1	2	1	4	0	7	2	5	2	2

18.4 Na základě náhodného výběru o rozsahu 24 chceme ohodnotit sílu vztahu mezi stářím a krevním tlakem testovaných žen. Data ukazuje tabulka 18.6. Na hladině významnosti 0,05 otestujte, zda jsou veličiny V a T nekorelované.

Tabulka 18.6: Věk žen v v letech a jejich krevní tlak t v mm Hg

v	46	52	30	34	60	45	63	39	49	41	31	55
t	130	140	113	119	145	133	151	117	145	135	115	136

v	46	50	30	63	51	42	35	70	45	50	25	37
t	133	141	110	137	132	135	111	165	125	135	112	115

18.5 Tabulka 18.7 ukazuje data o obratu O prodejny (v milionech Kč) a velikosti její prodejní plochy S . Obchody mají srovnatelnou polohu v témž městě a také podobnou návštěvnost.

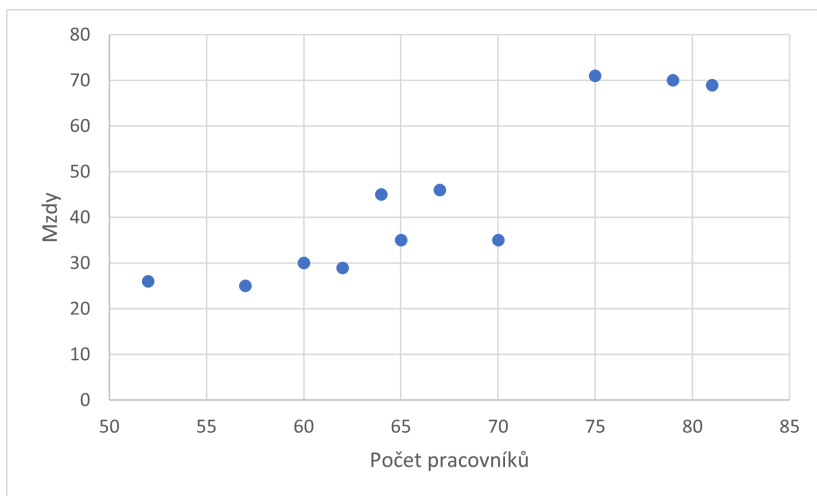
- Spočítejte korelační koeficient.
- Na hladině významnosti 0,01 testujte hypotézu, že mezi velikostí prodejní plochy a výší obratu je pozitivní závislost.

Tabulka 18.7: Obrat o v mil. Kč a velikost s prodejní plochy obchodu v m²

s	55	52	66	58	52	45	68	50	49	60
o	8,53	8,30	11,52	9,50	8,99	7,90	11,49	8,20	8,55	10,72

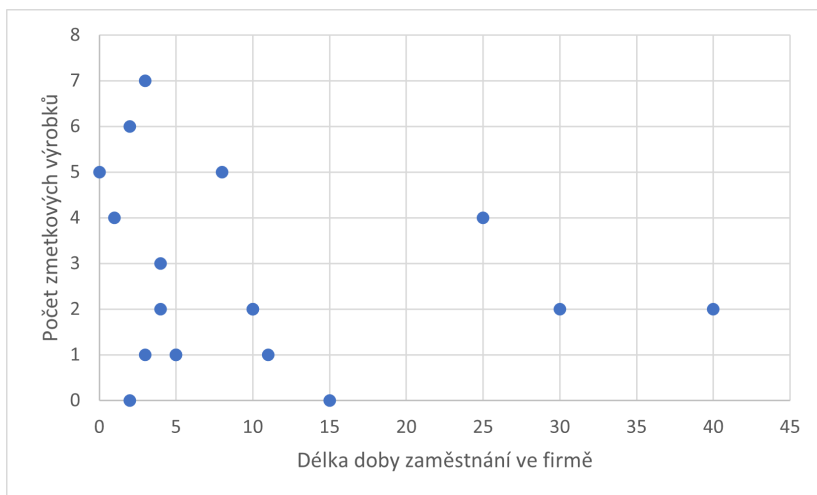
Řešení úloh

18.1 Obrázek 18.2 ukazuje graf závislosti vyplacených mezd na množství pracovníků. Pearsonův korelační koeficient $r_P \doteq 0,987$, Spearmanův korelační koeficient $r_S \doteq 0,991$. Oba koeficienty i jejich blízkost prokazují vysokou pozitivní korelovanost množství pracovníků ve výzkumu a vývoji s množstvím vyplacených mezd.



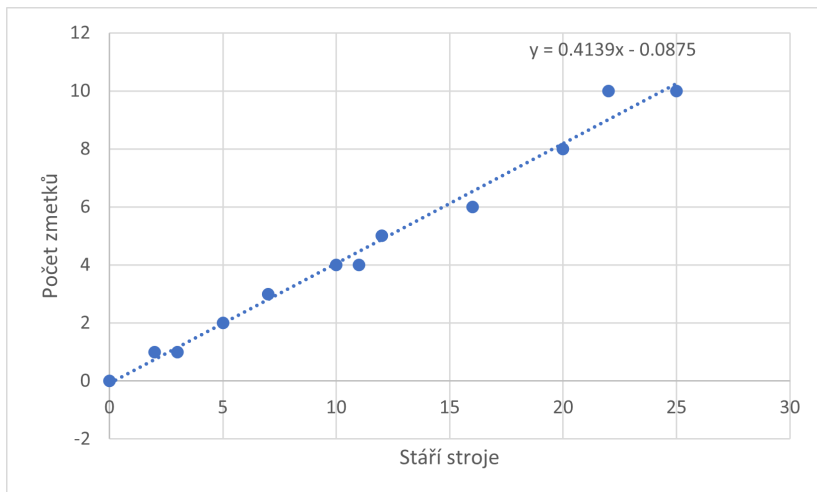
Obrázek 18.2: Bodový graf závislosti vyplacených mezd v mld. Kč na počtu pracovníků v tisících

18.2 Z hodnot $r_P \doteq -0,190$, $r_S \doteq -0,426$ usuzujeme na malou (zápornou) korelovanost hodnot proměnných. Vizuálně pozorujeme také vysokou rozptýlenost dat (viz obrázek 18.3). Protože $|r_S|$ nepřevyšuje kritickou hodnotu 0,471 6, nelze zamítnout hypotézu o nekorelovanosti veličin: nelze jednoznačně soudit, že s délkou doby zaměstnání ve firmě klesá počet zmetkových výrobků.



Obrázek 18.3: Bodový graf závislosti počtu zmetků na délce doby zaměstnání ve firmě v rocích

18.3 Z hodnot $r_P \doteq 0,993$, $r_S \doteq 0,994$ usuzujeme na velmi vysokou kladnou korelovanost hodnot proměnných, tedy lze soudit, že čím je stroj starší, tím produkuje více zmetků. To potvrzuje i nízká rozptýlenost dat (viz obrázek 18.4). Protože $|r_S|$ převyšuje kritickou hodnotu 0,580 4, zamítáme hypotézu o nekorelovanosti obou veličin.



Obrázek 18.4: Bodový graf závislosti počtu zmetků na stáří stroje v letech

18.4 Předpokládáme, že oba znaky jsou v populaci rozděleny normálně. $r_P \doteq 0,912$. Testování nezávislosti: $H_0 : \rho_{V,T} = 0$, $H_1 : \rho_{V,T} \neq 0$, $|t| \doteq 7,991 > t_{22}(0,975) \doteq 2,16 \Rightarrow$ zamítáme H_0 o nekorelovanosti veličin V a T , veličiny V a T jsou korelované a tedy mezi věkem a krevním tlakem žen je určitá lineární vazba.

18.5

- a) $r_P \doteq 0,952$;
- b) testování pozitivní závislosti: $H_0 : \rho_{O,S} = 0$, $H_1 : \rho_{O,S} > 0$, za předpokladu normálního rozdělení S , O je testová statistika $t \doteq 8,76$, kritická hodnota $t_8(0,95) \doteq 1,86$, $t > t_k \Rightarrow$ zamítáme H_0 o nekorelovanosti veličin S , O , veličiny S , O jsou kladně korelované, lze říci, že s nárůstem velikosti prodejní plochy roste také obrat prodejný.

Kapitola 19

Regresní analýza

V praxi jsou některé náhodné veličiny snadno dostupné například jednoduchým měřením, zatímco jiné veličiny se určují obtížně nebo se o nich dozvíme až s velkým časovým odstupem. Pokud mezi těmito dvěma druhy veličin existuje nějaký vztah, lze z jedné odhadnout druhé, resp. předpovědět je. Při specifikaci funkčního vztahu jde především o prokázání tohoto vztahu, o odhady jeho parametrů, o empirickou reprezentaci velkých datových souborů, o interpolaci chybějících hodnot a jejich předpovědi apod. Pro tyto účely slouží metody lineární regrese, jejichž základy jsou popsány v této kapitole.

19.1 Lineární regrese s jednou vysvětlující proměnnou

Regresní model

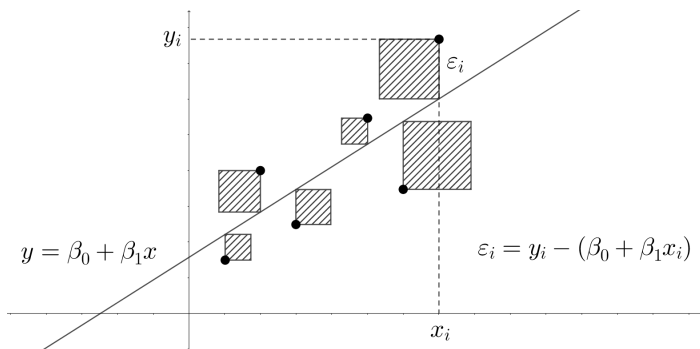
$$Y = f(x)$$

vysvětluje závislost veličiny Y na hodnotách proměnné x prostřednictvím regresní funkce f . Cílem regrese je najít regresní funkci f , známe-li n pozorovaných dvojic

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n),$$

kde x_i jsou hodnoty nezávislé (vysvětlující) proměnné x a y_i jsou hodnoty závislé (hodnoty vysvětlované veličiny Y).

Předpokládáme, že hodnoty y_i jsou naměřeny s určitou chybou e_i . Pro odvození všech testů a intervalových odhadů v průběhu celé této kapitoly klademe na chyby e_i předpoklad, že mají normální rozdělení s nulovou střední hodnotou a konstantním rozptylem, tj. $N(0; \sigma^2)$. Pro odvození bodových



Obrázek 19.1: Metoda nejmenších čtverců

odhadů tento předpoklad není nutný. Jinak řečeno, máme n pozorování vysvětlované veličiny Y v n známých hodnotách vysvětlující proměnné x , tudíž máme n rovnic:

$$Y_i = f(x_i) + e_i, \quad i = 1, 2, \dots, n.$$

Nejjednodušší formou regrese je jednoduchá lineární regrese, která předpokládá lineární závislost mezi dvěma veličinami vyjádřenou vztahem

$$f(x) = \beta_0 + \beta_1 x.$$

Cílem lineární regrese je nalezení parametrů β_0 a β_1 . Regresní parametr β_1 udává, o kolik se změní $f(x)$ při změně x o jednotku. Jde tedy o hodnotu závislou na jednotkách, ve kterých měříme x a y .

Úkol nalézt rovnici regresní přímky provedeme pomocí metody nejmenších čtverců (viz obrázek 19.1). Tato metoda spočívá v tom, že hledáme parametry β_0 a β_1 , pro něž je součet čtverců chyb modelu minimální. Tedy hledáme minimum funkce L (podle anglického *loss function*, ztrátová funkce)

$$L(\beta_0, \beta_1) = \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 x_i))^2.$$

Tudíž řešíme soustavu rovnic

$$\frac{\partial L(\beta_0, \beta_1)}{\partial \beta_0} = 0, \quad \frac{\partial L(\beta_0, \beta_1)}{\partial \beta_1} = 0.$$

Po úpravách obdržíme tyto odhady:

$$b_1 = \frac{\sum (x_i - \bar{x}) Y_i}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i Y_i - n \bar{x} \bar{Y}}{\sum x_i^2 - n \bar{x}^2}, \quad b_0 = \bar{Y} - b_1 \bar{x}, \quad (19.1)$$

kde $\bar{x} = \frac{1}{n} \sum x_i$ a $\bar{Y} = \frac{1}{n} \sum Y_i$. Odhady b_0, b_1 jsou nejlepší nestranné odhady, tzn. že odhady b_0, b_1 jsou nestranné ($\mathbb{E}b_0 = \beta_0$, $\mathbb{E}b_1 = \beta_1$) a mají nejmenší rozptyl ze všech nestranných odhadů.

Minimum funkce L

$$S_e = L(b_0, b_1) = \sum (Y_i - (b_0 + b_1 x_i))^2 = \sum Y_i^2 - b_0 \sum Y_i - b_1 \sum x_i Y_i$$

se nazývá *reziduální součet čtverců*. Odhad rozptylu chyb σ^2 je

$$S^2 = \frac{S_e}{n-2},$$

jeho odmocnina S se nazývá *směrodatná chyba regrese*. Konečně *celkový součet čtverců*

$$S_T = \sum (Y_i - \bar{Y})^2$$

vyjadřuje celkovou variabilitu vysvětlované proměnné.

Vhodnost modelu posuzujeme *koeficientem determinace*

$$R^2 = 1 - \frac{S_e}{S_T} = \frac{S_T - S_e}{S_T},$$

který vyjadřuje, jaká část celkové variability S_T je vysvětlena regresním modelem. Chyba S_e pak obsahuje to, co regresní model nedokázal vysvětlit. Koeficient determinace můžeme také počítat podle vzorce

$$R^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2},$$

kde $\hat{Y}_i = \hat{f}(x_i) = b_0 + b_1 x_i$ je regresní odhad hodnoty regresní funkce v bodě x_i . Je zřejmé, že čím blíže je R^2 jedné, tím lépe regresní model vystihuje naměřená data. Někdy se uvádí: je-li koeficient determinace větší než 0,85, můžeme říci, že model je vhodně zvolen. V oborech, které obvykle pracují pouze s pozorovanými (neexperimentálními) daty (např. v ekonometrii nebo sociálních vědách), se však tak vysokých hodnot koeficientu determinace dosahuje spíše výjimečně; přesto i v nich modely s výrazně nižšími koeficienty determinace mají svoji interpretační hodnotu.

Nejčastěji se zabýváme otázkou, zda je možné model zjednodušit tak, že hodnoty Y_i vůbec nezávisí na x_i . Tudíž testujeme hypotézu

$$H_0: \beta_1 = 0 \quad \text{proti} \quad H_1: \beta_1 \neq 0.$$

Za platnosti H_0 má testová statistika

$$T = \frac{b_1}{S} \sqrt{\sum x_i^2 - n\bar{x}^2} \sim t_{n-2}$$

Studentovo t rozdělení o $n-2$ stupních volnosti. Tudíž pokud $|T| \geq t_{n-2}(\frac{\alpha}{2})$, zamítneme hypotézu H_0 na hladině významnosti α . Zamítneme-li hypotézu H_0 tohoto testu, pak jsme vlastně potvrdili lineární závislost Y_i na x_i , zastřenou náhodnými chybami e_i .

Testovat můžeme i hypotézu $H_0: \beta_0 = 0$ proti $H_1: \beta_0 \neq 0$. V případě nezamítnutí H_0 jsme prokázali, že regresní přímka prochází počátkem. Pro testy významnosti regresních koeficientů se používá souhrnný název dílčí t -testy.

Standardním způsobem můžeme vytvořit intervalový odhad pro parametr β_1 o spolehlivosti $1 - \alpha$:

$$\left[b_1 - \frac{t_{n-2} \left(1 - \frac{\alpha}{2}\right) S}{\sqrt{\sum x_i^2 - n\bar{x}^2}}, b_1 + \frac{t_{n-2} \left(1 - \frac{\alpha}{2}\right) S}{\sqrt{\sum x_i^2 - n\bar{x}^2}} \right],$$

Intervalový odhad o spolehlivosti $1 - \alpha$ pro parametr β_0 je

$$[b_0 - t_{n-2}s_{b_0}; b_0 + t_{n-2}s_{b_0}], \quad (19.2)$$

kde s_{b_0} je směrodatná chyba odhadu koeficientu b_0 .

Hodnoty výše uvedených součtů čtverců (S_e , S_T) prezentuje statistický software většinou ve formě tabulky ANOVA, která se vztahuje k testování hypotézy, zda zvolená závislost mezi veličinami vůbec existuje. Oproti tom dílčí t -testy se používají pro zjištění statistické významnosti jednotlivých regresních koeficientů.

Někdy hledáme intervalový odhad pro $\beta_0 + \beta_1 x$ pro nějakou předepsanou hodnotu x . Ten lze odvodit jako interval

$$\left[b_0 + b_1 x - t_{n-2} \left(1 - \frac{\alpha}{2}\right) S \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum x_i^2 - n\bar{x}^2}}, \right. \\ \left. b_0 + b_1 x + t_{n-2} \left(1 - \frac{\alpha}{2}\right) S \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum x_i^2 - n\bar{x}^2}} \right].$$

Tento interval překrývá hodnotu $\beta_0 + \beta_1 x$ s pravděpodobností $1 - \alpha$. Sestrojíme-li takovéto intervaly pro všechna $x \in [\min x_i; \max x_i]$, vytvoříme tzv. pás spolehlivosti kolem regresní přímky. Hranice pásu jsou tvořeny dvěma větvemi hyperboly.

Vztah mezi regresí a korelací hodnot dvou znaků X a Y vyjadřuje vztah Pearsonova korelačního koeficientu ke koeficientu b_1 regresní přímky:

$$b_1 = R_P \frac{S_Y}{S_X}.$$

Tabulka 19.1: Data o počtu hodin x_i v hodinách provozu linky a nákladech Y_i v tisících Kč na její údržbu

x_i	275	350	250	325	375	400	300
Y_i	149	170	140	164	192	200	165

Tedy z korelačního koeficientu R_P se dá vypočítat růstový koeficient b_1 regresní přímky a obráceně. Znamená to, že interpretační síla prosté regrese je shodná s interpretační silou korelační analýzy.

Příklad 19.1.  Tabulka 19.1 obsahuje záznamy o počtu hodin provozu výrobní linky (x_i) a o nákladech na její údržbu (Y_i) v tisících Kč za prvních sedm měsíců roku.

- V MS Excel vytvořte bodový graf vyjadřující závislost nákladů na údržbu (Y_i) na počtu hodin provozu výrobní linky x_i (tento graf se nazývá korelační pole) a zvolte vhodný typ lineárního regresního modelu.
- Ověřte oprávněnost použití vybraného modelu (koeficient determinace a dílčí t -testy).
- Pokud se zvýší provoz linky o 1 hodinu, o kolik se zvýší náklady na její údržbu?
- Interpretujte sestavený regresní model.

Řešení. Nejprve budeme hledat regresní přímku $Y = b_0 + b_1x$. Dosadíme-li do vzorců 19.1, dostaneme $b_0 = 42,75$ a $b_1 = 0,387$ a rovnice regresní přímky tedy je $y = 42,75 + 0,387x$. Nyní spočtíme reziduální součet čtverců $S_e = 148,821$, tudíž odhad rozptylu chyb e_i je $s^2 = 29,76$. Celkový součet čtverců můžeme snadno spočítat jako $S_T = (n - 1)S_Y^2$, kde S_Y^2 je výběrový rozptyl Y . $S_T = 2\,771,71$. Tudíž koeficient determinace $R^2 \doteq 0,946\,3$.

Regresní přímku vložíme do bodového grafu v MS Excel výběrem *Spojnice trendu* v možnostech pod štítkem se znaménkem $+$. Hodnoty b_0 a b_1 lze odečíst z rovnice spojnice trendu, kterou lze vložit zaškrtnutím políčka *Zobrazit rovnici v grafu* na kartě *Formát spojnice trendu*. Tamtéž najdeme také možnost volby *Zobrazit v grafu hodnotu spolehlivosti R na druhou*, po jejímž využití se v grafu zobrazí koeficient determinace.

Pokud bychom hledali průsečík b_0 regresní přímky s osou y , aniž bychom hledali rovnici regresní přímky, je možné v MS Excel vhodně využít funkci

= INTERCEPT. Zde je $b_0 = 42,75$.

= INTERCEPT(C4:C10;B4:B10)

Nejkomplexnějším nástrojem v MS Excel pro regresní analýzu je analytický

VÝSLEDEK	
<i>Regresní statistika</i>	
Násobné R	0,97278316
Hodnota spolehlivosti R	0,94630708
Nastavená hodnota spolehlivosti R	0,9355685
Chyba stř. hodnoty	5,45566547
Pozorování	7

Obrázek 19.2: Hlavní výstup analytického nástroje Regrese v MS Excel pro příklad 19.1

ANOVA					
	<i>Rozdíl</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Významnost F</i>
Regrese	1	2622,892857	2622,893	88,12215	0,000231306
Rezidua	5	148,8214286	29,76429		
Celkem	6	2771,714286			

Obrázek 19.3: Analýza rozptylu ANOVA pro příklad 19.1

nástroj *Regrese*, který je součástí doplňku *Analytické nástroje*. Po aktivaci doplňku zvolte pole *Regrese* na kartě *Analýza dat*. Po vyplnění požadovaných údajů a nastavení možností výstupu, povinným údajem je oblast s daty závisle proměnné y a hodnotami jedné nebo více nezávisle proměnných x . Výstup nástroje *Regrese* pro příklad 19.1 ukazuje obrázek 19.2.

První číslo (Násobné R) udává Pearsonův korelační koeficient, jehož druhá mocnina se rovná koeficientu determinace, druhá hodnota (Hodnota spolehlivosti R) představuje koeficient determinace, třetí číslo (Nastavená hodnota spolehlivosti R) znamená upravený koeficient determinace, který se může použít, porovnáváme-li modely s různým množstvím parametrů, čtvrté číslo (Chyba střední hodnoty) udává směrodatnou chybu regrese (odhad směrodatné odchylky σ) a poslední položka (Pozorování) znamená rozsah výběru (počet pozorování).¹

Pokud by bylo potřeba samostatně vypočítat směrodatnou chybu regrese, lze v MS Excel využít funkci `= STEYX`. Pro příklad 19.1 píšeme v MS Excel `= STEYX(C4:C10;B4:B10)` $\doteq 5,456$.

Dalším výstupem může být tabulka s analýzou rozptylu ANOVA empirických hodnot závisle proměnné Y : celkový součet čtverců S_T je souč-

¹Exaktní definice uvedených pojmů lze nalézt v odborné literatuře, například [3],[4].

	Koeficienty	Chyba stř. hodnoty	t Stat	Hodnota P	Dolní 95%	Horní 95%
Hranice	42,75	13,56100163	3,1524	0,02531	7,890336	77,60966
Soubor X	0,38714286	0,041240954	9,3873	0,000231	0,28113	0,493156

Obrázek 19.4: Tabulka odhadů koeficientů regresní přímky a intervalových odhadů pro příklad 19.1

tem čtverců vyrovnaných hodnot regrese² a reziduální složky S_e (viz obrázek 19.3). Celkový součet čtverců generuje v MS Excel také funkce `= DEVSQ`.

Pro příklad 19.1 je příslušný výpočet `= DEVSQ(C4:C10)` $\doteq 2\,771,714$, jak je také uvedeno v tabulce na obrázku 19.3.

Základním výstupem analytického nástroje *Regrese* je tabulka s odhady regresních parametrů b_0 a b_1 (Koeficienty), odhady směrodatných odchylek těchto odhadů (v tabulce jako chyba střední hodnoty), testová statistika t -testu (t Stat), minimální hladina významnosti p -hodnota (Hodnota P), na které lze ještě zamítnout H_0 , a 95% intervaly spolehlivosti 95 % (Dolní 95%, Horní 95%). (Viz obrázek 19.4.)

Nástroj *Regrese* produkuje rovněž tabulku reziduí (viz obrázek 19.5) a jejich normovaných variant. Data lze využít pro grafické posouzení předpokladů lineární regrese, a to zejména nezávislosti jednotlivých residuí, homoskedasticity a lineárnosti závislosti Y a X . Vykazuje-li graf residuí nějaký trend, či *pattern*, pak je zřejmě některý z předpokladů porušen, a pro testování grafického vývoje v čase a pro výpočet Durbinova³–Watsonova⁴ testu, kterým se zde však nezabýváme.

Nyní se zabýváme hypotézou $H_0 : \beta_1 = 0$. Spočteme (nebo v tabulce na obrázku 19.4 přečteme) statistiku $t \doteq 9,38$ a porovnáme ji s hodnotou kvantilu (p -hodnota) $t_5(0,975) \doteq 2,57$ `= T.INV(0,975;5)`. Tudíž zamítáme hypotézu H_0 na 5% hladině významnosti. Jak koeficient determinace, tak tento test nám potvrdil vhodnost tohoto lineárního modelu.

Podívejme se ještě na intervalové odhady. Intervalový odhad o spolehlivosti 95 % pro parametr β_1 je $[0,281\,1; 0,493\,1]$. Odtud je také vidět, že zamítáme hypotézu H_0 . Pás spolehlivosti kolem regresní přímky není na obrázku 19.6 vyznačen.

Zbývá ještě interpretace modelu: absolutní člen b_0 modelu reprezentuje fixní měsíční náklady (náklady vynaložené v každém měsíci bez ohledu na

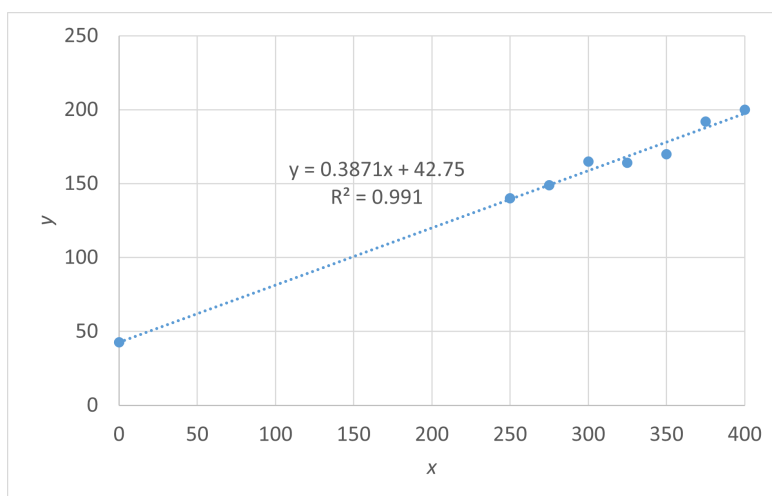
²Tato veličina není v teoretické části zavedena, používá se však v regresní analýze k posouzení, jak dobře regresní rovnice odpovídá pozorovaným datům. Blíže k tomu v Nápovědě MS Excel.

³James Durbin (1923–2012), britský statistik a ekonometr.

⁴Geoffrey S. Watson (1921–1998), australský statistik.

REZIDUA			
Pozorování	Očekávaná Y	Rezidua	Normovaná rezidua
1	149,214286	-0,214285714	-0,043026509
2	178,25	-8,25	-1,656520593
3	139,535714	0,464285714	0,093224103
4	168,571429	-4,571428571	-0,917898857
5	187,928571	4,071428571	0,817503669
6	197,607143	2,392857143	0,480462683
7	158,892857	6,107142857	1,226255504

Obrázek 19.5: Tabulka reziduí pro příklad 19.1

Obrázek 19.6: Závislost provozních nákladů na době provozu (body zobrazují naměřené hodnoty, plná čára představuje odhadnutou regresní přímku $Y = b_0 + b_1x$)

délku provozu linky) a lineární člen b_1x je odhadem variabilních nákladů přímo úměrných délce provozu linky. $\triangle$

19.2 Lineární regrese s více vysvětlujícími proměnnými

Regrese patří k základním statickým metodám. Jejím cílem je najít regresní funkci, která se snaží vysvětlit vznik většího počtu pozorovaných náhodných veličin $Y_1, Y_2, \dots, Y_n$ pomocí známých vlivů x_{ij} a pomocí poměrně malého

počtu parametrů $\beta_0, \beta_1, \beta_2, \dots, \beta_k$. Za lineární regresi budeme považovat regresi, ve které je závislost na parametrech $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ lineární.

Na jednu vysvětlovanou proměnnou Y máme k dispozici k vysvětlujících proměnných. Tedy v tomto odstavci budeme pracovat s modelem

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k. \quad (19.3)$$

Pokud máme n pozorování, dostaneme pak n rovnic o $k + 1$ neznámých ve tvaru

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + e_i, \quad (19.4)$$

kde $i = 1, 2, \dots, n$ a e_i jsou náhodné chyby. Pro odvození všech testů a intervalových odhadů v průběhu celé této kapitoly také klademe na chyby e_i předpoklad, že mají normální rozdělení $N(0; \sigma^2)$. Lineární regrese má tedy celkem čtyři předpoklady: linearitu vztahu mezi $\mathbf{X}$ a Y , nezávislost pozorování, homoskedasticitu (shodu rozptylu σ^2 pro všechna pozorování) a normalitu chyb. Pro větší rozsahy výběru si můžeme dovolit předpoklad normality opomenout díky CLV. Ostatní předpoklady jsou ale pro modelování důležité a budeme je ověřovat v grafu reziduí. Pro odvození bodových odhadů tento předpoklad není nutný.

Maticový zápis tohoto modelu má tvar

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

kde

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & \dots & X_{1k} \\ 1 & X_{21} & \dots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \dots & X_{nk} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

Cílem lineární regrese je odhadnout parametry modelu $\beta_0, \beta_1, \beta_2, \dots, \beta_k$. Pro odhad těchto parametrů se nejčastěji používá *metoda nejmenších čtverců*, která spočívá v minimalizaci funkce L

$$L(\beta_0, \beta_1, \dots, \beta_k) = \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik}))^2. \quad (19.5)$$

Nutnou podmínkou pro existenci extrému je nulovost parciálních derivací. Vzhledem k tomu, že daná funkce je ve svém definičním oboru konvexní,

je to i postačující podmínka. Zderivujeme-li danou funkci podle všech proměnných a položíme-li parciální derivace rovné nule, dostaneme následující soustavu rovnic

$$\begin{aligned}\frac{\partial L(\beta_0, \beta_1, \dots, \beta_k)}{\partial \beta_0} &= -2 \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik})) = 0 \\ \frac{\partial L(\beta_0, \beta_1, \dots, \beta_k)}{\partial \beta_j} &= -2 \sum_{i=1}^n X_{ij} (Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik})) = 0,\end{aligned}$$

kde $j = 1, 2, \dots, k$.

Po menších úpravách obdržíme soustavu

$$\begin{aligned}n\beta_0 + \beta_1 \sum_{i=1}^n X_{i1} + \beta_2 \sum_{i=1}^n X_{i2} + \dots + \beta_k \sum_{i=1}^n X_{ik} &= \sum_{i=1}^n Y_i \\ \beta_0 \sum_{i=1}^n X_{i1} + \beta_1 \sum_{i=1}^n X_{i1}^2 + \beta_2 \sum_{i=1}^n X_{i2} X_{i1} + \dots + \beta_k \sum_{i=1}^n X_{ik} X_{i1} &= \sum_{i=1}^n Y_i X_{i1} \\ &\vdots \\ \beta_0 \sum_{i=1}^n X_{ik} + \beta_1 \sum_{i=1}^n X_{ik} X_{i1} + \beta_2 \sum_{i=1}^n X_{ik} X_{i2} + \dots + \beta_k \sum_{i=1}^n X_{ik}^2 &= \sum_{i=1}^n Y_i X_{ik}.\end{aligned}$$

Vyřešením této soustavy získáme odhady $b_0, b_1, \dots, b_k$ parametrů $\beta_0, \beta_1, \beta_2, \dots, \beta_k$. Výše uvedená soustava se nazývá *soustava normálních rovnic*. Maticový zápis soustavy normálních rovnic je

$$(\mathbf{X}^T \mathbf{X}) \cdot \boldsymbol{\beta} = \mathbf{X}^T \mathbf{Y}. \quad (19.6)$$

Je-li matice $(\mathbf{X}^T \mathbf{X})$ regulární, existuje k ní matice inverzní (označme ji $(\mathbf{X}^T \mathbf{X})^{-1}$), pak odhad parametrů $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_k)^T$ je

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}. \quad (19.7)$$

Minimum funkce L nazýváme *reziduální součet čtverců* a vypočteme jej

$$S_e = L(\mathbf{b}) = \sum (Y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \dots + b_k x_{ik}))^2 = \sum (Y_i - \hat{Y}_i)^2,$$

kde $\hat{Y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \dots + b_k x_{ik}$ je regresní odhad hodnoty Y_i . Odhad rozptylu chyb σ^2 je $S^2 = \frac{S_e}{n-k-1}$ a nazýváme jej *reziduální rozptyl*, jeho odmocninu pak jako směrodatná chyba regrese. *Celkový součet čtverců* $S_T = \sum (Y_i - \bar{Y})^2$ vyjadřuje celkovou variabilitu vysvětlované proměnné.

Vhodnost modelu posuzujeme koeficientem determinace

$$R^2 = 1 - \frac{S_e}{S_T} = \frac{S_T - S_e}{S_T},$$

který vyjadřuje, jaká část celkové chyby S_T je vysvětlena regresním modelem. (Chyba S_e obsahuje to, co regresní model nedokázal vysvětlit). Koeficient determinace můžeme také počítat podle vzorce

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2},$$

kde $\hat{Y}_i = \hat{f}(x_i) = b_0 + b_1x_{i1} + \dots + b_kx_{ik}$ je regresní odhad Y_i . Je zřejmé, že čím je R^2 blíže jedné, tím lépe regresní model vystihuje naměřená data.

Metodou nejmenších čtverců získáme bodové odhady regresních parametrů $\beta_0, \beta_1, \beta_2, \dots, \beta_k$. Někdy nás však zajímají i intervalové odhady o spolehlivosti $1 - \alpha$ konstruované pro parametry $\beta_0, \beta_1, \beta_2, \dots, \beta_k$. Intervalový odhad o spolehlivosti $1 - \alpha$ pro parametr β_i je interval

$$\left[b_i - t_{n-k-1}(\alpha) \cdot s \sqrt{(\mathbf{X}^T \mathbf{X})_{ii}^{-1}}; b_i + t_{n-k-1}(\alpha) \cdot S \sqrt{(\mathbf{X}^T \mathbf{X})_{ii}^{-1}} \right], \quad (19.8)$$

kde $(\mathbf{X}^T \mathbf{X})_{ii}^{-1}$ je prvek matice $(\mathbf{X}^T \mathbf{X})^{-1}$ nacházející se na i -tém řádku a i -tém sloupci.

Je zřejmé, že čím méně budeme mít vysvětlujících proměnných, tím bude model jednodušší. Proto se nejčastěji zabýváme otázkou, zda je možné model zjednodušit tak, aby hodnoty Y_i vůbec nezávisely na x_{ij} . Tudíž testujeme hypotézu

$$H_0: \beta_j = 0 \text{ proti } H_1: \beta_j \neq 0$$

Za platnosti H_0 má testová statistika

$$T = \frac{b_j}{S \cdot \sqrt{(\mathbf{X}^T \mathbf{X})_{jj}^{-1}}} \sim t_{n-k-1}. \quad (19.9)$$

Studentovo rozdělení t o $n - k - 1$ stupňích volnosti. Tudíž pokud

$$|t| \geq t_{n-k-1} \left(1 - \frac{\alpha}{2} \right),$$

zamítneme hypotézu H_0 na hladině významnosti α . Zamítneme-li hypotézu H_0 tohoto testu, pak jsme vlastně potvrdili lineární závislost Y_i na j -té vysvětlující proměnné, zastřenou náhodnými chybami e_i .

Někdy se ptáme, zda je možné model zjednodušit o více než jeden parametr. V takovém případě nepoužijeme dva předchozí testy, protože jejich společná hladina by nebyla α , ale použijeme následující test: testujeme hypotézu

$$H_0: \beta_{j_1} = \beta_{j_2} = \dots = \beta_{j_l} = 0, \quad 1 \leq j_1, \dots, j_l \leq k,$$

proti alternativě, že zjednodušený model neplatí (tj. alespoň jedno $\beta_{j_i} \neq 0$). Číslo l zde označuje počet parametrů, které se pokoušíme z modelu vypustit. Maticový zápis zjednodušeného modelu má tvar

$$\mathbf{Y} = \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} + \tilde{\mathbf{e}},$$

kde matice $\tilde{\mathbf{X}}$ vznikne z matice $\mathbf{X}$ vynecháním sloupců příslušejícím parametrům $\beta_{j_1}, \beta_{j_2}, \dots, \beta_{j_l}$, vektor $\tilde{\boldsymbol{\beta}}$ vznikne z vektoru $\boldsymbol{\beta}$ vynecháním parametrů $\beta_{j_1}, \beta_{j_2}, \dots, \beta_{j_l}$ a podobně vznikne i $\tilde{\mathbf{e}}$.

Parametry zjednodušeného modelu $\tilde{\boldsymbol{\beta}}$ odhadneme pomocí

$$\tilde{\mathbf{b}} = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \mathbf{Y}. \quad (19.10)$$

Poté spočteme reziduální součet čtverců pro zjednodušený model

$$\tilde{S}_e = \sum (Y_i - \hat{\tilde{Y}}_i)^2,$$

kde $\hat{\tilde{Y}}_i$ je regresní odhad Y_i ve zjednodušeném modelu. Je zřejmé, že $\tilde{S}_e \geq S_e$, neboť S_e je minimum funkce $L(\boldsymbol{\beta})$ bez jakýchkoli omezení na vektor $\boldsymbol{\beta}$, zatímco $\tilde{S}_e$ je minimum funkce $L(\boldsymbol{\beta})$ za podmínky $\beta_{j_1} = \beta_{j_2} = \dots = \beta_{j_l} = 0$.

Za platnosti H_0 má pak testová statistika F-rozdělení o l a $n - k - 1$ stupňů volnosti:

$$F = \frac{(\tilde{S}_e - S_e)/l}{S_e/(n - k - 1)} \sim F_{l, n-k-1}.$$

Pokud $f \geq F_{l, n-k-1}(1-\alpha)$, zamítneme hypotézu H_0 na hladině významnosti α , a tudíž model nemůžeme zjednodušit.

Příklad 19.2. V 60. letech proběhla ve Velké Británii následující studie a ve 30 hrabstvích byly naměřeny veličiny: A představující změnu populace za posledních deset let, B počet zaměstnanců v zemědělství, C velikost daní z nemovitostí, D procento obyvatel majících telefon, E procento obyvatel žijících na vesnici a F průměrný věk. Těmito veličinami měla být vysvětlena veličina Y , která udává procento obyvatel žijících pod hranicí chudoby.

Řešení. Na základě dat dostupných v online helpu programu Statistica⁵ dostáváme lineární regresní model

$$Y_i = \beta_0 + \beta_A A_i + \beta_B B_i + \beta_C C_i + \beta_D D_i + \beta_E E_i + \beta_F F_i + e_i.$$

⁵<https://docs.tibco.com/pub/stat/14.0.0/doc/html/UsersGuide/GUID-4F43645C-C174-48E2-99C5-45BD091A60B81.html>

Matice $\mathbf{X}$ bude obsahovat 7 sloupců, kde v prvním budou samé jedničky, ve druhém budou hodnoty veličiny A , ve třetím B atd. Nyní podle vzorce (19.7) spočteme odhad jednotlivých parametrů

$$\begin{aligned}\mathbf{b} &= (b_0; b_A; b_B; b_C; b_D; b_E; b_F)^T = \\ &= (31,26; -0,39; 0,000\,7; 1,23; -0,083; 0,16; -0,42)^T.\end{aligned}$$

Dále spočteme reziduální součet čtverců $S_e = 265,66$ a celkový součet čtverců $S_T = 1197,72$. Odtud dostáváme, že $R^2 = 0,78$.

Nyní nás zajímá, jestli některé proměnné můžeme z modelu vypustit, tj. testujeme $H_0: \beta_j = 0$. Za tímto účelem spočteme podle vzorce (19.9) pro každou proměnnou hodnotu statistiky T .

$$\begin{aligned}\mathbf{t} &= (t_0; t_A; t_B; t_C; t_D; t_E; t_F)^T = \\ &= (2,35; -4,87; 1,69; 0,38; -0,63; 2,67; -1,64)^T.\end{aligned}$$

Tyto hodnoty porovnáme s kvantilem $t_{23}(0,975) = 2,068$

$$= \text{T.INV}(0,975;23)$$

a vidíme, že hypotézu nulovosti parametrů zamítáme u $\beta_0, \beta_A, \beta_E$. Tyto testy nám říkají, že můžeme z modelu vypustit proměnnou B, C, D nebo F . Ovšem nevíme, zda můžeme vypustit všechny proměnné najednou. Na tuto otázku nám odpoví následující F -test.

Uvažujme tedy zjednodušený model (podmodel)

$$Y_i = \beta_0 + \beta_A A_i + \beta_E E_i + e_i.$$

Pro podmodel spočteme odhad jednotlivých parametrů

$$\tilde{\mathbf{b}} = (\tilde{b}_0, \tilde{b}_A, \tilde{b}_E)^T = (16,67; -0,40; 0,13)^T.$$

Dále spočteme reziduální součet čtverců podmodelu $\tilde{S}_e \doteq 393,03$. Odtud dostáváme, že $R^2 = 0,67$. Nyní sestojíme statistiku F a spočteme její hodnotu

$$f = \frac{(30 - 6 - 1)(\tilde{S}_e - S_e)}{3S_e} = 3,67,$$

vypočítáme kritickou hodnotu $F_{3,23}(0,05) \doteq 3,03$

$$= \text{F.INV}(0,95;3;23)$$

a porovnáme je. Jelikož $3,67 > 3,03$, H_0 zamítáme, neboli nemůžeme vypustit všechny čtyři proměnné zároveň.

Musíme tedy některou proměnnou do podmodelu přidat. Přidejme proměnnou B , protože $t_B > t_C, t_D, t_F$ (tj. proměnná B je v modelu významnější než C, D a F). Uvažujme tedy podmodel

$$Y_i = \beta_0 + \beta_A A_i + \beta_B B_i + \beta_E E_i + e_i.$$

Pro podmodel spočteme odhad jednotlivých parametrů

$$\tilde{\mathbf{b}} = (\tilde{b}_0, \tilde{b}_A, \tilde{b}_B, \tilde{b}_E)^T = (10,99; -0,40; 0,001; 0,19)^T.$$

Dále spočteme reziduální součet čtverců podmodelu $\tilde{S}_e \doteq 318,83$. Odtud dostáváme, že $R^2 \doteq 0,73$. Nyní sestrojíme statistiku F a spočteme její hodnotu

$$f = \frac{(30 - 6 - 1)(\tilde{S}_e - S_e)}{4S_e} = 1,15.$$

Vypočítáme kritickou hodnotu $F_{4,23}(0,05) \doteq 2,80$.

$$= \text{F.INV}(0,95;4;23)$$

Z nerovnosti $1,15 \not\geq 2,80$ plyne, že H_0 nezamítáme, neboli z původního modelu můžeme vypustit proměnné C, D a F . Popíšeme tedy veličinu Y proměnnými A, B a E . $\triangle$

Doplnit do textu, ve kterém bude řešení příkladu v MS Excel: hodnota statistiky F-test v ANOVA slouží ke zjištění statistické významnosti těchto koeficientů současně.

19.3 Polynomiální regrese, kvadratická regrese

Nejjednodušším případem polynomiální regrese je tzv. kvadratická regrese. Pod pojmem kvadratická regrese míníme model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i, \quad i = 1, 2, \dots, n,$$

kde $e_i \sim N(0; \sigma^2)$, $n \geq 4$. Zde veličiny Y_i závisí kvadraticky na veličinách X_i .

Položíme-li $Z_i = X_i^2$, $i = 1, 2, \dots, n$, dostáváme model

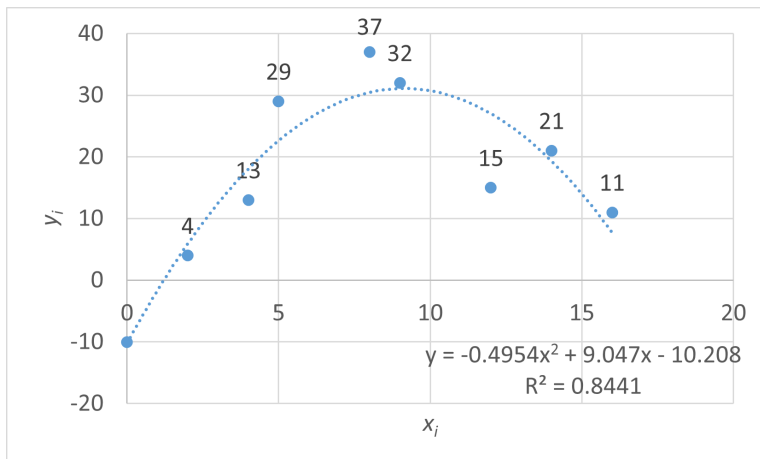
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + e_i, \quad i = 1, 2, \dots, n.$$

V tomto modelu závisí náhodná veličina Y_i lineárně na veličinách X_i a Z_i . Neboli úloha kvadratické regrese byla převedena na úlohu lineární regrese se dvěma vysvětlujícími proměnnými.

Příklad 19.3.  V tabulce 19.2 jsou data o zisku firmy v závislosti na vynaložených nákladech na reklamu.

Tabulka 19.2: Data o vynaložených nákladech v mil. Kč na reklamu a docíleného zisku

Výdaje v mil. Kč	0	2	4	5	8	9	12	14	16
Zisk v mil. Kč	-10	4	13	29	37	32	15	21	11



Obrázek 19.7: Bodový graf s křivkou kvadratické regrese a zobrazenou rovnicí trendu a koeficientem determinace

b_1	b_2	b_0
9.046994	-0.495438	-10.20839192

Obrázek 19.8: Výřez z MS Excel s využitím funkce `= LINREGRESE`

- Najděte a pomocí MS Excel znázorněte model vyjadřující závislost zisku firmy v milionech Kč na vynaložených finančních nákladech na reklamu v milionech Kč. Vhodnost zvoleného regresního modelu ověřte pomocí koeficientu determinace a testu ANOVA.
- Odhadněte zisk, pokud náklady na reklamu budou 6 milionů Kč.
- Při jakých nákladech na reklamu bude firma dosahovat maximálního zisk?

Řešení. Bodový graf na obrázku 19.7 vytvořený v MS Excel graficky znázorňuje závislost zisku firmy na výši finančních nákladů na reklamu. Regresním modelem této závislosti je část paraboly s přibližným analytickým vyjádřením $y = -0,5 \cdot x^2 + 9x - 10$ (po zaokrouhlení koeficientů na celá čísla).

Koeficienty v rovnici regresní kvadratické funkce lze v MS Excel nalézt pomocí funkce `= LINREGRESE`, v níž je prvním parametrem matice

ANOVA		
		Rozdíl
	Regrese	2
	Rezidua	6
	Celkem	8
Koefficienty		
b_0	Hranice	-10.2083919
b_1	Soubor X 1	9.046993641
b_2	Soubor X 2	-0.49543769

Obrázek 19.9: Výřez z MS Excel s tabulkou ANOVA nástroje Analýza dat části Regrese

VÝSLEDEK	
Regresní statistika	
Násobné R	0.918742955
Hodnota spolehlivosti R	0.844088617
Nastavená hodnota spolehlivosti R	0.792118156
Chyba stř. hodnoty	6.722010763
Pozorování	9

Obrázek 19.10: Výřez z MS Excel s tabulkou Regresní statistika nástroje Analýza dat část Regrese

dat y_i , druhým parametrem je matice obsahující data ze sloupců x_i^2 a x_i v tomto pořadí. Třetím a čtvrtý parametr je vždy PRAVDA: třetí specifikuje model s absolutním členem a čtvrtý výstup rozšířený o další charakteristiky odhadů. Funkce je maticová a potvrdí se současným stiskem kláves Ctrl+Shift+Enter.

= LINREGRESE(matice1;matice2;PRAVDA;PRAVDA).

Výsledek použití této maticové funkce ukazuje obrázek 19.8.

Koeficienty rovnice kvadratické regrese lze získat také z nástroje Analýza dat v části Regrese. Pro náš konkrétní případ jsou koeficienty zvýrazněny na obrázku 19.9. Koeficient determinace je $R^2 \doteq 0,844$. Informaci o něm lze získat buď zobrazením přímo v grafu MS Excel nebo v tabulce Regrese nástroje Analýza dat (viz obrázek 19.10). Koeficient determinace R^2 stojí na druhém místě shora.

MS	F	Významnost F
733.888158	16.24169958	0.00378995
45.1854287		

Obrázek 19.11: Výřez z MS Excel s tabulkou ANOVA nástroje Analýza dat části Regrese

Hodnota P
0.110407665
0.001364245
0.002131367

Obrázek 19.12: Výřez z MS Excel v tabulce Regrese nástroje Analýza dat v části ANOVA

Ověření vhodnosti zvoleného polynomickeho modelu otestujeme pomocí p -hodnoty testu ANOVA, která je součástí tabulky Regrese z Analýzy dat (viz obrázek 19.11). Testujeme hypotézy:

H_0 : regresní model není statisticky významný.

proti alternativě

H_1 : regresní model je statisticky významný.

Protože $p = 0,0038 < 0,05$, zamítáme H_0 o statistické nevýznamnosti zvoleného regresního modelu jako celku.

Testovat je možné také významnost jednotlivých koeficientů kvadratické regresní funkce. K tomuto účelu slouží p -hodnoty t -testů ve sloupci s názvem Hodnota P (viz obrázek 19.12) nástroje Regrese z Analýzy dat. První hodnota shora vypovídá o statistické významnosti koeficientu b_0 , zde je $p = 0,11 > 0,05$, tedy nezamítáme nulovou hypotézu o tom, že tento koeficient je statisticky nevýznamný. Druhá a třetí hodnota vypovídají o statistické významnosti koeficientů b_1 a b_2 , přitom oba jsou v tomto příkladu menší než hladina významnosti $\alpha = 0,05$, tedy b_1 i b_2 jsou pro vytvoření kvadratického regresního modelu statisticky významné.

Odhad zisku při nákladech 6 milionů Kč zjistíme dosazením do rovnice regresní funkce a dostaneme $y(6) = 26,2$ Kč. Pro určení hodnoty nákladů pro maximální zisk lze využít například diferenciálního počtu a dostaneme přibližně 9,13 milionů Kč. $\triangle$

Podobně jako v případě kvadratické regrese budeme postupovat i u regrese vyšších stupňů. Ještě je potřeba odpovědět na otázku volby stupně

regresního polynomu. Uvažujme model

$$Y_i = \beta_0 + \beta_1 X_i + \cdots + \beta_p X_i^p + e_i, \quad i = 1, 2, \dots, n,$$

kde stupeň p regresního polynomu není znám. Počet parametrů tohoto modelu označme $k = p + 1$.

Uvažujme, že skutečný stupeň polynomu je p_0 a tedy skutečný počet parametrů modelu je $k_0 = p_0 + 1$. Označme S_k^2 reziduální rozptyl modelu s k parametry (reziduální rozptyl je definován na str. 372). Dá se ukázat, že

$$\mathbb{E}S_k^2 > \sigma^2 \text{ pro } k < k_0,$$

a

$$\mathbb{E}S_k^2 = \sigma^2 \text{ pro } k \geq k_0.$$

Je tudíž potřeba najít bod, kde se posloupnost s_k^2 mění z klesající posloupnosti na oscilující. Toto je obtížná úloha, proto ji převedeme na úlohu hledání minima posloupnosti. Vytvoříme posloupnost

$$A_k = S_k^2 \left(1 + \frac{k}{\sqrt[n]{n}} \right),$$

která větším k přidává větší váhu. Hodnotu k , pro kterou je A_k minimální, pak vezmeme jako odhad skutečného počtu parametrů k_0 .

19.4 Nelineární regrese

Uvažujme nelineární regresní model

$$Y_i = f(X_i, \beta) + e_i, \quad i = 1, 2, \dots, n,$$

kde f je regresní funkce a β je vektor neznámých parametrů. Odhad parametrů β metodou nejmenších čtverců dostaneme minimalizací výrazu

$$g(\beta) = \sum_{i=1}^n (Y_i - f(X_i, \beta))^2.$$

Tuto úlohu iteračně řeší různé statistické a matematické programy. Tyto programy ovšem vyžadují počáteční aproximaci vektoru $\mathbf{b}$ (odhad parametru β).

Počáteční aproximaci můžeme snadno získat u tzv. linearizovatelných modelů, tj. modelů, které se dají převést na lineární model. Jako příklad uveďme model, jehož regresní funkce je exponenciální:

$$Y_i = \beta_0 e^{\beta_1 X_i} + e_i, \quad i = 1, 2, \dots, n.$$

Tabulka 19.3: Data o stáří strojů (v letech) a nákladech na jejich údržbu (ve stovkách Kč)

Stáří stroje (x_i)	2	1,2	14,8	8,3	8,4	3
Náklady (y_i)	12	8	76,4	17	21,3	10
Stáří stroje (x_i)	4,8	15,6	16,1	11,5	14,2	14
Náklady (y_i)	12,5	97,3	88	25	38,6	47,3

Při počáteční aproximaci si můžeme dovolit zapomenout na chyby e_i a model zlogaritmovat

$$\ln Y_i = \ln \beta_0 + \beta_1 X_i, \quad i = 1, 2, \dots, n.$$

Zavedeme-li nové parametry $\alpha_0 = \ln \beta_0$ a $\alpha_1 = \beta_1$, dostaneme lineární regresní model

$$\ln Y_i = \alpha_0 + \alpha_1 X_i, \quad i = 1, 2, \dots, n,$$

který vyřešíme podle kapitoly 19.2 a dostaneme odhady a_0, a_1 . Za počáteční aproximaci odhadu parametrů původního modelu pak vezmeme odhady

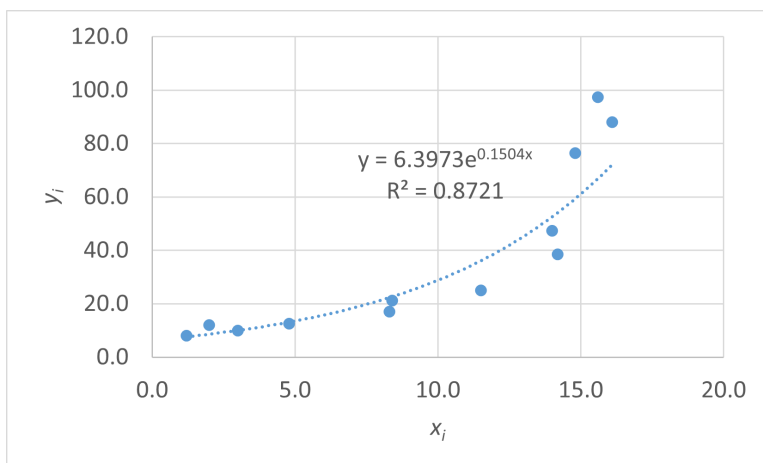
$$b_0 = e^{a_0}, \quad b_1 = a_1.$$

Na závěr uvedme příklady některých linearizovatelných modelů:

- 1) $Y = e^{\beta_0 + \beta_1 X}$,
- 2) $Y = \beta_0 X^{\beta_1}$,
- 3) $Y = \ln(\beta_0 + \beta_1 X)$,
- 4) $Y = \frac{1}{\beta_0 + \beta_1 X}$.

Příklad 19.4.  Tabulka 19.3 obsahuje data o výši nákladů y na údržbu (ve stovkách Kč) a údaje o stáří stroje (v letech). Užitím vhodného regresního modelu popište závislost nákladů na údržbu na stáří stroje.

Řešení. V MS Excel vytvoříme bodový graf (viz obrázek 19.13). Z rozmístění bodů grafu lze odhadnout exponenciální regresní model. Do grafu si znázorníme exponenciální regresní křivku, její rovnici a koeficient determinace. Rovnice regresní křivky je $y = 6,4 \cdot e^{0,15 \cdot x}$ po zaokrouhlení na setiny. Koeficienty b_1, b_2 regresní funkce $y = b_1 \cdot e^{b_2 x}$ lze v MS Excel nalézt také pomocí funkce  = **LINREGRESE**. K tomu potřebujeme sloupec s hodnotami $\ln y$ a úpravu rovnice $y = b_1 \cdot e^{b_2 x}$ do tvaru $\ln y = \ln b_1 + b_2 x$. Užití maticové funkce  = **LINREGRESE** ukazuje obrázek 19.14, kde první dvě pole jsou výsledky maticové funkce a poslední pole b_1 vypočítáme jako e umocněné na vygenerovanou hodnotu $\ln b_1$.



Obrázek 19.13: Bodový graf závislosti nákladů na údržbu stroje na jeho stáří

LINREGRESE		
b_2	ln(b_1)	b_1
0.1504281	1.855881	6.3973327

Obrázek 19.14: Výřez z MS Excel s využitím funkce LINREGRESE

	Koeficienty
Hranice	1.855881132
Soubor X 1	0.150428145

Obrázek 19.15: Výřez z MS Excel z tabulky Regrese nástroje Analýza dat

MS	F	Významnost F
7.859582041	113.0815083	9.02369E-07
0.06950369		

Obrázek 19.16: Výřez z MS Excel v části ANOVA z tabulky Regrese nástroje Analýza dat

Oba koeficienty b_2 i $\ln(b_1)$ lze vygenerovat v tabulce Regrese z nástroje Analýza dat (viz obrázek 19.15). Koeficient determinace má podle téže tabulky hodnotu $R^2 \doteq 0,872$.

Při testování statistické významnosti zvoleného exponenciálního regresního modelu opět vyjdeme z informací části ANOVA v tabulce Regrese nástroje Analýza dat (viz obrázek 19.16). Zjišťujeme, že p -hodnota významnosti zvoleného modelu je $p \doteq 9,02 \cdot 10^{-7} < 0,05$, tedy zamítáme nulovou hypotézu, podle které je zvolený model statisticky nevýznamný. $\triangle$

Poznámka 19.1. *Jelikož jsme předpokládali aditivní model, tj. residua jsou k regresní funkci přičítána, je postup při linearizaci jen aproximací přesného výpočtu. Často ovšem předpokládáme multiplikativní model, tj. residua jsou s regresní funkcí násobena. V tomto případě převádí linearizace multiplikativní exponenciální model přímo na lineární regresi a výpočet s pomocí linearizace je přesný. Multiplikativní modely se často využívají ve finančních analýzách, kde výkyvy jsou úměrné hodnotě podkladového aktiva.*

19.5 Úlohy

19.1 Tabulka 19.4 ukazuje párová data výšky a hmotnosti 20 osob. Určete rovnici regresní přímky.

Tabulka 19.4: Data o výšce v v cm a hmotnosti m osob v kg

v	175	175	184	180	173	173	184	179	168	183	178
m	75	73	74	82	77	70	88	68	60	82	79
v	185	180	175	173	180	185	170	178	166	168	160
m	86	64	70	64	91	84	81	72	67	66	45

19.2 Zjišťovalo se, kolik mg kyseliny mléčné je v 100 ml krve u matek prvorodiček (hodnoty x_i) a u jejich novorozenců (hodnoty y_i) těsně po porodu. 19.5 ukazuje získané hodnoty.

Tabulka 19.5: Data o obsahu kyseliny mléčné v krvi matek

x_i	40	64	34	15	57	45	27	35	51
y_i	33	46	23	12	56	40	18	32	47

- Popište závislost y_i na x_i regresním modelem. Ověřte vhodnost modelu pomocí koeficientu determinace.
- Zjistěte, jaká bude hodnota kyseliny mléčné v 100 ml krve novorozence, jestliže u matky byla 60 mg.
- O kolik mg vzroste kyselina mléčná v 100 ml krve u novorozence, jestliže u matky vzrostla o 5 mg?

19.3 V tabulce 19.6 jsou uvedeny údaje o hodnotě produkce v 100 000 Kč (proměnná y) a o výši investic v 10 000 Kč (proměnná x) v roce 1998 v souboru 12 vybraných soukromých strojírenských firem s počtem zaměstnanců větším než 24.⁶

Tabulka 19.6: Data o investicích a produkci

Firma	1	2	3	4	5	6
y_i	52,8	48,4	54,2	50	54,9	53,9
x_i	16,3	16,8	18,5	16,3	17,9	17,4
y_i	53,1	52,4	53	52,9	53,1	60,1
x_i	16,1	16,2	17	16,7	17,5	19,1

⁶Úloha včetně dat a řešení převzata z [9], s. 188.

- Stanovte rovnici regresní přímky modelující závislost hodnoty produkce na výši investic.
- Zjistěte, o kolik v průměru vzroste hodnota produkce, jestliže výše investice vzroste o 10 000 Kč?

19.4 Bylo sledováno, jaký vliv na produkci rajčat má výška rostliny. Výsledky jsou v tabulce: 19.7 ukazuje získané hodnoty.

Tabulka 19.7: Data o výšce a produkci rajčat

Výška v mm	210	330	180	260	280	220	270	260	290	310
Produkce v g	270	420	260	350	390	310	380	330	360	410

- Určete regresní funkci, která modeluje závislost produkce na výšce rostliny. Ověřte kvalitu modelu (pomocí koeficientu determinace a testu ANOVA).
- Jakou produkci můžeme očekávat u rostliny, která má výšku 320 mm?
- Jakou změnu produkce můžeme v průměru očekávat, dojde-li ke zvýšení rostliny o 5 cm?

19.5 Byl sledován vztah mezi výrobními náklady v tisících Kč na hektar (c), který byl osázen pšenicí a hektarovým výnosem pšenice (v v tunách). Výzkumu se zúčastnilo 11 soukromých zemědělců. Výsledky jsou v tabulce 19.8.

Tabulka 19.8: Data o výrobních nákladech a výnosu

c	29,5	40,5	23,8	34	46,1	41,5	31,9	38,1	31	27	33
v	3,5	4	2,9	3,7	4,5	4,3	3,5	3,9	3,9	3,2	3,9

- Najděte model vyjadřující závislost hektarového výnosu pšenice na výrobních nákladech na hektar. Vhodnost modelu ověřte pomocí koeficientu determinace a testu ANOVA.
- Zjistěte, o kolik vzroste hektarový výnos, jestliže se výrobní náklady zvýší o 5 tisíc?

19.6 Byl vyvinut nový druh insulinu a zkoumá se závislost snížení hladiny cukru v krvi pacienta na množství podaného insulinu určitou dobu před měřením. Osmi náhodně vybraným pacientům byla naočkováána různá množství insulinu (n v ml) a po určité době bylo těmto pacientům změřeno snížení cukru v krvi (s v %). Výsledky měření ukazuje tabulka 19.9.

Tabulka 19.9: Data o množství podaného insulinu a snížení cukru v krvi

n	150	200	250	300	350	400	450	500
s	8	12	30	20	55	58	44	65

- Znázorněte korelační pole a zvolte vhodný typ lineárního regresního modelu pro popis závislosti snížení hladiny cukru na množství podaného inzulínu.
- Ověřte oprávněnost použití vybraného modelu (pomocí koeficientu determinace, pomocí ANOVA).
- Odhadněte, o kolik procent se sníží hladina cukru pacienta, jemuž se podá 325 ml insulínu?

19.7 Trabant, ikonický symbol automobilové historie východního Německa, byl dlouho spojen s jednoduchostí a praktičností socialistického inženýrství. U tohoto automobilu byla v 80. letech minulého století měřena spotřeba paliva y_i (v litrech na 100 km) v závislosti na jeho rychlosti x_i (km/h). Zkoušky probíhaly na rovině za stejných podmínek. Vůz jel stále se zařazeným 4. rychlostním stupněm. Tabulka 19.10 ukazuje naměřené výsledky.

Tabulka 19.10: Data o rychlosti a spotřebě automobilu

x_i	40	50	60	70	80	90	100
y_i	6,1	5,8	6	6,5	6,8	8,1	10

- Sestavte regresní model vyjadřující závislost spotřeby paliva na rychlosti a ověřte jeho vhodnost (kvalitu modelu).
- Odhadněte spotřebu, pokud automobil pojede rychlostí 85 km/h.
- Při jaké rychlosti bude mít automobil minimální spotřebu?

19.8 Prodejna potravinových doplňků provedla 22 měření v náhodně vybraných dnech, v nichž sledovala, kolik osob navštíví během prodejní doby prodejnu (x), a jaká je v tento den tržba (y v tisících Kč a zaokrouhleno). Současně se sleduje, zda při rostoucím počtu zákazníků je stávající počet prodávačů dostatečný. Tabulka 19.11 ukazuje naměřené výsledky.

Tabulka 19.11: Data o tržbě a počtu zákazníků

x_i	18	19	24	25	26	27	28	29	30	32	33
y_i	11	12	13	13	14	15	16	17	18	19	19
x_i	35	36	37	40	42	46	47	49	52	55	58
y_i	20	20	21	22	21	21	20	19	19	18	17

- Najděte regresní funkci, která vystihuje závislost tržby na počtu zákazníků. Určete vhodnost modelu pomocí koeficientu determinace.
- Odhadněte pomocí nalezené regresní funkce denní tržbu, jestliže prodejnu navštíví 23, resp. 38 zákazníků.

- c) Stanovte optimální počet zákazníků, kterému je obsluha schopna se věnovat tak, aby nezačalo s rostoucím počtem zákazníků docházet k poklesu tržby.

Řešení úloh

19.1 $y = 0,4743x + 141,12$

19.2

- a) $y = 0,8974x - 2,5814$, $R^2 = 0,8811$,
b) 51,26 mg,
c) o 4,49 mg.

19.3

- a) $y = 16,1509 + 2,1622x$,
b) o 216 220 Kč.

19.4

- a) $y = 1,1467x + 48,709$, $R^2 = 0,9204$, ANOVA test: $F(1,8) = 92,415$,
 $p = 0,000011$,
b) 415,65 g,
c) 57,335 g.

19.5

- a) $y = 0,0646x + 1,544$, $R^2 = 0,8724$, ANOVA test: $F(1,9) = 61,55$,
 $p = 0,000026$,
b) o 323 kg.

19.6

- a) $y = -15,96 + 0,16x$,
b) $R^2 = 0,805$, ANOVA test: $F(1,6) = 24,73$, $p = 0,0025$,
c) o 36,5 %.

19.7

- a) $y = 0,0019x^2 - 0,2073x + 11,393$, $R^2 = 0,9849$,
b) 7,5 l,
c) 54,5 km/h.

19.8

- a) $y = -0,0173x^2 + 1,498x - 11,736$, $R^2 = 0,9122$,
- b) asi 13 566 Kč, asi 20 207 Kč,
- c) asi 43 zákazníků.

Kapitola 20

Testy dobré shody

V předchozích kapitolách jsme se seznámili s některými testy, přičemž jsme mohli pozorovat, že tyto testy jsou vázány na předpoklad, že rozdělení základního souboru, z něhož byl výběrový soubor pořízen, je určitého typu. V této kapitole se tudíž budeme zabývat testy, které nám odhalí, zda náhodný výběr má konkrétní rozdělení nebo nikoli.

20.1 Pearsonův χ^2 test

Pearsonův χ^2 test je univerzálním nástrojem k ověřování skutečnosti, zda náhodný výběr pochází z nějakého diskrétního či spojitého rozdělení, mluvíme o tzv. χ^2 -testu dobré shody.

Mějme náhodný výběr $X_1, \dots, X_n$, kde veličiny X_j , $j = 1, \dots, n$, mohou nabývat hodnot $1, \dots, k$. Veličinám X_i označujícím počet výskytů výsledku i se říká empirické četnosti. Náhodný vektor $(X_1, \dots, X_k)^T$ má multinomické rozdělení. Budeme testovat hypotézu H_0 , že skutečné hodnoty pravděpodobností multinomického rozdělení jsou právě rovny číslům $p_1, \dots, p_k$. Veličinám np_i budeme říkat teoretické četnosti. Za platnosti hypotézy H_0 má statistika

$$\chi^2 = \sum_{i=1}^k \frac{(X_i - np_i)^2}{np_i} \sim \chi_{k-1}^2$$

asymptoticky rozdělení χ^2 o $k - 1$ stupních volnosti. Jakmile dostaneme

$$\chi^2 \geq \chi_{k-1}^2(1 - \alpha),$$

zamítneme hypotézu H_0 na hladině α .

Tabulka 20.1: Výsledky testování generátoru náhodných čísel [1]

Cifra	0	1	2	3	4	
x_i	59 889	59 796	59 969	60 056	60 303	
p_i	0,1	0,1	0,1	0,1	0,1	
np_i	60 000	60 000	60 000	60 000	60 000	
$\frac{(x_i - np_i)^2}{np_i}$	0,205 35	0,693 6	0,016 0	0,052 3	1,530 15	
Cifra	5	6	7	8	9	Celkem
x_i	60 048	60 234	59 750	60 224	59 731	600 000
p_i	0,1	0,1	0,1	0,1	0,1	1
np_i	60 000	60 000	60 000	60 000	60 000	600 000
$\frac{(x_i - np_i)^2}{np_i}$	0,0384	0,912 6	1,041 67	0,836 3	1,206 1	6,532 33

Je třeba mít na zřeteli, že test χ^2 je asymptotický, a proto ho lze doporučit jen při dostatečně velkém rozsahu výběru n . V literatuře se obvykle uvádí, že musí platit

$$np_i \geq 5q \quad \text{pro libovolné } i = 1, \dots, k, \quad k \geq 3,$$

kde q je podíl tříd, pro něž platí $np_i < 5$.

Tento test se může používat například při ověřování spravedlivosti hrací kostky, při kontrole generátorů náhodných čísel (každá cifra 0, 1, ..., 9 by se měla objevovat s pravděpodobností $\frac{1}{10}$ – viz následující příklad 20.1) a v řadě dalších případů.

Příklad 20.1.  Při testování spravedlivosti generátoru náhodných čísel byla zkoumána řada šesticiferných náhodných čísel o délce 100 000. Tudíž počet všech cifer je $n = 600\,000$. Zjištěné počty výskytů jednotlivých cifer náhodných čísel uvádí tabulka 20.1. Testováním na hladině významnosti 0,05 ověřte, zda zjištěné empirické pravděpodobnosti výskytu jednotlivých cifer odpovídají teoretickým pravděpodobnostem házení spravedlivou kostkou.

Řešení. Testujeme nulovou hypotézu H_0 , podle které se empirické pravděpodobnosti rovnají teoretickým pravděpodobnostem. Z tabulky 20.1 vyplývá, že hodnota testové statistiky je $\chi^2 = 6,532\,33$, kterou porovnáme s kritickou hodnotou $\chi^2_9(0,05) = 16,92$.

= CHISQ.INV(0,95;9)

Na základě zjištěných dat nelze zamítnout nulovou hypotézu, že generátor je náhodný, tedy hrací kostka je se spolehlivostí 95 % spravedlivá. $\triangle$

Pro řešení příkladu 20.1 a podobných příkladů je vhodné využít funkci

= CHISQ.TEST v MS Excel.

= CHISQ.TEST(aktuální;očekávané) $\doteq 0,68568$. Hodnota testového kritéria je větší než kritická hodnota pro hladinu významnosti 0,05, proto nezamítáme H_0 , tedy že se se spolehlivostí 95 % rovnají očekávané a teoretické pravděpodobnosti výskytu jednotlivých cifer.

20.2 Test normality

Ve statistice se testy normality používají k ověření, zda je soubor dat dobře modelován normálním rozdělením, a k výpočtu toho, jak je pravděpodobné, že náhodná proměnná, která je základem souboru dat, bude normálně rozdělena.

Nechť $Z_1, \dots, Z_n$ je náhodný výběr. Chceme testovat hypotézu H_0 , že jde o výběr z $N(\mu; \sigma^2)$, kde parametry μ a σ^2 nejsou známy. Nejprve vytvoříme třídy

$$(-\infty, b_1), [b_1, b_2), [b_2, b_3), \dots, [b_{k-2}, b_{k-1}), [b_{k-1}, \infty),$$

kde $k \geq 4$. Pro stručnost označme i -tou třídu symbolem J_i . Empirické četnosti jednotlivých tříd opět označíme X_i . Pravděpodobnost p_i , že daná veličina Z_j , $j = 1, \dots, n$ padne do J_i , je rovna

$$p_i = p_i(\mu; \sigma) = \int_{J_i} f(x) dx = \int_{J_i} \frac{1}{\sqrt{2\pi}\sigma} e^{\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]} dx.$$

Kdybychom znali parametry μ a σ^2 , pak by úloha byla převedena na případ multinomického rozdělení se známými parametry $p_1, \dots, p_k$. Toho by se dalo využít například při testu, zda náhodný výběr má rozdělení $N(0; 1)$.

V obecném případě ovšem parametry neznáme, tudíž tento postup není vhodný. V tomto případě je tedy nutné najít vhodné odhady μ a σ^2 . Není vhodné zvolit za μ a σ^2 klasické odhady, tedy průměr a výběrový rozptyl, protože by se tím podstatně změnilo rozdělení statistiky χ^2 . Musíme tedy nalézt odhady μ a σ^2 iteračně tak, aby odhady splňovaly soustavu rovnic

$$\mu = \frac{1}{n} \sum_{i=1}^k \frac{X_i}{p_i} \int_{J_i} x f(x) dx, \quad \sigma^2 = \frac{1}{n} \sum_{i=1}^k \frac{X_i}{p_i} \int_{J_i} (x - \mu)^2 f(x) dx,$$

kde jako počáteční aproximaci pro μ a σ^2 zvolíme průměr a výběrový rozptyl. Nezapomeňme, že na μ i σ závisejí p_i i $f(x)$, které jsou na pravých stranách těchto rovnic. Řešení soustavy označme $\hat{\mu}$ a $\hat{\sigma}$. Tyto odhady použijeme pro výpočet pravděpodobností p_i . Statistika

$$\chi^2 = \sum_{i=1}^k \frac{[X_i - np_i(\hat{\mu}, \hat{\sigma})]^2}{np_i(\hat{\mu}, \hat{\sigma})} \sim \chi_{k-3}^2$$

má pak rozdělení χ^2 o $k-3$ stupních volnosti. Pokud vyjde $\chi^2 > \chi_{k-3}^2(1-\alpha)$, zamítáme hypotézu H_0 na hladině α .

Příklad 20.2.  Bylo naměřeno padesát hodnot určité veličiny, jejichž četnosti ukazuje tabulka 20.2. Pomocí χ^2 -testu dobré shody na hladině významnosti 0,05 ověřte, zda je možné považovat tento náhodný výběr za výběr z populačního souboru s normálním rozdělením.

Tabulka 20.2: Hodnoty veličiny a jejich četnosti X_i

Hodnota veličiny	1	2	3	4	5	6	7
Četnost x_i	2	6	8	15	7	7	5

Řešení. Měřená veličina je spojitá náhodná veličina X . Odhadem μ je výběrový aritmetický průměr $\bar{x} \doteq 4,2$, odhadem σ^2 je výběrový rozptyl $s^2 \doteq 2,61$, tj. $s \doteq 1,62$. Testujeme

H_0 : Náhodná veličina X má normální rozdělení $X \sim N(4,2; 1,62)$

proti alternativě

H_1 : Náhodná veličina X nemá normální rozdělení $X \sim N(4,2; 1,62)$.

Výběr dat rozdělíme do sedmi intervalů. Výpočet testové statistiky ukazuje tabulka 20.3.

Tabulka 20.3: Hodnoty veličiny a jejich četnosti

b_{j-1}	b_j	n_j	p_j	np_j	np_j	n_j	$\frac{(n_j - np_j)^2}{np_j}$
$-\infty$	1	2	0,023 8	1,192 9	-	-	-
1	1,857	0	0,049 7	2,486 0	-	-	-
1,857	2,714	6	0,105 4	5,268 1	8,947 0	8	0,100 2
2,714	3,571	8	0,169 6	8,481 7	8,481 7	8	0,027 4
3,571	4,428	15	0,207 5	10,375 9	10,375 95	15	2,060 8
4,428	5,285	7	0,192 9	9,644 8	9,644 8	7	0,725 3
5,285	6,142	7	0,136 2	6,812 1	12,550 6	12	0,024 2
6,142	6,999	0	0,073 1	3,655 7	-	-	-
6,999	∞	5	0,041 7	2,082 3	-	-	-
Celkem		50	1	50	50	50	2,937 8

Protože vyšly čtyři teoretické četnosti menší než 5, snížili jsme počet intervalů na 5. Hodnota testového kritéria je po zaokrouhlení 2,937 8. Odhadovali jsme dva parametry (střední hodnota a rozptyl), tedy výsledný počet stupňů volnosti daného χ^2 -testu je $\nu = 5 - 2 - 1 = 2$. Pro zvolenou hladinu významnosti 0,05 je kritická hodnota $\chi_{0,05}^2(2) \doteq 5,99$.

= CHISQ.INV(0,95;2)

Protože $\chi^2 < \chi_{0,95}^2(5)$, nezamítáme H_0 a normální rozdělení $N(4,2; 1,62)$ je vhodným modelem pro popis náhodné veličiny X .

Funkce = CHISQ.TEST(aktuální; očekávané) V MS Excel poslouží v řešené úloze k určení p -hodnoty testu. Zde dostaneme $p \doteq 0,09 > 0,05$, tedy opět nezamítáme H_0 .

△

20.3 Test Poissonova rozdělení

Nechť $X_1, \dots, X_n$ je náhodný výběr z nějakého rozdělení na množině nezáporných celých čísel. Budeme testovat hypotézu H_0 , že jde o výběr z Poissonova rozdělení $Po(\lambda)$, neboli že pozorování $x_1, x_2, \dots$ pochází ze souboru dat, která mají Poissonovo rozdělení, kde parametr λ není znám. Testujeme proti alternativní hypotéze H_1 , podle které výběrová data $x_1, x_2, \dots$ nepocházejí z dat s Poissonovým rozdělením. Test provedeme obdobně jako u testu normality. Nejprve vytvoříme třídy. Jedna z možností je tato: do první třídy se zařadí veličiny, které jsou menší nebo rovny něja-

Tabulka 20.4: Počty volání

k	0	1	2	3	4	5	6	7	8	9	10	>10
n_k	1	6	8	10	6	6	7	1	1	2	2	0

kému číslu r . Další třídy jsou postupně tvořeny samostatnými hodnotami $r+1, r+2, \dots, r+k-2$. Poslední třída obsahuje hodnoty větší nebo rovné číslu $r+k-1$. Tím je vytvořeno k tříd, kde $k \geq 3$ a jejichž četnosti označíme $Y_r, Y_{r+1}, \dots, Y_{r+k-1}$. Označme

$$q_i = \mathbb{P}[X_j = i] = \frac{\lambda^i e^{-\lambda}}{i!}, \quad i = 0, 1, 2, \dots$$

Pak pravděpodobnosti jednotlivých tříd jsou

$$\begin{aligned} p_r &= \sum_{i=0}^r q_i, \\ p_i &= q_i \quad \text{pro } i = r+1, \dots, r+k-2, \\ p_{r+k-1} &= \sum_{i=r+k-1}^{\infty} q_i. \end{aligned}$$

Pravděpodobnosti p_i opět závisí na parametru λ , který neznáme a který musíme odhadnout. Podobně jako u testu normality vyřešíme iteračně rovnici

$$\lambda = \frac{1}{n} \left[Y_r \frac{\sum_{i=0}^r i q_i}{\sum_{i=0}^r q_i} + \sum_{i=r+1}^{r+k-2} i Y_i + Y_{r+k-1} \frac{\sum_{i=r+k-1}^{\infty} i q_i}{\sum_{i=r+k-1}^{\infty} q_i} \right],$$

kdy za počáteční aproximaci λ zvolíme průměr hodnot x_j , $j = 1, \dots, n$.

Řešení rovnice označíme $\hat{\lambda}$. Statistika

$$\chi^2 = \sum_{i=r}^{r+k-1} \frac{[Y_i - np_i(\hat{\lambda})]^2}{np_i(\hat{\lambda})} \sim \chi_{k-2}^2 \quad (20.1)$$

má pak rozdělení χ^2 o $k-2$ stupních volnosti. Pokud vyjde $\chi^2 > \chi_{k-2}^2(1-\alpha)$, zamítáme hypotézu H_0 na hladině α .

Příklad 20.3.  Pomocí χ^2 -testu dobré shody otestujte na hladině významnosti 0,05, zda počet telefonních volání X na cetrálu hotelu v době 15 minut má Poissonovo rozdělení $\text{Po}(3)$. Tabulka 20.4 ukazuje počty volání na telefonní cetrálu (k) a počet časových intervalů n_k s k voláními během 15 minut.

Tabulka 20.5: Pozorované a očekávané četnosti telefonních volání

i	1 $(-\infty; 1]$	2 $(1; 2]$	3 $(2; 3]$	4 $(3; 4]$	5 $(4; \infty)$
n_i	7	8	10	6	19
$\mathbb{E}_i$	9,957	11,202	11,202	8,401 6	9,222 2

Řešení. Nejdříve sestavíme tabulku 20.5 s intervalovým rozdělením četností (volíme 5 intervalů) a s pozorovanými (n_i) a očekávanými četnostmi ($\mathbb{E}_i$) pro $i = 1, \dots, 5$. Očekávané četnosti $\mathbb{E}_i$ jsou střední hodnoty Poissonova rozdělení $\text{Po}(3)$, například

$$\begin{aligned}\mathbb{E}_1 &= n \cdot p_1 = n \cdot \mathbb{P}(X \in (-\infty; 1]) = n \cdot (\mathbb{P}(X = 0; X = 1)) = \\ &= 50 \cdot 0,199 \doteq 9,957,\end{aligned}$$

kde $\mathbb{P}(X = 0) = \frac{e^{-\lambda}\lambda^0}{0!} = \frac{e^{-3}}{1} \doteq 0,050$ apod. pro $i = 2 \dots, 5$.

Hodnota testové statistiky se vypočítá podle vztahu (20.1):

$$\chi^2 = \sum_{i=1}^5 \frac{(n_i - \mathbb{E}_i)^2}{\mathbb{E}_i} = \frac{(7 - 9,957)^2}{9,957} + \dots + \frac{(19 - 9,222)^2}{9,222} \doteq 12,976.$$

Vypočítáme kritickou hodnotu testu jako kvantil rozdělení χ^2 s $n - 1$ stupni volnosti:

$$\chi_4^2(0,05) \doteq 9,488.$$

$$= \text{CHISQ.INV}(0,95;4)$$

Hodnotu kvantilu porovnáme s testovou statistikou:

$$12,88 > \chi_4^2(0,05).$$

Na zvolené hladině významnosti 0,05 tedy zamítáme nulovou hypotézu H_0 , tj. během 15 minut nebudou průměrně tři volání.

K testování lze využít také $= \text{CHISQ.TEST}$, kterým určíme p -hodnotu testu: $= \text{CHISQ.TEST}(7, \dots, 19; 9,957, \dots, 9,222)$. Dostáváme $p \doteq 0,0114 < 0,05$, tedy zamítáme H_0 na zvolené hladině významnosti. $\triangle$

20.4 Kolmogorovův–Smirnovův¹ jednovýběrový test shody

Kolmogorovův–Smirnovův test využíváme pro určení shody „tvaru“ empirického a populačního rozdělení určité spojité náhodné veličiny. Test se hodí

¹Vladimir Ivanovič Smirnov (1887–1974), ruský matematik.

i pro výběry malého rozsahu. Nulová hypotéza H_0 je tvrzení, že náhodný výběr pochází z určitého rozdělení se spojitou distribuční funkcí $F(x)$, alternativní hypotéza H_1 vyjadřuje, že výběr nepochází z daného rozdělení.

Nejprve zavedme pojem empirická distribuční funkce.

Definice 20.1 (Empirická distribuční funkce)

Nechť $X_1, \dots, X_m$ je náhodný výběr z rozdělení, které má distribuční funkci F . Pro $i = 1, \dots, m$ zavedme náhodné veličiny

$$\xi_i(x) = \begin{cases} 1, & \text{je-li } X_i < x, \\ 0, & \text{je-li } X_i \geq x. \end{cases}$$

Pak empirická distribuční funkce je

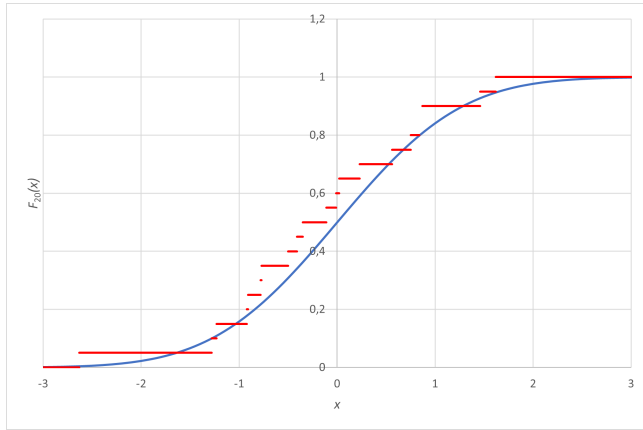
$$F_m(x) = \frac{1}{m} \sum_{i=1}^m \xi_i(x).$$

Empirická distribuční funkce (často nazývaná také empirická kumulativní distribuční funkce) je „skoková“ funkce s výškou skoku $\frac{1}{m}$ v každé z m empirických hodnot a je odhadem kumulativní distribuční funkce, která generovala hodnoty ve výběru.

Příklad 20.4. Generátor náhodných čísel normovaného normálního rozdělení $N(0; 1)$ nám dal následujících dvacet hodnot seřazených vzestupně: $-2,63, -1,28, -1,23, -0,92, -0,91, -0,78, -0,77, -0,50, -0,41, -0,35, -0,11, -0,01, 0,02, 0,23, 0,56, 0,75, 0,84, 0,87, 1,46, 1,62$. Sestavte empirickou distribuční funkci a porovnejte ji s grafem distribuční funkce normovaného normálního rozdělení.

Řešení. Je zadáno celkem 20 hodnot, tedy výška skoku je $\frac{1}{20} = 0,05$. Šířku skoku určují hranice tvořené zadanými hodnotami. Obrázek 20.1 ukazuje distribuční funkci normálního rozdělení $N(0; 1)$ (modrá čára) a empirickou distribuční funkci vytvořenou z výše uvedených hodnot (červené úsečky).

△



Obrázek 20.1: Distribuční funkce $\Phi(x)$ (modrá čára) a empirická distribuční funkce normovaného normálního rozdělení $F_{20}(x)$ (červené úsečky)

Následující věta nám říká, že empirická distribuční funkce je dobrou aproximací distribuční funkce.

Věta 20.1. *Pro každé x platí*

$$F_m(x) \rightarrow F(x)$$

skoro jistě pro $m \rightarrow \infty$.

Označíme-li

$$D_m = \sup_x |F_m(x) - F(x)|,$$

pak platí

$$\mathbb{P} \left(\lim_{m \rightarrow \infty} D_m = 0 \right) = 1.$$

Nechť nyní $X_1, \dots, X_n$ je náhodný výběr z nějakého rozdělení se spojitou distribuční funkcí. Chceme testovat hypotézu H_0 , že tato distribuční funkce je F . Nechť F_m je empirická distribuční funkce odpovídající výběru $X_1, \dots, X_m$. Věta 20.1 nám říká, že velké hodnoty veličiny D_m budou svědčit proti hypotéze H_0 . Je-li m malé, najdeme kritické hodnoty $D_m(\alpha)$ v tabulkách. Při větších hodnotách m se kritické hodnoty aproximují výrazem

$$D_m(\alpha) \doteq \sqrt{\frac{1}{2m} \ln \frac{2}{\alpha}}. \quad (20.2)$$

Tedy H_0 zamítáme v případě, že $D_m \geq D_m(\alpha)$.

Vzhledem k monotónii $F(x)$ nám při výpočtu veličiny D_m stačí omezit se pouze na krajní body intervalů konstantnosti empirické distribuční funkce

F_m . Je-li tudíž x skok F_m , pak vyšetříme hodnotu rozdílu zleva $F_m(x) - F(x)$ a zprava $\lim_{y \rightarrow x+} F_m(y) - F(y)$.

Podobně jako u χ^2 testu musíme znát přesně distribuční funkci $F(x)$. Tudíž tento test můžeme bez modifikace použít pro testování, zda náhodný výběr je z rozdělení $N(0; 1)$, $R(0; 1)$ apod. Testujeme-li ovšem například normalitu náhodného výběru, nemůžeme odhadnout parametry a ty dosadit do distribuční funkce $F(x)$. Pokud bychom to tak udělali, změnilo by se rozdělení testové statistiky D_m , a tedy i kritické hodnoty, při nichž zamítáme hypotézu. Ovšem tyto změněné kritické hodnoty byly určeny pomocí simulačních studií a jsou tabelovány ve speciálních tabulkách. Testy normality pomocí Kolmogorovova-Smirnova testu jsou také implementovány ve statistických softwarech.

Příklad 20.5.  V příkladu 20.4 zjistěte hodnotu statistiky D_m . Určete empirické a teoretické hodnoty distribuční funkce. Kolikátý skok je maximální? Využijte MS Excel.

Řešení. Jak již bylo řečeno, vyšetříme všechny body, ve kterých má empirická distribuční funkce skok, a to jak limity zleva, tak zprava. Maximální hodnota asi 0,142 vyjde u třináctého skoku pro hodnotu 0,02 při limitě zprava:

$$|F_{20}(0,02) - F(0,02)| = |0,65 - 0,50798| \doteq 0,142.$$

Kritická hodnota $D_{20}(0,05) = 0,294$, tedy nezamítáme hypotézu, že výběr je z normovaného normálního rozdělení. Aproximace kritické hodnoty byla vypočtena podle vzorce (20.2) je $D_{20}(0,05) = 0,304$. $\triangle$

20.5 Úlohy

20.1 Tabulka 20.6 ukazuje 150 výsledků měření určité veličiny. Pomocí χ^2 -testu dobré shody na hladině významnosti 0,05 ověřte, zda je možné považovat tento náhodný výběr za výběr z populace s normálním rozdělením.

Tabulka 20.6: Hodnoty veličiny a jejich četnosti

Hodnota veličiny	20	21	22	23	24	25	26	27
Četnost	5	15	20	35	29	32	10	4

20.2 Tabulka 20.7 ukazuje 51 výsledků hodů hrací kostkou. Pomocí χ^2 -testu dobré shody na hladině významnosti 0,05 ověřte, zda je kostka spravedlivá.

Tabulka 20.7: Výsledky hodů hrací kostkou a jejich četnosti

Výsledek hodu	1	2	3	4	5	6
Četnost	10	10	2	25	0	4

20.3 Jsou dána data 8, 9, 10, 15, 17 a 22. Pomocí jednovýběrového Kolmogorova-Smirnova testu zjistíte na hladině významnosti 0,05, zda tato data pocházejí ze souboru hodnot s normálním rozdělením.

20.4 Skupina náhodně vybraných 90 žáků základní školy odpovídala v dotazníku na otázku: „Který z vyučovaných předmětů máš nejraději?“ Dotazník přinesl tyto žákovské preference: matematika – 35 žáků, fyzika – 28 žáků a chemie – 27 žáků. Rozhodněte, zda mezi oblibou předmětů matematika, fyzika a chemie jsou statisticky významné rozdíly.

20.5 V určitém roce zemřelo v České republice na následky dopravních nehod 114 dětí. Tato úmrtí byla rozdělena do jednotlivých měsíců roku tak, jak uvádí tabulka 20.8. Rozhodněte, zda počet úmrtí je ve všech měsících roku stejný, anebo zda se počty úmrtí statisticky významně liší.

Tabulka 20.8: Počty zemřelých dětí při dopravních nehodách v jednotlivých měsících určitého roku

Měsíc	1	2	3	4	5	6	7	8	9	10	11	12
Počet úmrtí	7	7	8	10	10	9	14	12	10	12	7	8

20.6 Řetězec cukráren, který nabízí čtyři druhy zmrzliny, otevřel provozovnu v nové lokalitě. Ve stávajících provozovnách řetězce byla dosud struktura

prodeje podle druhů zmrzliny následující: vanilková 62 %, čokoládová 18 %, jahodová 12 %, pistáciová 8 %. Po otevření provozovny v nové lokalitě máme záznam o následujícím prodeji: vanilková 120, čokoládová 40 jahodová 18, pistáciová 22. Vyjádřete se pomocí statistického testu ke shodě či odlišnosti struktury prodeje v nové lokalitě oproti dosavadním prodejům řetězce.

20.7 S ohledem na šetření vytíženosti personálu se vedení nemocnice zajímá o to, zda jsou pacienti nemocnice propouštěni rovnoměrně během týdne. Počty propouštěných pacientů během jednoho týdne ukazuje tabulka 20.9. Pomocí vhodného testu ověřte, zda jsou pacienti propouštěni rovnoměrně.

Tabulka 20.9: Počty propuštěných pacientů během jednoho týdne

Den v týdnu	Po	Út	St	Čt	Pá	So	Ne
Počet propuštěných pacientů	78	90	94	89	110	84	44

20.8 S ohledem na šetření vytíženosti personálu se vedení nemocnice zajímá o to, zda jsou pacienti nemocnice propouštěni rovnoměrně během pracovních dnů týdne. Počty propouštěných pacientů během pracovních dnů jednoho týdne ukazuje tabulka 20.10. Pomocí vhodného testu ověřte, zda jsou pacienti v pracovních dnech propouštěni rovnoměrně.

Tabulka 20.10: Počty propuštěných pacientů během pracovních dnů jednoho týdne

Den v týdnu	Po	Út	St	Čt	Pá
Počet propuštěných pacientů	78	90	94	89	110

20.9 V programu MS Excel jsme pomocí generátoru náhodných čísel vygenerovali sto náhodných čísel. Ověřte, zda jsou číslice v těchto číslech použity se stejnou pravděpodobností. Tabulka 20.11 obsahuje četnostní zastoupení jednotlivých číslic v generovaných číslech.

Tabulka 20.11: Počty číslic obsažených v generovaných číslech

Číslice	0	1	2	3	4	5	6	7	8	9
Počet	75	87	90	102	97	86	113	89	95	99

20.10 V tabulce 20.12 je uvedeno rozdělení počtu bodových vad zjištěných při zkoumání dvaceti výrobků. Rozhodněte, zda je možno považovat počet vad na výrobku za náhodnou veličinu typu Poissonovo rozdělení.

Tabulka 20.12: Počty výrobků s příslušným počtem bodových vad

Počet vad	0	1	2	3	4	5	6	7
Počet výrobků	2	3	2	4	2	2	4	1

Řešení úloh

20.1 Měřená veličina je náhodná veličina X , která je spojitá. Odhadem μ je výběrový aritmetický průměr $\bar{x} \doteq 23,49$, odhadem σ^2 je výběrový rozptyl $\text{var } X \doteq 2,68$, $s \doteq 1,64$. H_0 : Náhodná veličina X má normální rozdělení $X \sim N(23,49; 2,68)$. H_1 : Náhodná veličina X nemá normální rozdělení $X \sim N(23,49; 2,68)$. Výběr dat rozdělíme do osmi intervalů; intervalové četnosti jsou rovny absolutním. Výpočet testové statistiky ukazuje tabulka 20.13.

Tabulka 20.13: Hodnoty veličiny a jejich četnosti

x_{j-1}	x_j	n_j	p_j	np_j	np_j	n_j	$\frac{(n_j - np_j)^2}{np_j}$
$-\infty$	20	0	0,016 4	2,466 7	-	-	-
20	20,875	5	0,038 5	5,768 9	8,235 6	5	0,071 8
20,875	21,75	15	0,088 6	13,290 95	13,290 95	15	0,219 8
21,75	22,625	20	0,154 4	23,166 0	23,166 0	20	0,432 7
22,625	23,5	35	0,203 7	30,551 05	30,551 05	35	0,647 9
23,5	24,375	29	0,203 2	30,486 2	30,486 2	29	0,072 4
24,375	25,25	32	0,153 5	23,018 8	23,018 8	32	3,504 2
25,25	26,125	10	0,087 7	13,150 5	13,150 5	10	0,754 8
26,125	27	4	0,037 9	5,683 8	8,100 8	4	0,349 96
27	∞	0	0,016 1	2,417 1	-	-	-
Celkem							6,053 4

Protože vyšly dvě teoretické četnosti menší než 5, snížili jsme počet intervalů na 8. Hodnota testového kritéria je 6,053 4. Odhadovali jsme dva parametry (střední hodnota a rozptyl), tedy výsledný počet stupňů volnosti daného χ^2 -testu je $\nu = 8 - 2 - 1 = 5$. Pro zvolenou hladinu významnosti 0,05 je kritická hodnota $\chi^2_5(0,05) \doteq 11,07$. = CHISQ.INV(0,95;5) Protože $\chi^2 < \chi^2_5(0,05)$, nezamítáme H_0 a normální rozdělení $N(23,49; 2,68)$ je vhodným modelem pro popis náhodné veličiny X .

20.2 Hodnota testové statistiky $\chi^2 \doteq 48,4$, což je víc než hodnota testového kritéria $\chi^2(5) \doteq 11,07$, hrací kostka není spravedlivá. p -hodnota je asi $3,45 \cdot 10^{-12} < 0,05$.

20.3 $n = 6$, testová statistika $D_6 \doteq 0,466$, $D_6(0,05) \doteq 0,519$, nezamítáme nulovou hypotézu, že data, z nichž pochází výběr, mají normální rozdělení.

20.4 $\chi^2 \doteq 1,266$, $\chi^2_2(0,05) \doteq 5,991$.

20.5 $\chi^2 \doteq 6$, $p \doteq 0,873\,362$.

20.6 $\chi^2 \doteq 4,32$, kritická hodnota $7,815$, $p \doteq 0,228\,589$.

20.7 $\chi^2 \doteq 29,388$, $p \doteq 0,000\,051$.

20.8 $\chi^2 \doteq 5,822$, $p \doteq 0,212\,834$

20.9 $\chi^2 \doteq 10,397\,6$, $\chi^2_9(0,05) \doteq 16,918$.

20.10 $\chi^2 \doteq 1,239\,086$, $\chi^2_1(0,05) \doteq 3,841\,5$, $p \doteq 0,265\,496$.

Kapitola 21

Kontingenční tabulky

Tato kapitola je věnována základním testům v kontingenčních tabulkách, kterých je celá řada. Dříve, než tyto testy uvedeme, definujeme pojem kontingenční tabulka.

Uvažujme náhodný vektor $Z = (X, Y)$, který má diskrétní rozdělení. Náhodná veličina X nabývá hodnot $1, \dots, r$ a náhodná veličina Y nabývá hodnot $1, \dots, c$. Náhodné veličiny X a Y představují znaky nějakého statistického souboru (např. pohlaví, dosažené vzdělání apod.). Hodnotu znaku sice uvažujeme kladnou celočíselnou, ale ve skutečnosti hodnoty znaku nemusí být číselné, jak je zřejmé z příkladů uvedených v závorce. V mnohých případech přiřazujeme čísla $1, 2, \dots$ jen jako označení. Například dosažené vzdělání: 1 – základní, 2 – střední, 3 – vysokoškolské. Znaky mohou být tudíž

- kvalitativní,
- diskrétní kvantitativní,
- spojitě kvantitativní s hodnotami sloučenými do skupin.

Pro náhodný vektor $Z = (X, Y)$ označme

$$\begin{aligned} p_{ij} &= \mathbb{P}(X = i, Y = j), \\ p_{i.} &= \mathbb{P}(Y = i) = \sum_{j=1}^c p_{ij}, \\ p_{.j} &= \mathbb{P}(Z = j) = \sum_{i=1}^r p_{ij}. \end{aligned}$$

Předpokládejme, že se uskutečnil výběr o rozsahu n z tohoto rozdělení. Počet případů, kdy se ve výběru vyskytla dvojice (i, j) , označme n_{ij} (jde

Tabulka 21.1: Matice pravděpodobností (vlevo) a kontingenční tabulka (vpravo)

	$\mathbf{Z}$			$\mathbf{Z}$	
Y	$1 \dots c$	Σ	Y	$1 \dots c$	Σ
1	$p_{11} \dots p_{1c}$	$p_{1.}$	1	$n_{11} \dots n_{1c}$	$n_{1.}$
$\dots$	$\dots \dots \dots$	$\dots$	$\dots$	$\dots \dots \dots$	$\dots$
r	$p_{r1} \dots p_{rc}$	$p_{r.}$	1	$n_{r1} \dots n_{rc}$	$n_{r.}$
Σ	$p_{.1} \dots p_{.c}$	1	Σ	$n_{.1} \dots n_{.c}$	n

o absolutní četnost). Náhodné veličiny n_{ij} mají sdružené multinomické rozdělení s parametrem n a pravděpodobnostmi p_{ij} . Kontingenční tabulku pak definujeme jako matici (n_{ij}) . Kontingenční tabulka je uvedena v tabulce 21.1 společně s maticí pravděpodobností (p_{ij}) , přičemž

$$n_{i.} = \sum_{j=1}^c n_{ij}, \quad n_{.j} = \sum_{i=1}^r n_{ij}, \quad n = \sum_{i=1}^r \sum_{j=1}^c n_{ij}$$

a platí

$$n = \sum_{j=1}^c n_{.j} = \sum_{i=1}^r n_{i.} = \sum_{i=1}^r \sum_{j=1}^c n_{ij}.$$

Máme-li data uspořádaná do kontingenční tabulky, kdy kategorie jednoho znaku určují řádky a kategorie druhého znaku sloupce, jak je vidět z tabulky 21.1, můžeme testovat následující hypotézy:

- hypotéza nezávislosti dvou náhodných veličin X a Y ,
- hypotéza homogenity multinomických rozdělení,
- hypotéza symetrie.

21.1 Test nezávislosti

Na prvcích souboru sledujeme dva znaky. Naším cílem je testovat nulovou hypotézu

H_0 : náhodné veličiny X (1. znak) a Y (2. znak) jsou nezávislé

proti alternativní hypotéze

H_1 : náhodné veličiny X a Y nejsou nezávislé.

Věta 21.1. *Veličiny X a Y jsou nezávislé tehdy a jen tehdy, platí-li*

$$p_{ij} = p_{i.}p_{.j}, \quad i = 1, \dots, r, \quad j = 1, \dots, c.$$

Vzhledem k tvrzení věty (21.1) můžeme hypotézu nezávislosti přepsat do tvaru

$$H_0: p_{ij} = p_{i.}p_{.j} \text{ pro } i = 1, \dots, r, \quad j = 1, \dots, c.$$

Za platnosti hypotézy H_0 má statistika

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} \quad (21.1)$$

asymptoticky rozdělení χ^2 s počtem stupňů volnosti $(r-1)(c-1)$. Vzorec (21.1) lze přepsat do tvaru

$$\chi^2 = n \sum_{i=1}^r \sum_{j=1}^c \frac{n_{ij}^2}{n_{i.}n_{.j}} - n. \quad (21.2)$$

Hypotézu H_0 o nezávislosti veličin X a Y zamítneme v případě, že

$$\chi^2 \geq \chi_{(r-1)(c-1)}^2(1-\alpha)$$

Ke shodě s limitním rozdělením se vyžaduje, aby všechny teoretické četnosti $\frac{n_{i.}n_{.j}}{n}$ byly větší než 5. Obvykle se požaduje, aby nejméně 80 % teoretických četností bylo větších než 5 a všechny teoretické četnosti výskytu byly větší než 1. Pokud tato podmínka není splněna, spojují se obvykle některé řádky nebo sloupce. Toto ovšem nejde u tzv. čtyřpolních tabulek, což jsou kontingenční tabulky 2×2 . V takovém případě se používá Fisherův faktoriálový test.

K řešení následujícího příkladu je využit speciálně upravený soubor MS Excel, ve kterém je podrobně rozpracován postup výpočtu χ^2 -test.

Příklad 21.1.  Testujeme nezávislost mezi výsledky testů z matematiky a oborem, na který se uchazeč hlásí. Studenti se mohou hlásit na bakalářský obor Finanční matematika, na pětileté magisterské studium učitelství matematiky pro základní školy a na pětileté magisterské studium učitelství matematiky pro střední školy. Obory jsou seřazeny z hlediska obtížnosti studia od nejlehčího (bakalářský obor FM) k nejtěžšímu (pětileté magisterské studium učitelství pro SŠ). Vychází otázka, zda při výběru studia tuto skutečnost uchazeči zohledňují vzhledem ke svým dosavadním studijním výsledkům. Jednoduše řečeno, zda „lepší“ studenti se hlásí na těžší obor

Tabulka 21.2: Kontingenční tabulka empirických a teoretických četností výběru oboru a výsledků v testu

Hodnocení	Aprobace			Celkem
	Fin. mat.	Učitel ZŠ	Učitel SŠ	
1 (60–80 bodů)	9 (9,7)	7 (17,9)	40 (28,4)	56
2 (40–59 bodů)	10 (17,2)	31 (31,6)	58 (50,2)	99
3 (20–39 bodů)	17 (13)	29 (24)	29 (38)	75
4 (0–19 bodů)	14 (10)	25 (18,5)	19 (29,4)	58
Celkem	50	92	146	288

a „horší“ studenti na lehčí. Otestujme, zda existuje závislost mezi výsledky testů z matematiky a oborem, na který se uchazeč hlásí.

Uchazeč může získat z testu maximálně 80 bodů. Veličina X (výsledek testu) nabývá čtyř hodnot, a to 1 – počet získaných bodů 60–80, 2 – počet získaných bodů 40–59, 3 – počet získaných bodů 20–39, 4 – počet získaných bodů 0–19. Veličina Y (studijní obor) nabývá tří hodnot: 1 – finanční matematika, 2 – učitelství pro ZŠ, 3 – učitelství pro SŠ. Veškeré údaje jsou v tabulce 21.2 včetně teoretických četností (čísla v závorkách).

Řešení. Teoretické četnosti vypočítáme postupně jako součin řádkového a sloupcového součtu pro příslušné pole tabulky dělený celkovým rozsahem souboru.

Hodnota testové statistiky $\chi^2 \doteq 27,56$ překračuje kritickou hodnotu $\chi^2_6(0,05) \doteq 12,59$, proto zamítáme H_0 o nezávislosti.

= CHISQ.INV(0,95;6)

Také p -hodnota testu $p \doteq 0,000114$ ukazuje na nezbytnost zamítnout H_0 .

= CHISQ.TEST(aktuální;očekávané)

Tím je statisticky prokázána závislost (H_1) mezi výsledkem testu z matematiky a oborem, na který se student hlásí.

△

21.2 Test homogenity multinomických rozdělení

Tento test se někdy nazývá také jako test shodnosti struktury nebo test o shodnosti struktury. Testujeme shodnost jednoho ze sledovaných znaků za různých podmínek, které vyjadřují kategorie druhého znaku. Například nás může zajímat, zda věková struktura hospitalizovaných pacientů je ve

Tabulka 21.3: Výsledky radiologického vyšetření

Radiologické hodnocení / Léčba	Streptomycin	Kontrolní	Celkem
Významné zlepšení	28 (16,45)	4 (15,55)	32
Střední/malé zlepšení	10 (11,82)	13 (11,18)	23
Beze změn	2 (2,57)	3 (2,43)	5
Střední/malé zhoršení	5 (8,74)	12 (8,26)	17
Významné zhoršení	6 (6,17)	6 (5,83)	12
Úmrtí	4 (9,25)	14 (8,75)	18
Celkem	55	52	107

dvou nemocnicích stejná. Obecně tato nulová hypotéza zní:

H_0 : pravděpodobnosti $q_{i1}, \dots, q_{ic}$ nezávisí na řádkovém indexu i ,

neboli všechny řádky matice (q_{ij}) jsou stejné, kde teoretické pravděpodobnosti $q_{i1}, \dots, q_{ic}$ přísluší relativním marginálním četnostem v i -tém řádku kontingenční tabulky $\frac{n_{i1}}{n_i}, \dots, \frac{n_{ic}}{n_i}$, přičemž $q_{i1} + \dots + q_{ic} = 1$, a dále předpokládáme, že marginální řádkové četnosti n_i jsou předem stanoveny.

Při testování homogenity budeme opět vycházet ze statistiky χ^2 počítané podle vzorce 21.1 nebo 21.2. Za platnosti hypotézy H_0 má statistika χ^2 asymptoticky rozdělení χ^2 s počtem stupňů volnosti $(r-1)(c-1)$. Hypotézu H_0 o homogenitě multinomických rozdělení zamítneme v případě, že $\chi^2 \geq \chi^2_{(r-1)(c-1)}(1-\alpha)$.

Příklad 21.2.  Kontingenční tabulka 21.3 ukazuje výsledky lékařského experimentu ze čtyřicátých let minulého století, který se zabýval účinkem streptomycinu při léčbě plicní tuberkulózy. Údaje z radiologického hodnocení po šesti měsících byly porovnány s tím, zda pacient patřil do léčebné, nebo kontrolní skupiny. Existuje vztah mezi léčbou a výsledkem?



Řešení. Vzhledem k tomu, že testová statistika se opírá o teoretické četnosti $\frac{n_{i.}n_{.j}}{n}$, musíme tyto četnosti vypočítat. Jsou uvedeny v závorkách vedle skutečných četností v tabulce 21.3.

Testová statistika má hodnotu $\chi^2 = 26,96$. Protože $\chi^2_5(0,95) = 11,07$,

 `CHISQ.INV(0,95;5)`

Tabulka 21.4: Čtyřpolní tabulka

n_{11}	n_{12}	$n_{1.}$
n_{21}	n_{22}	$n_{2.}$
$n_{.1}$	$n_{.2}$	n

Tabulka 21.5: Čtyřpolní tabulka obezity rodičů obézních dětí

	Otec obézní	Otec neobézní	Celkem
Matka obézní	15	9	24
Matka neobézní	7	19	26
Celkem	22	28	50

platí $\chi^2 \geq \chi_1^2(1 - \alpha)$, tudíž hypotézu homogenity zamítáme, tzn. na hladině významnosti asymptoticky rovné 0,05 jsme prokázali, že existuje vztah mezi léčbou a výsledkem.

Užitím `= CHISQ.TEST(aktuální;očekávané)` vypočítáme $p \doteq 0,000\,058$, která je však menší než $\alpha = 0,05$, tedy zamítáme H_0 o nezávislosti výsledků obou sledovaných skupin. $\triangle$

21.3 Test χ^2 ve čtyřpolních tabulkách

Jak již bylo poznamenáno výše, v případě $r \times c = 2 \times 2$ mluvíme o tzv. čtyřpolní tabulce. Tato tabulka má tvar jako tabulka 21.4. Ve čtyřpolní tabulce můžeme opět testovat nezávislost a homogenitu. Testová statistika zůstává stejná jako v případě kontingenční tabulky $r \times c$. Vzhledem k tomu, že sčítací indexy nabývají pouze dvou hodnot, lze testovou statistiku χ^2 zjednodušit do následujícího tvaru:

$$\chi^2 = n \frac{(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}.$$

Pokud $\chi^2 \geq \chi_1^2(1 - \alpha)$, zamítáme hypotézu nezávislosti. Stejným způsobem testujeme i homogenitu dvou binomických rozdělení (zobecněním binomického rozdělení je multinomické), jestliže řádkové (nebo sloupcové) marginální četnosti jsou pevné.

Příklad 21.3.  V náhodném výběru padesáti obézních dětí ve věku 6–14 let byla u každého dítěte zjištěna obezita u matky a obezita u otce. Údaje jsou zaznamenány v tabulce 21.5. Zajímá nás, zda obezita rodičů spolu souvisí.

Řešení. Po dosazení do testové statistiky dostaneme

$$\chi^2 = 50 \cdot \frac{(15 \cdot 19 - 9 \cdot 7)^2}{24 \cdot 22 \cdot 26 \cdot 28} \doteq 6,41.$$

Kritická hodnota $\chi_1^2(0,05) \doteq 3,84$.

= CHISQ.INV(0,95;1)

Protože $\chi^2 \geq \chi_1^2(1 - \alpha)$, zamítneme hypotézu nezávislosti, tzn. obezita rodičů spolu významně souvisí.

Zamítnutí H_0 vyplývá také z testu = CHISQ.TEST, podle které je $p \doteq 0,0113 < 0,05$. $\triangle$

21.4 Fisherův faktoriálový (exaktní) test

Jak již bylo poznamenáno výše, ke shodě s limitním rozdělením χ_1^2 se vyžaduje, aby všechny teoretické četnosti $\frac{n_{i \cdot} \cdot n_{\cdot j}}{n}$ byly větší než 5. Pokud tato podmínka není splněna, dochází ke spojování řádků, popř. sloupců. Toto ovšem nelze u čtyřpolní tabulky, a proto se používá Fisherův faktoriálový test. Tento test umožňuje ověřit hypotézu nezávislosti i při malých četnostech. Provedení testu probíhá v následujících krocích.

- 1) Vytvoříme soubor všech kontingenčních tabulek se stejnými marginálními četnostmi jako má původní kontingenční tabulka.
- 2) U každé tabulky souboru vypočteme pravděpodobnosti

$$P = \frac{n_{1 \cdot}! \cdot n_{2 \cdot}! \cdot n_{\cdot 1}! \cdot n_{\cdot 2}!}{n! \cdot n_{11}! \cdot n_{12}! \cdot n_{21}! \cdot n_{22}!}$$

a číslo $d = \ln b$, kde $b = \frac{n_{11}n_{22}}{n_{12}n_{21}}$. Číslo d se nazývá logaritmická interakce dané tabulky.

- 3) Sečteme pravděpodobnosti $\mathbb{P}$ tabulek se stejnými marginálními četnostmi, jako má výchozí tabulka, jejichž logaritmické interakce jsou v absolutní hodnotě větší nebo rovny číslu $|d|$ (logaritmická interakce dané tabulky).
- 4) Je-li součet těchto pravděpodobností menší nebo roven číslu α (hladina testu), hypotézu nezávislosti zamítneme.

Celou proceduru ukážeme na následujícím příkladu.

Příklad 21.4.  U 24 náhodně vybraných žáků se zjišťovalo, zda mají dobrý či špatný prospěch v matematice a zda se učí nebo neučí hrát na nějaký hudební nástroj. Zjištěné výsledky jsou uvedeny v tabulce 21.6. Má se ověřit hypotéza, že prospěch v matematice a okolnost, že se dítě učí hrát na nějaký hudební nástroj, na sobě nezávisí.

Tabulka 21.6: Výsledky studentů v matematice v porovnání se skutečností, zda se učí hrát na hudební nástroj

Prospěch v matematice / Hudba	Učí se	Neučí se	Celkem
Dobrý	6	4	10
Špatný	1	13	14
Celkem	7	17	24

Tabulka 21.7: Dílčí pravděpodobnosti

$\begin{array}{r} 0 \quad 10 \\ 7 \quad 7 \\ \hline d = -\infty \\ \mathbb{P} \doteq 0,009\,916 \end{array}$	$\begin{array}{r} 1 \quad 9 \\ 6 \quad 8 \\ \hline d = -1,91 \\ \mathbb{P} \doteq 0,086\,766 \end{array}$	$\begin{array}{r} 2 \quad 8 \\ 5 \quad 9 \\ \hline d = -0,80 \\ \mathbb{P} \doteq 0,260\,297 \end{array}$	$\begin{array}{r} 3 \quad 7 \\ 4 \quad 10 \\ \hline d = 0,07 \\ \mathbb{P} \doteq 0,347\,063 \end{array}$
$\begin{array}{r} 4 \quad 6 \\ 3 \quad 11 \\ \hline d = 0,89 \\ \mathbb{P} \doteq 0,220\,858 \end{array}$	$\begin{array}{r} 5 \quad 5 \\ 2 \quad 12 \\ \hline d = 1,79 \\ \mathbb{P} \doteq 0,066\,258 \end{array}$	$\begin{array}{r} 6 \quad 4 \\ 1 \quad 13 \\ \hline d = 2,97 \\ \mathbb{P} \doteq 0,008\,495 \end{array}$	$\begin{array}{r} 7 \quad 3 \\ 0 \quad 14 \\ \hline d = \infty \\ \mathbb{P} \doteq 0,000\,347 \end{array}$

Řešení. Vytvoříme všechny tabulky se stejnými marginálními četnostmi, jako má výchozí tabulka. U každé tabulky vypočteme logaritmickou interakci d a pravděpodobnost $\mathbb{P}$ (viz tabulku 21.7).

Výchozí kontingenční tabulka má absolutní hodnotu logaritmické interakce rovnu 2,97, tudíž sčítáme pravděpodobnosti těch tabulek, které mají d v absolutní hodnotě větší nebo rovnu hodnotě 2,97. Součet těchto pravděpodobností je 0,018 758. Vzhledem k tomu, že tento součet není větší než $\alpha = 0,05$, zamítneme hypotézu o nezávislosti. $\triangle$

21.5 McNemarův test

Při statistické analýze kontingenčních tabulek nemusí být vždy cílem provést klasický test nezávislosti nebo homogenity. Další test, který může být proveden v rámci čtyřpolní kontingenční tabulky, je McNemarův¹ test. Tento test se provádí v případě,

PO STATISTICKÉ
TERAPII JE STUDENT
VĚTŠINOU K NEPOZNÁNÍ



¹Quinn Michael McNemar (1900–1986), americký psycholog a statistik.

Tabulka 21.8: Tabulka absolutních četností

Před zásahem	Po zásahu		Celkem
	+	−	
+	n_{11}	n_{12}	$n_{1.}$
−	n_{21}	n_{22}	$n_{2.}$
Celkem	$n_{.1}$	$n_{.2}$	n

Tabulka 21.9: Tabulka pravděpodobností

Před zásahem	Po zásahu		Celkem
	+	−	
+	p_{11}	p_{12}	$p_{1.}$
−	p_{21}	p_{22}	$p_{2.}$
Celkem	$p_{.1}$	$p_{.2}$	1

kdy se na souboru n náhodně vybraných objektů sleduje přítomnost nebo nepřítomnost výskytu nějakého znaku. Posléze se udělá na témže souboru nějaký zákrok a opět se zjistí přítomnost či nepřítomnost sledovaného znaku u jednotlivých objektů souboru. Cílem bude zjistit, zda zákrok změnil pravděpodobnost výskytu znaku.

Označme symbolem $+$ výskyt sledovaného znaku a symbolem $-$ případy, kdy se znak nevyskytl. Obdržíme tabulky 21.8 a 21.9 následujících tvarů, přičemž X = před zásahem, Y = po zásahu, $X = +, -$ a $Y = +, -$. Dále například $p_{11} = \mathbb{P}(X = +, Y = +)$.

Testujeme hypotézu

$$H_0: p_{1.} = p_{.1}.$$

Tato hypotéza je ekvivalentní s hypotézou

$$H_0: p_{12} = p_{21}.$$

Tzn. procento pozitivního výsledku před zásahem je stejné jako po zásahu.

Testová statistika má tvar

$$\chi^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

a má asymptoticky χ_1^2 rozdělení.

Hypotézu H_0 zamítáme v případě, že $\chi^2 \geq \chi_1^2(1 - \alpha)$. Aproximaci pomocí asymptotického rozdělení χ^2 -kvadrát o 1 stupni volnosti můžeme použít,

Tabulka 21.10: Porovnání účinnosti léku A a B

Lék B	Lék A		Celkem
	Úspěch	Neúspěch	
Úspěch	1	3	4
Neúspěch	9	5	14
Celkem	10	8	18

pokud $(n_{12} + n_{21}) \geq 8$. Jestliže není splněna podmínka, nemůže se použít výše zmíněná statistika. Test, který se používá při malých hodnotách $(n_{12} + n_{21})$, můžeme najít například v [1].

Příklad 21.5.  Pozorujeme náhodný výběr 18 pacientů, kteří byli léčeni dvěma různými anti-hypertenzivy A a B. Každý pacient dostával po dobu jednoho měsíce lék A a po odeznění jeho případných účinků dostával po dobu jednoho měsíce lék B. Výsledek byl klasifikován jako úspěch nebo neúspěch. Máme otestovat, zda procenta úspěšnosti jsou u obou léků shodná. Výsledky pozorování jsou uvedeny v tabulce 21.10.



Řešení. Po dosazení do příslušné testové statistiky obdržíme

$$\chi^2 = \frac{(3 - 9)^2}{3 + 9} = 3.$$

Protože příslušná kritická hodnota je $\chi_1^2(0,05) \doteq 3,84$

 `CHISQ.INV(0,95;1)`

a $\chi^2 < \chi_1^2(0,05)$, hypotézu H_0 nezamítneme, tzn. že na základě zkoumaných dat nelze prokázat rozdíl v působení obou léků. $\triangle$

21.6 Test symetrie

Uvažujme čtvercovou kontingenční tabulku 21.11 typu $c \times c$. Příslušná tabulka pravděpodobností má tvar tabulky 21.12. Budeme testovat hypotézu

$$H_0: p_{ij} = p_{ji} \quad \text{pro všechny dvojice } (i, j), \quad i, j = 1, 2, \dots, c.$$

Tabulka 21.11: Tabulka typu $c \times c$

$n_{11} \dots n_{1c}$	$n_{1.}$
$n_{21} \dots n_{2c}$	$n_{2.}$
$\dots$	$\dots$
$n_{c1} \dots n_{cc}$	$n_{c.}$
$n_{.1} \dots n_{.c}$	n

Tabulka 21.12: Tabulka pravděpodobností

$p_{11} \dots p_{1c}$	$p_{1.}$
$p_{21} \dots p_{2c}$	$p_{2.}$
$\dots$	$\dots$
$p_{c1} \dots p_{cc}$	$p_{c.}$
$p_{.1} \dots p_{.c}$	p

Tabulka 21.13: Četnosti manželství při různých původních rodinných stavech partnerů [1]

Ženich/Nevěsta	Svobodná	Ovdovělá	Rozvedená	Celkem
Svobodný	75 564	824	3 463	79 851
Ovdovělý	1 370	904	798	3 072
Rozvedený	4 603	590	2943	8 136
Celkem	81 537	2 318	7 204	91 059

Jde o zobecnění případu 2×2 . McNemarův test je tedy speciálním případem tohoto testu symetrie. Testová statistika má v tomto případě tvar

$$\chi^2 = \sum_{i < j} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}.$$

Za platnosti nulové hypotézy H_0 (hypotéza symetrie) má statistika χ^2 asymptoticky χ^2 rozdělení o $\frac{c(c-1)}{2}$ stupňů volnosti. Hypotézu symetrie zamítneme, jestliže $\chi^2 \geq \chi_1^2(1 - \alpha)$.



Příklad 21.6.  V tabulce 21.13 jsou údaje o rodinném stavu snoubenců. Je třeba rozhodnout, zda pravděpodobnost uzavření sňatku mezi svobodným ženichem a ovdovělou nevěstou je stejná jako pravděpodobnost uzavře-

ní sňatku mezi svobodnou nevěstou a ovdovělým ženichem a že analogická rovnost platí i pro pravděpodobnost ostatních kombinací původních rodinných stavů partnerů.

Řešení. Dosazením do testové statistiky pro test symetrie obdržíme

$$\chi^2 = \frac{(824 - 1\,370)^2}{824 + 1\,370} + \frac{(3\,463 - 4\,603)^2}{3\,463 + 4\,603} + \frac{(798 - 590)^2}{798 + 590} \doteq 328,17.$$

Vzhledem k tomu, že $\chi^2 \geq \chi_3^2(0,95) \doteq 7,81$,

= CHISQ.INV(0,95;3)

hypotézu symetrie zamítáme.

△

21.7 Úlohy

21.1 Pomocí testu nezávislosti otestujte na hladině významnosti 5%, zda je očkování proti chřipce účinné. Otestováno bylo celkem 500 osob, z toho se 120 dalo proti chřipce očkovat, z nich onemocnělo 24. Z neočkovaných onemocnělo 97. Sestavte čtyřpolní tabulku s absolutními četnostmi. Zjistěte, zda je rozdíl mezi podílem osob, které onemocněly, ač byly očkovány, a osobami, které onemocněly, ač nebyly očkovány, signifikantní.



21.2 U čtyřset studentů vysoké školy byl zjišťován navštěvovaný ročník a míra konzumace alkoholu (viz tabulku 21.14). Testujte nulovou hypotézu, že mezi navštěvovaným ročníkem a mírou konzumace alkoholu není žádná závislost. Použijte 5% hladinu významnosti.



Tabulka 21.14: Empirické a očekávané četnosti věkových skupin obyvatel v letech 2011 a 2021 (hodnoty v závorce jsou očekávané četnosti po zaokrouhlení na jedno desetinné místo)

Skupina	Ročník				Celkem
	1	2	3	4	
Piják	29 (38)	41 (34,1)	33 (31,4)	28 (27,5)	131
Konzument	32 (39,4)	29 (35,4)	36 (32,6)	39 (28,6)	136
Abstinent	55 (38,6)	34 (34,6)	27 (31,9)	17 (27,9)	133
h_j	116	104	96	84	400

21.3 U studentů zanechávajících studium byl zjišťován důvod zanechání studia a typ absolvované střední školy. Výsledek je uveden v kontingenční tabulce 21.15. Na 5% hladině významnosti otestujte nezávislost v této kontingenční tabulce.

Tabulka 21.15: Empirické četnosti důvodů zanechání studia podle typu školy (G – gymnázium, SEŠ – Střední ekonomická škola, O – ostatní, 1 – obtížnost studia, 2 – zdravotní důvody, 3 – sociální důvody, 4 – ostatní

Typ SŠ	Důvod zanechání studia				h_i
	1	2	3	4	
G	5	35	10	0	50
SEŠ	30	5	0	5	40
O	10	0	0	0	10
h_j	45	40	10	5	100



21.4 Pomocí testu homogenity multinomických rozdělení otestujte na hladině významnosti 5%, zda je shodné rozdělení počtu obyvatel v jednotlivých kategoriích za roky 2011 a 2021. Potřebná data ukazuje tabulka 21.16.

Tabulka 21.16: Empirické a očekávané četnosti věkových skupin obyvatel v letech 2011 a 2021

Věková skupina / Rok	2011	2021	Celkem
< 14	1,49 (1,58)	1,69 (1,60)	3,18
16–64	7,27 (6,93)	6,68 (7,02)	13,95
> 65	1,64 (1,88)	2,15 (1,91)	3,79
Celkem	10,4	10,52	20,92

21.5 Tabulka 21.17 ukazuje počty dopravních nehod motocyklistů s určitým stupněm zdravotních následků v závislosti na dojezdové vzdálenosti rychlé záchranné služby. Na hladině významnosti 95 % testujte nezávislost vážnosti zranění na dojezdové vzdálenosti rychlé záchranné služby.



Tabulka 21.17: Kontingenční tabulka s počty jednotlivých druhů zranění a dojezdovými časy rychlé záchranné služby v minutách

	Druh zranění				
Dojezdový čas	Žádné	Lehké	Těžké	Smrtelné	Celkem
≤ 10	1 768	807	189	47	2 811
10–20	946	1 387	746	53	3 132
> 20	115	438	288	16	857
Celkem	2 829	2 632	1 223	116	6 800

21.6 Očkování proti chřipce se zúčastnilo 460 dospělých, z nichž 240 dostalo očkovací látku proti chřipce (očkováná skupina) a 220 dostalo placebo (kontrolní skupina). Na konci experimentu onemocnělo 100 lidí chřipkou. Dvacet z nich bylo z očkové skupiny a osmdesát z kontrolní skupiny. Pozorované četnosti jsou v tabulce 21.18. Je dostatečný důkaz proto, že očkovací látka byla účinná, nebo mohly rozdíly mezi očkovanou a kontrolní skupinou vzniknout náhodou? Zvolte hladinu významnosti $\alpha = 0,001$.

Tabulka 21.18: Tabulka absolutních četností a relativních četností v % nemocných chřipkou

Chřipka	Očkování	Placebo	Celkem
Ano	20 (8,3)	80 (36,4)	100 (21,7)
Ne	220 (91,7)	140 (63,6)	360 (78,3)
Celkem	240 (100,0)	220 (100,0)	460 (100,0)

21.7 Kontingenční tabulka 21.19 ukazuje výsledky velmi slavného lékařského experimentu ze čtyřicátých let 20. století, který se zabýval účinkem streptomycinu při léčbě plicní tuberkulózy. Údaje z radiologického hodnocení po šesti měsících byly porovnány s tím, zda pacient patřil do léčené, nebo kontrolní skupiny. Existuje statisticky významný vztah mezi léčbou a výsledkem ($\alpha = 0,01$)?

Tabulka 21.19: Výsledky léčby tuberkulózy

Radiologické hodnocení	Streptomycin	Kontrolní	Celkem
Významné zlepšení	28	4	32
Střední zlepšení	10	13	23
Beze změn	2	3	5
Střední zhoršení	5	12	17
Významné zhoršení	6	6	12
Smrt	4	14	18
Celkem	55	52	107

21.8 U 20 náhodně vybraných studentů se zjišťovalo, zda klima ve škole považují za dobré či špatné a zda se chtějí či nechťejí zúčastnit teambuildingové akce. Zjištěné výsledky jsou uvedeny v tabulce 21.20. Má se ověřit hypotéza, že hodnocení klima studenty a účast na teambuildingové akci na sobě nezávisí.

Tabulka 21.20: Účast studentů na akci v závislosti na hodnocení školního klimatu

Klima / Akce	Zúčastní se	Nezúčastní se	Celkem
Dobré	5	3	8
Špatné	6	6	12
Celkem	11	9	20

21.9 O pracovní místo se uchází celkem 20 uchazečů, kteří se podrobí psychologickému testu a budou ohodnoceni také na základě pohovoru s vedoucím. V obou hodnoceních budou uchazeči ohodnoceni buď „úspěš“ (hodnocení 1), nebo „neúspěš“ (hodnocení 0). Výsledky ukazuje tabulka 21.21. Pomocí McNemarova testu otestujte na hladině významnosti 5 % symetrii (obtížnost) obou částí přijímacího řízení. (Závislost výsledku u pohovoru na úspěchu v testu je očekávatelná.

Tabulka 21.21: Hodnocení uchazečů v přijímacím řízení

Pohovor / Psych. test	0	1	Celkem
0	3	6	9
1	3	8	11
Celkem	6	14	20

21.10 Tabulka 21.22 ukazuje výsledky studentského hodnocení kvality stravování ve dvou menzách A a B ve škále „špatné“, „průměrné“ a „dobré“. Na hladině významnosti 0,05 testujte hypotézu, že je mezi hodnoceními obou stravovacích míst symetrie a že lze obě hodnocení označit za srovnatelná.



Tabulka 21.22: Hodnocení kvality stravování ve dvou menzách A a B

Menza A/B	Špatné	Průměrné	Dobré	Celkem
Špatné	23	27	15	65
Průměrné	45	86	92	223
Dobré	9	110	115	234
Celkem	77	223	222	522

21.11 Známý anglický přírodovědec Francis Galton sledoval v 1000 případech barvu očí otce a jeho syna. Barva očí je zakódována následujícím způsobem:

1 – světle modrá, 2 – modrozelená nebo šedá, 3 – tmavě šedá nebo světle hnědá, 4 – tmavě hnědá. Výsledky pozorování ukazuje tabulka 21.23. Rozhodněte o hypotéze symetrie barvy očí u otce a syna.²

Tabulka 21.23: Tabulka s výsledky pozorování barvy očí

Barva očí syna	Barva očí otce				Celkem
	1	2	3	4	
1	194	70	41	30	335
2	83	124	41	36	284
3	25	34	55	23	137
4	56	36	43	109	224
Celkem	358	264	180	198	1000

21.12 Při malém průzkumu, byl sledován sexuální život manželských dvojic před a po narození dítěte. Uvedené výsledky jsou v tabulce 21.24. Na hladině významnosti 5 % otestujte, zda má narození potomka vliv na sexuální život manželů.

²Úloha včetně dat převzata z [3], s. 298.

Tabulka 21.24: Tabulka s výsledky pozorování sexuálního života manželských párů před a po narození dítěte

Před / Po	Neuspokojivý	Uspokojivý	Velmi dobrý	Celkem
Neuspokojivý	24	9	6	39
Uspokojivý	6	16	11	33
Velmi dobrý	1	8	20	29
Celkem	31	33	37	101

Řešení úloh

21.1 Tabulka 21.25 je čtyřpolní. Rozdíl mezi sledovanými skupinami osob činí 4,2 %. H_0 : Oba jevy jsou navzájem nezávislé. Testová statistika $\chi^2 = 500 \cdot \frac{(24 \cdot 283 - 97 \cdot 96)^2}{120 \cdot 380 \cdot 121 \cdot 379} \doteq 1,52$, $\chi^2(0,95; 1) \doteq 3,84$, $1,52 < 3,84$, tedy nezamítáme H_0 , tj. procentní rozdíl není na hladině významnosti 0,05 statisticky významný, nebo také: účinnost očkovací látky proti chřipce nebyla statisticky prokázána.

Tabulka 21.25: Četnosti zdravých/nemocných a očkovaných/neočkovaných osob při výskytu chřipky

Očkování / Nemocnost	Zdraví	Nemocní	Celkem
Očkování	96	24	120
Neočkování	283	97	380
Celkem	379	121	500

21.2 Test nezávislosti v kontingenční tabulce: H_0 : Mezi navštěvovaným ročníkem a mírou konzumace alkoholu není závislost. H_1 : Mezi navštěvovaným ročníkem a mírou konzumace alkoholu je závislost. $p = 0,05$, kritická hodnota $\chi_p^2((m-1)(n-1)) = \chi_{0,05}^2(2 \cdot 3) \doteq 12,59$, $\chi^2 \doteq 22,291 > \chi_{0,05}^2(6)$, zamítáme H_0 , tedy na hladině významnosti 5 % je mezi navštěvovaným ročníkem studentů a mírou konzumace alkoholu závislost.

21.3 $\chi^2 \doteq 68,65$, kritická hodnota $\chi_{0,05}^2(6) \doteq 12,5916 < \chi^2$, zamítáme H_0 o nezávislosti v kontingenční tabulce, tj. mezi typem střední školy a příčinou zanechání studia je statisticky významná závislost.

21.4 H_0 : Rozdělení četností je shodné. Testová statistika $\chi^2 \doteq 1,69$, $\chi^2(0,95; 1) \doteq 3,84$, $1,69 < 3,84$, tedy nezamítáme H_0 , tj. rozdíly mezi roky 2011 a 2021 ve věkových skupinách obyvatel nejsou na hladině 0,05 statisticky významné.

21.5 H_0 : Vážnost zranění a délka dojezdové doby rychlé záchranné služby jsou nezávislé. H_1 : Vážnost zranění a délka dojezdové doby rychlé záchranné služby nejsou nezávislé.

Hodnota testové statistiky je $\chi^2 \doteq 1\,073,5$, což je výrazně více než hodnota testového kritéria rozdělení $\chi^2_{(3-1) \cdot (4-1)} = \chi^2_6 \doteq 12,59$ při zvolené hladině významnosti 0,05, zamítáme tedy H_0 o nezávislosti obou veličin, což potvrzuje také p -hodnota $p \doteq 1,124 \cdot 10^{-228} < 0,05$.

21.6 H_0 : Procento výskytu chřipky je v očkované a kontrolní skupině stejné. Očekávané četnosti jsou uvedeny v závorkách v tabulce 21.26.

Tabulka 21.26: Tabulka absolutních a relativních četností v % nemocných chřipkou

Chřipka	Očkování	Placebo	Celkem
Ano	52,17 (21,7)	47,83 (21,7)	100 (21,7)
Ne	187,83 (78,3)	172,17 (78,3)	360 (78,3)
Celkem	240 (100,0)	220 (100,0)	460 (100,0)

Pokud mezi očkováním a chřipkou neexistuje žádný vztah, potom by se pozorované a očekávané četnosti výskytu měly navzájem blížit a případná odchylka by byla způsobena pouze náhodou. Nejlepší způsob, jak zjistit rozdíly mezi pozorovanými a očekávanými četnostmi výskytu, je spočítat χ^2 statistiku, kde tentokrát sčítáme přes všechna políčka v tabulce. $\chi^2 = 53$, počet stupňů volnosti $df = (m - 1) \cdot (n - 1) = 1$, $\chi^2_{0,001}(1) = 10,83 < \chi^2$, tzn. že pravděpodobnost, že by tak velké pozorované procentuální rozdíly ve výskytu chřipky mohly vzniknout jen náhodou, pokud by neexistovaly žádné skutečné rozdíly mezi očkovací látkou a placebem, je menší než 0,1 %. Proto můžeme učinit závěr, že očkovací látka je účinnější než placebo.

21.7 Pokud není mezi skupinami žádný rozdíl, získáme očekávané četnosti, které jsou v tabulce 21.27.

Tabulka 21.27: Výsledky léčby tuberkulózy - očekávané četnosti (%)

Radiologické hodnocení	Streptomycin	Kontrolní	Celkem
Významné zlepšení	28 (16,45)	4 (15,55)	32
Střední zlepšení	10 (11,82)	13 (11,18)	23
Beze změn	2 (2,57)	3 (2,43)	5
Střední zhoršení	5 (8,74)	12 (8,26)	17
Významné zhoršení	6 (6,17)	6 (5,83)	12
Smrt	4 (9,25)	14 (8,75)	18
Celkem	55	52	107

$\chi^2 = 26,96$, $df = (6 - 1)(2 - 1) = 5$, pro 5 stupňů volnosti je 1% kritická hodnota rovna kvantilu $\chi^2_{1-0,01}(5) \doteq 15,09$. Hodnota vypočtené testové statistiky $\chi^2 \doteq 26,96$ značně překračuje kritickou hodnotu 15,09, a proto můžeme učinit závěr, že na hladině významnosti $\alpha = 0,01$ existuje dostatečný důkaz vztahu mezi léčbou a rentgenovým výsledkem.

21.8 Vytvoříme všechny tabulky se stejnými marginálními četnostmi, jako má výchozí tabulka. U každé tabulky vypočteme logaritmickou interakci d a pravděpodobnost $\mathbb{P}$ (viz tabulku 21.28).

Tabulka 21.28: Dílčí pravděpodobnosti

0 8 11 1 $d = -\infty$ $\mathbb{P} \doteq 0,000\,071$	1 7 10 2 $d = -3,55$ $\mathbb{P} \doteq 0,003\,144$	2 6 9 3 $d = -2,19$ $\mathbb{P} \doteq 0,036\,675$	3 5 8 4 $d = -1,20$ $\mathbb{P} \doteq 0,165\,039$
4 4 7 5 $d = -0,33$ $\mathbb{P} \doteq 0,330\,078$	5 3 6 6 $d = 0,51$ $\mathbb{P} \doteq 0,308\,073$	6 2 5 7 $d = 1,43$ $\mathbb{P} \doteq 0,132\,031$	7 1 4 8 $d = 2,64$ $\mathbb{P} \doteq 0,023\,577$

Výchozí kontingenční tabulka má absolutní hodnotu logaritmické interakce rovnu 0,51, tudíž sčítáme pravděpodobnosti těch tabulek, které mají d v absolutní hodnotě větší nebo rovnu hodnotě 0,51. Součet těchto pravděpodobností je 0,66861. Vzhledem k tomu, že tento součet je větší než $\alpha = 0,05$, nezamítáme hypotézu o nezávislosti, tedy hodnocení klimatu není závislé na účasti na teambuildingové akci.

21.9 Obě části přijímacího řízení budou stejně obtížné, jestliže pravděpodobnosti $p_{11} + p_{12} = p_{11} + p_{21}$, tedy $p_{12} = p_{21}$. Protože $n_{12} + n_{21} = 9 > 8$, lze využít McNemarův test. $H_0 : p_{12} = p_{21}$, $H_1 : p_{12} \neq p_{21}$. Testová statistika $\chi^2 = \frac{(6-3)^2}{9} = 1$, $\chi^2(0,95; 1) \doteq 3,84$, $1 < 3,84$, tedy nezamítáme H_0 , tj. na hladině významnosti 0,05 se neprokázalo, že by byly obě části přijímacího řízení různě obtížné.

21.10 $H_0 : p_{ij} = p_{ji}$ pro všechny dvojice $(1, j), i, j = 1, 2, 3$,

$$\chi^2 = \sum_{i>j} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}} = \frac{(45 - 27)^2}{45 + 27} + \frac{(9 - 15)^2}{9 + 15} + \frac{(110 - 92)^2}{110 + 92} = 7,6,$$

počet stupňů volnosti pro daný test je $c = \frac{3(3-1)}{2} = 3$, $\chi^2(0,95; 3) \doteq 7,81$. Protože $7,6 < 7,81$, nelze na hladině významnosti 0,05 zamítnout H_0 , tedy hodnocení obou stravovacích míst je srovnatelné.

21.11 $\chi^2 = 19,56$, $\chi^2_6(0,05) = 12,59$, protože $19,56 > 12,59$, zamítáme hypotézu symetrie, tj. zamítáme, že pravděpodobnost, že „otec má barvu očí i , syn má barvu očí j “, je jiná než pravděpodobnost, že „otec má barvu očí j , syn má barvu očí i “.

21.12 $\chi^2 = 4,64$, $df = \frac{k(k-1)}{2} = 3$, $\chi^2_3(0,05) = 7,81$, protože $4,64 < 7,81$, nelze zamítnout H_0 , tj. vliv narození dítěte na sexuální život manželů nelze na zvolené hladině významnosti prokázat.

Závěr

Publikace, kterou jste jako čtenáři a zájemci o pravděpodobnost a statistiku právě prošli, je kolektivním dílem pracovníků Katedry datových věd a informačních systémů na Zemědělské a technologické fakultě a Katedry matematiky Pedagogické fakulty Jihočeské univerzity v Českých Budějovicích. Snad se podařila naplnit jejich vize o učebnici, která bude studentům dobrým pomocníkem při studiu.



Hodnotu publikace lze spatřovat v zařazení více příkladů s podrobným řešením a více úloh k procvičení s příloženými výsledky, případně s pokyny a návody k postupu řešení. Podstatným doplňkem jsou také odkazy na funkce a nástroje MS Excel, které lze při řešení úloh z pravděpodobnosti a statistiky využít. Jde totiž o nejdostupnější software užívaný mezi studenty a veřejností.

Odborný výkladový text je doplněn mnoha poznámkami běžného jazyka, které mají za úkol zpřístupnit všem čtenářům význam a ukázat konkrétní použití probíraného učiva. Učebnice svým obsahem zdaleka téma pravděpodobnosti a statistiky nevyčerpává, doporučujeme tedy čtenáři dále studovat z publikací uvedených v seznamu literatury.

Autoři děkují recenzentům za podnětné připomínky!

Summary

The publication that you, as readers and those interested in probability and statistics, have just read is a collective work of staff from the Department of Data Science and Computing Systems at the Faculty of Agriculture and Technology, and the Department of Mathematics at the Faculty of Education of the University of South Bohemia in České Budějovice. Hopefully, it was possible to fulfill their vision with a textbook that will be a good helper for students during their studies.

The value of the publication can be seen in the inclusion of more examples with detailed solutions and more problems to practice with attached results, possibly with instructions and instructions for the solution procedure. References to MS Excel functions and tools, which can be used when solving problems from probability and statistics, are also an essential addition. It is the most accessible software used by students and the public.

The expert explanatory text is supplemented by many notes of common language, which have the task of making the meaning accessible to all readers and showing the specific use of the subject matter discussed. The content of the textbook does not cover the topic of probability and statistics, so we recommend that the reader continue to study from the publications listed in the bibliography.

The authors thank the reviewers for their stimulating comments!

Rejstřík

analýza

- korelační, 353

- regresní, 363

analýza rozptylu

- dvojného třídění, 333

- jednoduchého třídění, 325

argument souboru, 233

Bernoulliho proces, 48

Bernoulliho schéma, 48

borelovské množiny, 84

četnost, 235

- absolutní třídění, 236

- empirická, 389, 391

- kumulativní (součtová), 236

- kumulativní relativní, 236

- relativní, 22, 236

- relativní procentní, 236

- teoretická, 389, 405, 409

chyba

- druhého druhu, 286

- přípustná, 275

- prvního druhu, 286

- směrodatná, 275

- směrodatná regrese, 365, 368,
372

definice

- absolutně spojitě náhodné
veličiny, 78

absolutně spojitého rozdělení

- náhodného vektoru, 183

aritmetického průměru, 244

axiomatická definice

- pravděpodobnosti, 19

charakteristik variability, 249

diskrétní náhodné veličiny,
75

diskrétního

- pravděpodobnostního

- prostoru, 31

diskrétního rozdělení

- náhodného vektoru, 181

distribuční funkce, 71

distribuční funkce náhodného
vektoru, 179

druhů četností, 235

empirické distribuční funkce,
396

geometrického průměru, 244

harmonického průměru, 244

klasického

- pravděpodobnostního

- prostoru, 25

koeficientu šikmosti, 255

koeficientu špičatosti, 255

konfidenčního intervalu, 274

konvergence náhodných
veličin, 220

- kovariance náhodných veličin, 185
- kvantilové funkce, 98
- kvantilu, 247
- mediánu, 247
- modu, 248
- n -tého momentu, 104
- náhodné veličiny, 67
- náhodného vektoru, 179
- náhodného výběru, 260
- nezávislých náhodných jevů, 44
- Pearsonova výběrového korelačního koeficientu, 353
- po dvou nezávislých jevů, 45
- podmíněné pravděpodobnosti, 40
- pravděpodobnosti náhodného jevu, 24
- rozptylu, 101, 263
- σ -algebry, 21
- skupinově nezávislých jevů, 45
- směrodatné odchylky, 263
- Spearmanova korelačního koeficientu, 356
- spojitého pravděpodobnostního prostoru, 32
- střední hodnoty náhodné veličiny, 93
- teoretických a empirických charakteristik, 261
- variačního oboru a třídění, 238
- varianční a korelační matice náhodného vektoru, 187
- výběrového průměru, 263
- vzájemně nezávislých náhodných veličin, 191
- délka třídy, 240
- diagram
 - spojnicový, 237
 - úsečkový, 237
- disperze, 249
- falešná negativita, 286
- falešná pozitivita, 286
- Fatouovo lemma, 96
- formule úplné pravděpodobnosti, 58
- funkce
 - beta, 169
 - distribuční, 71
 - distribuční diskrétní, 210
 - distribuční empirická, 396
 - distribuční marginální, 184
 - distribuční náhodného vektoru, 181
 - distribuční sdružená, 179, 185
 - Euleroва, 169
 - kvantilová, 98
 - měřitelná, 84
 - náhodných veličin, 203
 - pravděpodobnostní, 70
 - pravděpodobnostní diskrétní náhodné veličiny, 75
 - regresní, 363, 370
- histogram, 241
- hladina testu, 286
- hodnota
 - p -hodnota, 287
 - kritická, 264
- hranice třídy
 - dolní, 240
 - horní, 240
- hustota rozdělení náhodné veličiny, 80
- marginální, 184
- sdružená, 185

- hustota rozdělení náhodného vektoru, 183
- hustota rozdělení pravděpodobnosti, 78
- hustota vektoru marginální, 184
- hypotéza, 60, 285
 - alternativní, 285
 - homogenity, 404
 - jednostranná, 285
 - nezávislosti náhodných veličin, 404
 - nulová, 285
 - oboustranná, 285
 - pravostranná, 285
 - symetrie, 404
- integrál
 - Eulerův prvního druhu, 169
 - Lebesgueův, 96
 - Lebesgueův–Stieltjesův, 95
 - Riemannův, 96
- interakce, logaritmická, 409
- interval
 - konfidenční, 274
 - spolehlivosti, 274
 - třídní, 240
- jevové pole, 21
- kauzalita, 357
- koeficient
 - determinace, 365, 372
 - korelační Spearmanův, 354
 - korelační výběrový, 353
 - pořadové korelace, 355
 - šikmosti (asymetrie), 255
 - Spearmanův korelační, 355
 - špičatosti, 255
 - spolehlivosti, 274
 - variační, 250
- kontingenční tabulka, 403
- konvergence
 - náhodných veličin, 220
 - podle pravděpodobnosti, 220
 - podle středu stupně 2, 220
 - skoro jistě, 220
- konvoluce, 208
 - binomických rozdělení, 211
 - exponenciálních rozdělení, 211
 - hustot, 210
 - normálních rozdělení, 211
 - Poissonových rozdělení, 211
 - rovnoměrných rozdělení, 210
- korelace, 357
- kovariance, 185
- křivka Gaussova, 161
- kvantil, 98, 264
 - výběrový, 247
- kvartil
 - dolní, 99
 - horní, 99
 - výběrový dolní, 247
 - výběrový horní, 247
- logaritmická interakce čtyřpolní tabulky, 409
- matice
 - korelační, 187
 - varianční, 187
- medián, 98, 247
- metoda
 - linearizace, 380
 - nejmenších čtverců, 364, 371
 - Neményiova, 332
 - Tukeyova, 327
 - Tukeyova mnohonásobného porovnání, 336
- míra, 83
 - Diracova, 113
 - Lebesgueova, 30, 84
 - Lebesgueova–Stieltjesova, 84

- pravděpodobnostní, 83
- model
 - linearizovatelný, 380
 - lineární, 371
 - nelineární, 380
 - regresní, 363
- modus, 100, 248
- moment
 - n -tý, 104
- moment náhodné veličiny, 104
 - absolutní, 104
 - centrální, 104
- náhodná proměnná, 67
- náhodná veličina, 67
 - absolutně spojitá, 78
 - diskrétní, 70, 75
- náhodné jevy
 - nezávislé, 44, 47
 - po dvou nezávislé, 45
 - skupinově (totálně) nezávislé, 45
- náhodné veličiny
 - nekorelované, 208
- náhodný jev, 21
 - doplňkový, 28
 - elementární, 19
 - jistý, 21
 - nemožný, 21
- náhodný výběr
 - oblastní (stratifikovaný), 260
 - prostý bez vracení, 259
 - prostý s vracením, 259
- náhodný vektor, 179
- náhodný výběr
 - systematický (mechanický), 260
- nerovnost
 - Cauchyho–Schwarzova–
 - Buňakovského, 186, 354
 - Čebyševova, 103, 221
- nezávislost
 - náhodných veličin po dvou, 191
 - náhodných veličin vzájemná, 191
 - složek náhodného vektoru, 192
- nezávislost náhodných jevů, 44
- obor
 - kritický, 286
 - nezamítnutí hypotézy, 286
 - zamítnutí hypotézy, 286
- očekávaná hodnota, 94
- odchylka
 - decilová, 251
 - kvantilová, 251
 - kvartilová, 251
 - percentilová, 251
 - průměrná, 250
 - směrodatná, 102, 176, 249
- odhad
 - bodový, 269
 - intervalový, 273
 - jednostranný intervalový, 275
 - konzistentní, 272
 - nestranný, 271
 - oboustranný intervalový, 275
 - rozptylu, 277
 - střední hodnoty intervalový, 278
 - vydatný, 273
- ogiva, 237
- pás spolehlivosti regresní přímky, 366
- podjev, 20
- podmodel, 375
- pokus
 - Bernoulliho, 48, 113
 - deterministický, 12
 - náhodný, 12
- polygon rozložení četností, 237

- pořadí, 313
- porovnání více výběrů, 325
- pravděpodobnost, 22
 - aposteriorní, 61
 - apriorní, 60
 - geometrická, 30
 - náhodného jevu, 24
 - podmíněná, 39
 - pravděpodobnostní míra, 24
- pravděpodobnostní prostor, 24
- problém
 - De Montmortův přiřazovací, 54
- proces
 - Poissonův, 132
 - událostí
 - v diskrétním čase, 132
 - ve spojitém čase, 132
- průměr
 - aritmetický, 244
 - geometrický, 244
 - harmonický, 244
 - useknutý, 246
 - výběrový, 263
 - winsorizovaný, 246
- regrese
 - kvadratická, 376
 - linearizovatelná, 380
 - lineární, 363
 - lineární, s více vysvětlujícími proměnnými, 370
 - nelineární, 380
 - polynomiální, 376
- regresní polynom, 380
- rozdělení
 - alternativní, 113
 - Cauchyho, 91, 171
 - empirické, 219
 - pravděpodobnosti, 70
 - relativních četností, 219
 - Simpsonovo, 211
 - rozdělení náhodné veličiny, 84
 - absolutně spojité, 97
 - asymptoticky normální, 223
 - Bernoulliho, 113
 - bimodální, 248
 - binomické, 115
 - Cauchyho, 95
 - diskrétní, 97
 - dvourozměrné normální
 - obecné, 190
 - Erlangovo, 151
 - exponenciální, 145
 - F, 171
 - Fisherovo-Snedecorovo, 171
 - Fisherovo-Snedecorovo, 172
 - gama, 153
 - geometrické, 121
 - hypergeometrické, 129
 - leptokurtické, 105, 256
 - logaritmicko-normální, 164
 - mesokurtické, 105, 256
 - mnohorozměrné, 189
 - multinomické, 189
 - negativně binomické, 126
 - normální normované
 - (standardizované), 159
 - normální obecné, 155
 - nula-jedničkové, 113
 - Pascalovo, 126
 - Pearsonovo χ^2 , 166
 - Pearsonovo χ^2 , 167
 - platykurtické, 105, 256
 - Poissonovo, 132
 - Pólyovo, 126
 - rovnoměrné spojité, 143
 - spojité, 78
 - Studentovo t , t rozdělení, 169
 - unimodální, 248
 - vícerozměrné normální, 190
 - rozdělení náhodného vektoru
 - absolutně spojité, 183

- diskrétní, 181
- rozpětí
 - decilové, 250
 - kvantilové, 250
 - kvartilové, 250
 - percentilové, 251
 - studentizované rozpětí, 328
 - variační, 239
- rozptyl, 101
 - populační, 249
 - reziduální, 327, 372
 - výběrový, 249
- rozsah souboru, 233
- shoda v pořadí, 313
- σ -algebra
 - borelovská, 32
- σ -algebra, 21
 - borelovská, 22, 30
- signifikance, 289
- šikmost, 105
- síla testu, 286
- šířka třídy, 240
- sloupcový graf, 237
- součet čtverců
 - celkový, 326, 365, 372
 - řádkový, 335
 - reziduální, 326, 365, 372
 - sloupcový, 335
- soustava normálních rovnic, 372
- špičatost, 105
- statistická indukce, 259
- statistická inference, 259
- statistická jednotka, 233
- statistický soubor, 233
 - uspořádaný, 247
 - výběrový, 259
 - základní, 259
- stochastika, 11
- střed třídy, 240
- střední hodnota, 93
- tabulka
 - čtyřpolní, 408
 - rozložení četností, 237
- teorie pravděpodobnosti, 11
- test
 - Durbin-Watsonův, 369
 - Fisherův faktoriálový, 409
 - Friedmanův, 337, 339
 - homogenity, 406
 - Kolmogorovův-Smirnovův
 - jednovýběrový test
 - shody, 395
 - Kruskalův-Wallisův, 327, 332
 - Levenův, 328
 - McNemarův, 410
 - neparametrický, 313
 - nezávislosti, 404
 - normality, 391
 - o shodnosti struktury, 406
 - o střední hodnotě pomocí
 - CLV, 303
 - párový t -test, 292
 - Pearsonův, 389
 - Poissonova rozdělení, 393
 - rovnosti dvou rozptylů, 297
 - symetrie, 412
 - t , dvouvýběrový, 295
 - t , jednovýběrový, 288
 - Wilcoxonův dvouvýběrový,
 - 319
 - Wilcoxonův jednovýběrový,
 - 316
 - znaménkový, 314
- testová statistika, 286
- třída, 240
- třídění
 - dvojné, 333
 - jednoduché, 233
 - trojné, 337
- úhrn argumentu, 233
- úloha

- Banachova, 54
- věta
 - o celkové pravděpodobnosti, 57, 58
 - o náhodném výběru z normálního rozdělení, 263
- varianční matice náhodného vektoru, 187
- věta
 - Bayesova, 60
 - centrální limitní (CLV), 222
 - de Moivreova-Laplaceova, 223
 - Lévyho-Lindebergova, 226
 - Ljapunovova, 228
 - Moivreova-Laplaceova, 223
 - o násobení pravděpodobností, 42
 - o nezávislosti jevů, 46
 - o vlastnostech distribuční funkce, 180
 - o vlastnostech střední hodnoty, 95
 - o vztahu mezi průměry, 245
- vzorec
 - Bayesův první, 60
- z-skór, 160
- zákon velkých čísel, 219
 - silný, 222, 263
 - slabý, 221
- zešikmení
 - doleva, 255
 - doprava, 255
- znak
 - souboru, 233
 - třídní, 240

Funkce a nástroje MS Excel pro pravděpodobnost a statistiku

= ABS (číslo) vrátí absolutní hodnotu vybraného čísla. Funkci lze využít například při výpočtu průměrné absolutní odchylky.

Analýza dat, nástroj na kartě Data, který slouží k testování hypotéz. Tento nástroj není na nástrojové liště v MS Excel běžně zobrazen. Je-li potřeba jej zobrazit, přejdeme do Soubor, Možnosti, Doplňky, Spravovat doplňky Excelu – Přejít. Poté se zobrazí samostatné okno, ve kterém zvolíme možnost Analytické nástroje a potvrdíme OK. Příkaz Analýza dat je pak k dispozici ve skupině Analýza na kartě Data. Analýza dat vrátí také hodnotu Pearsonova a Spearmanova korelačního koeficientu r_s pro získaná pořadí.



- **Anova: jeden faktor:** nástroj provede jednoduchou analýzu rozptylu dat pro dva nebo více vzorků, analýza testuje hypotézu, že každý vzorek je odebraný ze stejného rozdělení pravděpodobnosti. Pokud existují jenom dva vzorky, je možné použít funkci **= T.TEST**.
- **Anova: dva faktory s opakováním:** nástroj slouží k testování dat sebraných dvakrát na stejném souboru a k ověření příslušnosti každé dvojice k základnímu souboru. V opačném případě se nevyloučí alternativní hypotéza, podle které existují další vlivy, než byly sledovány.
- **Anova: dva faktory bez opakování:** jako u nástroje Anova: dva faktory s opakováním, zde se však předpokládá, že pro každou dvojici existuje pouze jedno pozorování.

- **Korelace:** slouží k výpočtu korelačního koeficientu mezi dvěma měřenými proměnnými.
- **Kovariance:** umožňuje zjistit závislost dvou měřených proměnných. Závislost znamená, že vysoké hodnoty jedné proměnné odpovídají vysokým hodnotám druhé proměnné (kladná kovariance), nebo že nízké hodnoty jedné proměnné odpovídají vysokým hodnotám proměnné druhé (záporná kovariance). Pokud jsou hodnoty obou proměnných nezávislé, bude kovariance blízká nule.
- **Popisná statistika:** slouží k určení počtu hodnot souboru dat, k určení součtu, minimální a maximální hodnoty, aritmetického průměru, mediánu, modu, směrodatné výběrové odchylky, výběrového rozptylu, šikmosti a špičatosti a rozdílu maxima a minima (variační šíře).
- **Regrese:** provede lineární regresi tak, že pomocí metody nejmenších čtverců proloží přímkou sadou pozorování. Regrese umožňuje analyzovat, jakým způsobem ovlivňují hodnoty jedné nebo více nezávislých proměnných hodnotu jedné závislé proměnné.
- **Dvouvýběrový F -test pro rozptyl:** nástroj provede F -test pro dva výběry a porovná jejich rozptyly na zvolené hladině významnosti. Hodnota f blíží se 1 potvrzuje rovnost rozptylů základních souborů.
- **Dvouvýběrový párový t -test na střední hodnotu:** slouží k testování středních hodnot souborů v každém výběru.
- **Dvouvýběrový t -test s rovností rozptylů:** provede Studentův t -test pro dva soubory za předpokladu, že data pochází z rozdělení se stejnými rozptyly.
- **Dvouvýběrový t -test s nerovností rozptylů:** nástroj provede Studentův t -test pro dva soubory dat za předpokladu, že dvě sady dat pochází z rozdělení s různými rozptyly.
- **Dvouvýběrový z -test na střední hodnotu:** test se používá k testování nulové hypotézy, že neexistuje rozdíl mezi dvěma středními hodnotami.

= **BINOM.DIST**(počet úspěchů;pokusy;pravděpodobnost úspěchu; PRAVDA/NEPRAVDA) vrátí hodnotu distribuční funkce F_X /pravděpodobnostní funkce p_X binomického rozdělení diskretní náhodné veličiny X .

= **CEILING.MATH**(číslo;významnost;režim) zaokrouhlí číslo nahoru na nejblíže celé číslo nebo na nejblíže násobek významnosti.

= **CONFIDENCE.NORM**(alpha;směrodatná odchylka;velikost) vrátí interval spolehlivosti pro střední hodnotu základního souboru pro normální rozdělení. Pod „velikost“ se rozumí rozsah výběrového souboru.

= **CORREL**(matice1;matice2) vrátí hodnotu Pearsonova korelačního koeficientu r_s .

= **COVARIANCE**(matice1;matice2) vrátí kovarianci, průměr součinů odchy-

lek pro každou dvojici bodů dat ve dvou množinách dat.

= **COVARIANCE.P**(m1;m2) vrátí kovarianci základního souboru, průměr součinů odchylek pro každou dvojici bodů dat ve dvou množinách dat.

= **COVARIANCE.S**(m1;m2) vrátí kovarianci výběru, průměr součinů odchylek pro každou dvojici datových bodů ve dvou množinách dat.

= **CONFIDENCE.NORM**(alpha;směrodatná odchylka;rozsah výběru) vrátí interval spolehlivosti pro střední hodnotu základního souboru pomocí normálního rozdělení.

= **CONFIDENCE.T**(alpha;směrodatná odchylka;velikost) vrátí interval spolehlivosti pro střední hodnotu základního souboru pro Studentovo t rozdělení. Pod „velikost“ se rozumí rozsah souboru.

= **CRITBINOM**(pokusy;pravděpodobnost;alfa) vrátí nejmenší hodnotu, pro kterou má součtové binomické rozdělení hodnotu větší nebo rovnu hodnotě kritéria. „Pokusy“ se myslí počet Bernoulliho pokusů a „pravděpodobnost“ se myslí pravděpodobnost každého úspěchu.

= **ČETNOSTI**(data;hodnoty) vypočte počet výskytů hodnot v oblasti hodnot a vrátí vertikální matici čísel, která má o jeden řádek více než soubor hodnot. Funkce slouží k tvorbě tabulky rozdělení četností a lze využít i při určování intervalové četnosti zadáním sloupce horních hranic intervalů.

Při zpracovávání a analýze dat bývá prvním krokem uspořádání dat podle velikosti, buď od nejmenšího k největšímu nebo obráceně, to lze provést kliknutím pravého tlačítka na myši na libovolnou buňku s daty uspořádanými ve sloupci a zvolit Seřadit a Od nejmenšího k největšímu, resp. Od největšího k nejmenšímu. Jinou možností se stejným výsledkem je zvolit možnost Seřadit na záložce Data.

= **DEVSQ**(číslo1;číslo2; ...) vrátí součet druhým mocnin odchylek zadaných hodnot od jejich střední hodnoty.

= **EXP**(číslo) vrátí hodnotu mocniny Eulerova³ čísla e ($e \doteq 2,718$).

= **EXPON.DIST**(x;λ;PRAVDA/NEPRAVDA) vrátí hodnotu distribuční funkce F_X /hustoty pravděpodobnost f_X exponenciálního rozdělení spojité náhodné veličiny X .

= **F.DIST**(x;volnost1;volnost2;PRAVDA/NEPRAVDA) vrátí levostrannou hodnotu pravděpodobnosti F rozdělení náhodné veličiny X .

= **F.DIST.RT**(x;volnost1;volnost2;PRAVDA/NEPRAVDA) vrátí pravostrannou hodnotu pravděpodobnosti F rozdělení náhodné veličiny X .

= **F.INV**(pravděpodobnost;volnost1;volnost2) vrátí hodnotu kvantilu distribuční funkce F_X F rozdělení náhodné veličiny X .

= **F.TEST**(m1;m2) vrátí výsledek F -testu jednostranné pravdě-

³Leonhard Paul Euler (1707–1783) švýcarská matematik a fyzik.

podobnosti rovnosti rozptylů dat v oblastech matic1 a matic2 (argument funkce není testové kritérium, takže nelze posoudit míru shody).

= **FAKTORIÁL**(číslo) určí hodnotu faktoriálu zvoleného čísla.

= **FORECAST.LINEAR**(x;pole y;pole x) slouží k výpočtu nebo odhadu budoucí hodnoty pomocí existujících hodnot. Budoucí hodnota je hodnota y pro danou hodnotu x . Existující hodnoty jsou známé jako hodnoty x a y a budoucí hodnota se předpovídá pomocí lineární regrese. x označuje datový bod, u kterého chceme předpovědět hodnotu y , „pole y “ je závislá matice nebo oblast dat, „pole x “ je nezávislá matice nebo oblast dat.

= **GAMMA.DIST**(x;alfa;beta;kumulativní) vrátí hodnotu Gamma rozdělení.

= **GAMMA.INV**(pravděpodobnost;alfa;beta) vrátí hodnotu inverzní funkce k distribuční funkci kumulativního rozdělení Gamma.

= **GEOMEAN**(číslo1;číslo2; ...) vrátí geometrický průměr množiny kladných čísel.

Grafy a diagramy v MS Excel lze vytvářet pomocí bohaté palety nástrojů na kartě *Vložit* v části *Grafy*.

= **HARMEAN**(číslo1;číslo2; ...) vrátí harmonický průměr množiny zadáných čísel.

= **HYPGEOM.DIST**(úspěch;celkem;základ úspěch;úspěch celkem; PRAVDA/NEPRAVDA) vrátí hodnotu distribuční funkce F_X (pravděpodobnostní funkce p_X) hypergeometrického rozdělení diskrétní náhodné veličiny X .

= **CHISQ.DIST**(x;volnost;PRAVDA/NEPRAVDA) vrátí levostrannou hodnotu pravděpodobnosti χ^2 rozdělení náhodné veličiny X .

= **CHISQ.DIST.RT**(x;volnost;PRAVDA/NEPRAVDA) vrátí pravostrannou hodnotu pravděpodobnosti χ^2 rozdělení náhodné veličiny X .

= **CHISQ.TEST**(aktuální;očekávané) vrátí test nezávislosti, vrátí hodnotu rozdělení χ^2 pro dané testové kritérium a příslušné stupně volnosti.

= **INTERCEPT**(pole y;pole x) vrátí y -ovou souřadnici průsečíku s osou y lineární regresní čáry proložené zadanými datovými body.

= **KOMBINACE**(počet;kombinace) určí hodnotu kombinací k -té třídy z n prvků.

Kontingenční tabulka, nástroj lze využít k tvorbě tabulek rozdělení četností. Tvorba kontingenční tabulky se zahajuje v záložce **Data** výběrem položky *Kontingenční tabulka*.

= **KURT**(číslo1;číslo2;...) vrátí hodnotu excesu množiny zadáných čísel (špičatost). Funkce KURT počítá koeficient špičatosti (označme a_4^*) podle jiného vzorce, než je v definici 12.9:

$$a_4^* = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_n} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}. \quad (21.3)$$

= LARGE(pole;k) vrátí k -tou největší hodnotu z množiny zadaných čísel.
= LINREGRESE(pole y;pole x;b;stat) vypočítá pomocí metody nejmenších čtverců statistické hodnoty pro přímkou $y = mx + b$, která nejlépe odpovídá uvedeným datům, a vrátí matici s parametry přímky m , b .

= LINTREND(pole y;pole x;nová x;b) vrátí hodnoty lineárního trendu odpovídajícího známým datovým bodům pomocí metody nejmenších čtverců.

= LN(číslo) vrátí přirozený logaritmus zadaného čísla.

= LOG(číslo) vrátí dekadický logaritmus zvoleného čísla.

= LOGZ(číslo;základ) vrátí hodnotu logaritmu zvoleného čísla o zadaném základu.

= LOGNORM.DIST(x ;střední hodnota;směrodatná odchylka; PRAVDA/NEPRAVDA) vrátí hodnotu distribuční funkce F_X /hustoty pravděpodobnosti f_X logaritmicko-normálního rozdělení spojitě náhodné veličiny X .

= LOGNORM.INV(pravděpodobnost; střední hodnota; směrodatná odchylka) vrátí inverzní funkci ke kumulativní distribuční funkci logaritmicko-normálního rozdělení náhodné veličiny X , kde $\ln X$ má normální rozdělení s parametry μ a σ^2 .

= MAX(číslo1;číslo2; ...) vrátí maximální hodnotu množiny zadaných čísel.

= MEDIAN(číslo1;číslo2; ...) vrátí medián (prostřední hodnotu) množiny zadaných čísel.

= MIN(číslo1;číslo2; ...) vrátí minimální hodnotu množiny zadaných čísel.

= MODE.MULT(číslo1;číslo2; ...) vrátí modus (nejčtenější hodnotu) množiny zadaných čísel.

= MODE.SNGL(číslo1;číslo2; ...) vrátí modus (nejčtenější hodnotu) množiny zadaných čísel.

= MULTINOMIAL(číslo1;číslo2; ...) určí hodnotu permutací s opakováním. Funkce slouží k výpočtu variací s opakováním, kde jednotlivá čísla jsou počty opakování týchž prvků.

= NEGBINOM.DIST(počet neúspěchů;počet úspěchů;pravděpodobnost úspěchu;PRAVDA/NEPRAVDA) vrátí hodnotu distribuční funkce F_X /pravděpodobnostní funkce p_X negativně binomického rozdělení diskrétní náhodné veličiny X .

= NORM.DIST(x ;střední hodnota;směrodatná odchylka; PRAVDA/NEPRAVDA) vrátí hodnotu distribuční funkce F_X /hustoty pravděpodobnosti f_X normálního rozdělení spojitě náhodné veličiny X .

= NORM.S.DIST(z ;PRAVDA/NEPRAVDA) vrátí hodnotu distribuční funkce F_X /hustoty pravděpodobnosti f_X standardizovaného normálního rozdělení spojitě náhodné veličiny X .

= **NORM.INV**(pravděpodobnost;střední;směrodatná odchylka) vrátí hodnotu kvantilové funkce F_X^{-1} , která přísluší zadané pravděpodobnosti náhodné veličiny X , která má normální rozdělení.

= **NORM.S.INV**(pravděpodobnost) vrátí hodnotu kvantilové funkce Φ_X^{-1} , která přísluší zadané pravděpodobnosti náhodné veličiny X , která má standardizované normální rozdělení.

= **ODMOCNINA**(číslo) vrátí druhou odmocninu ze zvolené hodnoty.

= **PEARSON**(matice1;matice2) vrátí hodnotu Pearsonova korelačního koeficientu r .

= **PERCENTIL.EXC**(matice;kvartil) vrátí k -tý percentil ze souboru dat, $k \in (0; 1)$.

= **PERCENTIL.INC**(matice;kvartil) vrátí k -tý percentil ze souboru dat, $k \in (0; 1)$.

= **PERMUTACE**(počet;permutace) určí počet variací k -té třídy z n prvků nebo počet permutací z n prvků.

= **PHI**(x) vrátí hodnotu funkce hustoty pro standardní směrodatnou odchylku.

= **POČET**(číslo1;číslo2; ...) vrátí počet prvků množiny zadaných čísel.

= **POISSON.DIST**(x ;střední;PRAVDA/NEPRAVDA) vrátí hodnotu distribuční funkce F_X /pravděpodobnostní funkce p_X Poissonova rozdělení diskrétní náhodné veličiny X .

= **POWER**(číslo;exponent) umocní dané číslo na zvolený exponent. Funkce se uplatní při výpočtu variací s opakováním.

= **PROB**(oblast hodnot x ;dolní limit;horní limit) vrátí pravděpodobnost, že hodnoty dané oblasti budou mezi dvěma limity. Pokud není zadán horní limit, vrátí funkce pravděpodobnost, že hodnoty ve zvolené oblasti jsou rovny dolnímu limitu.

= **PRŮMĚR**(číslo1;číslo2; ...) vrátí aritmetický průměr.

= **PRŮMODCHYLKA**(číslo1;číslo2; ...) vrátí průměrnou absolutní odchylku množiny zadaných čísel.

= **QUARTIL.INC**(matice;kvartil) = **QUARTIL.EXC**(matice;kvartil)
vrátí hodnotu kvartilu množiny dat vyjádřenou hodnotami percentilu, $k \in \{0; 1; 2; 3; 4\}$, kde $k = 0$ je minimální hodnota, $k = 1$ je první (dolní) kvartil (25% percentil), $k = 2$ je medián (50% percentil), $k = 3$ je třetí (horní) kvartil (75% percentil) a $k = 4$ je maximální hodnota.

= **RANDBETWEEN**(dolní;horní) vrátí náhodné celé číslo ze zadaného intervalu.

= **RANK.AVG**(číslo;odkaz;pořadí) vrátí pořadí čísla v seznamu čísel, tedy jeho relativní velikost vzhledem k ostatním hodnotám v seznamu. Pokud má stejné pořadí více hodnot, bude vráceno průměrné pořadí. Třetí argu-

ment 1 znamená řazení vzestupně, 2 znamená pořadí sestupně.

= SKEW(číslo1; číslo2; ...) vrátí zešikmení rozdělení množiny zadaných hodnot. Funkce SKEW počítá koeficient šikmosti podle jiného vzorce, než je v definici 12.9:)

$$a_3^* = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_n} \right)^3. \quad (21.4)$$

= SKEW.P(číslo1; číslo2; ...) vrátí zešikmení rozdělení hodnot základního souboru.

= SLOPE(pole y; pole x) vrátí směrnici lineární regresní čáry proložené zadanými datovými body.

= SMALL(pole; k) vrátí k -tou nejmenší hodnotu z množiny zadaných hodnot.

= SMODCH.P(číslo1; číslo2; ...) vrátí směrodatnou odchylku základního souboru.

= SMODCH.VÝBĚR.S(číslo1; číslo2; ...) vrátí výběrovou směrodatnou odchylku množiny zadaných čísel. Směrodatnou odchylku je možné vypočítat také jako druhou odmocninu z výběrového rozptylu pomocí **= ODMOCNINA**.

= SOUČIN.SKALÁRNÍ(pole1; ...) vrátí součet součinů odpovídajících si dat. Lze využít k výpočtu střední hodnoty EX diskrétní náhodné veličiny.

= STANDARDIZE(x; střední hodnota; směrodatná odchylka) vrátí normalizovanou hodnotu z rozdělení určeného střední hodnotou a směrodatnou odchylkou.

= STEYX(pole y; pole x) vrátí standardní chybu při výpočtu lineární regrese. Standardní chyba je určena množstvím chyb při odhadu y pro jednotlivá x . Funkce **= STEYX** využívá k výpočtu standardní chyby odhadu y vzorec

$$\sqrt{\frac{1}{n-2} \left[\sum (y - \bar{y})^2 - \frac{[\sum (x - \bar{x})(y - \bar{y})]^2}{\sum (x - \bar{x})^2} \right]},$$

kde x a y jsou nezávislá a závislá čísla ze vstupních matic a n je celkový rozsah souboru všech dat.

= SUMA(číslo1; číslo2; ...) vrátí součet prvků zadané číselné množiny.

= T.DIST(x; volnost; PRAVDA/NEPRAVDA) vrátí hodnotu levostranného Studentova t -rozdělení náhodné veličiny X .

= T.DIST.2T(x; volnost) vrátí hodnotu oboustranného Studentova t -rozdělení náhodné veličiny X .

= T.DIST.RT(x; volnost) vrátí hodnotu pravostranného Studentova t -rozdělení náhodné veličiny X .

= T.TEST(matice1;matice2;chvosty;typ) vrátí výsledek t -testu při testování rozdílnosti nebo shodnosti středních hodnot dat dvou výběrových souborů. Argument „chvosty“ určuje, zda se jedná o jednostranné či dvoustranné rozdělení. Argument „typ“ určuje typ testu.

= TRIMMEAN(matice;procenta) vrátí střední hodnotu vnitřní části množiny datových hodnot, která vznikla vynecháním určitého počtu procent dat z nejvyšších a nejnižších hodnot souboru. Funkce je vhodná pro případ useknutého průměru.

= VAR.P(číslo1;číslo2; ...) vrátí rozptyl základního souboru.

= VAR.S(číslo1;číslo2; ...) vrátí výběrový rozptyl množiny zadaných čísel.

= WEIBULL.DIST(x ;alpha;beta;PRAVDA/NEPRAVDA) vrátí hodnotu hustoty pravděpodobnosti, resp. distribuční funkce Weibullova rozdělení náhodné veličiny X .

= Z.TEST(matice; x ;sigma) vrátí jednostrannou p -hodnotu Z testu, kde „sigma“ značí směrodatnou odchylku datového souboru.

Statistické tabulky

Tabulka 21.29: Kritické hodnoty pro znaménkový test

Počet párů n	Hladina významnosti α		
	0,01	0,05	0,10
5	-	-	0
6	-	0	0
7	-	0	0
8	0	0	1
9	0	1	1
10	0	1	1
11	0	1	2
12	1	2	2
13	1	2	3
14	1	2	3
15	2	3	3
16	2	3	4
17	2	4	4
18	3	4	5
19	3	4	5
20	3	5	5



Tabulka 21.30: Kritické hodnoty k_1 a k_2 pro znaménkový test,
 $\mathbb{P}(Y \leq k_1) \leq \frac{\alpha}{2}$, $\mathbb{P}(Y \geq k_2) \leq \frac{\alpha}{2}$

$\alpha = 0,05 \quad \alpha = 0,01$				
N	k_1	k_2	k_1	k_2
6	0	6	-	-
7	0	7	-	-
8	0	8	0	8
9	1	8	0	9
10	1	9	0	10
11	1	9	0	11
12	2	10	1	11
13	2	11	1	12
14	2	12	1	13
15	3	12	2	13
16	3	13	2	14
17	4	13	2	15
18	4	14	3	15
19	4	15	3	16
20	5	15	3	17

$\alpha = 0,05 \quad \alpha = 0,01$				
N	k_1	k_2	k_1	k_2
21	5	16	4	17
22	5	17	4	18
23	6	17	4	19
24	6	18	5	19
25	7	18	5	20
26	7	19	6	20
27	7	20	6	21
28	8	20	6	22
29	8	21	7	22
30	9	21	7	23
31	9	22	7	24
32	9	23	8	24
33	10	23	8	25
34	10	24	9	25
35	11	24	9	26

$\alpha = 0,05 \quad \alpha = 0,01$				
N	k_1	k_2	k_1	k_2
36	11	25	9	27
37	12	25	10	27
38	12	26	10	28
39	12	27	11	28
40	13	27	11	29
41	13	28	11	30
42	14	28	12	30
43	14	29	12	31
44	15	29	13	31
45	15	30	13	32
46	15	31	13	33
47	16	31	14	33
48	16	32	14	34
49	17	32	15	34
50	17	33	15	35

Tabulka 21.31: Kritické hodnoty w_n jednovýběrového oboustranného a jednostranného Wilcoxonova testu, $\mathbb{P}(\min(S^+, S^-) \leq W_n(\alpha)) \leq \alpha$

n	Oboustranný test		Jednostranný test	
	$\alpha = 0,05$	$\alpha = 0,01$	$\alpha = 0,05$	$\alpha = 0,01$
5	–	–	0	–
6	0	–	2	–
7	2	–	3	0
8	3	0	5	1
9	5	1	8	3
10	8	3	10	5
11	10	5	13	7
12	13	7	17	9
13	17	9	21	12
14	21	12	25	15
15	25	15	30	19
16	29	19	35	23
17	34	23	41	27
18	40	27	47	32
19	46	32	53	37
20	52	37	60	43
21	58	42	67	49
22	65	48	75	55
23	73	54	83	62
24	81	61	91	69
25	89	68	100	76
26	98	75	110	84
27	107	83	119	92
28	116	91	130	101
29	126	100	140	110
30	137	109	151	120

Tabulka 21.32: Kritické hodnoty $W(0,05)$ pro dvouvýběrový Wilcoxonův test,
 $\mathbb{P}(\min(U_1, U_2) \leq W(0,05)) \leq 0,05$

m/n	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
4	-	-	0														
5	-	0	1	2													
6	-	1	2	3	5												
7	-	1	3	5	6	8											
8	0	2	4	6	8	10	13										
9	0	2	4	7	10	12	15	17									
10	0	3	5	8	11	14	17	20	23								
11	0	3	6	9	13	16	19	23	26	30							
12	1	4	7	11	14	18	22	26	29	33	37						
13	1	4	8	12	16	20	24	28	33	37	41	45					
14	1	5	9	13	17	22	26	31	36	40	45	50	55				
15	1	5	10	14	19	24	29	34	39	44	49	54	59	64			
16	1	6	11	15	21	26	31	37	42	47	53	59	64	70	75		
17	2	6	11	17	22	28	34	39	45	51	57	63	69	75	81	87	
18	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99
19	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106
20	2	8	14	20	27	34	41	48	55	62	69	76	83	90	98	105	112
21	3	8	15	22	29	36	43	50	58	65	73	80	88	96	103	111	119
22	3	9	16	23	30	38	45	53	61	69	77	85	93	101	109	117	125
23	3	9	17	24	32	40	48	56	64	73	81	89	98	106	115	123	132
24	3	10	17	25	33	42	50	59	67	76	85	94	102	111	120	129	138
25	3	10	18	27	35	44	53	62	71	80	89	98	107	117	126	135	145
26	4	11	19	28	37	44	53	62	71	80	89	98	107	117	126	135	145
27	4	11	20	29	38	48	57	67	77	87	97	107	117	127	137	147	158
28	4	12	21	30	40	50	60	70	80	90	101	111	122	132	143	154	164
29	4	13	22	32	42	52	62	73	83	94	105	116	127	138	149	160	171
30	5	13	23	33	43	54	65	76	87	98	109	120	131	143	154	166	177

Tabulka 21.33: Kritické hodnoty $W(0,01)$ pro dvouvýběrový Wilcoxonův test, $\mathbb{P}(\min(U_1, U_2) \leq W(0,05)) \leq 0,05$

m/n	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
5	-	-	-	0													
6	-	-	0	1	2												
7	-	-	0	1	3	4											
8	-	-	1	2	4	6	7										
9	-	0	1	3	5	7	9	11									
10	-	0	2	4	6	9	11	13	16								
11	-	0	2	5	7	10	13	16	18	21							
12	-	1	3	6	9	12	15	18	21	24	27						
13	-	1	3	7	10	13	17	20	24	27	31	34					
14	-	1	4	7	11	15	18	22	26	30	34	38	42				
15	-	2	5	8	12	16	20	24	29	33	37	42	46				
16	-	2	5	9	13	18	22	27	31	36	41	45	50				
17	-	2	6	10	15	19	24	29	34	39	44	49	54	60			
18	-	2	6	11	16	21	26	31	37	42	47	53	58	64	70		
19	0	3	7	12	17	22	28	33	39	45	51	57	63	69	74	81	
20	0	3	8	13	18	24	30	36	42	48	54	60	67	73	79	86	92
21	0	3	8	14	19	25	32	38	44	51	58	64	71	78	84	91	98
22	0	4	9	14	21	27	34	40	47	54	61	68	75	82	89	96	104
23	0	4	9	15	22	29	35	43	50	57	64	72	79	87	94	102	109
24	0	4	10	16	23	30	37	45	52	60	68	75	83	91	99	107	115
25	0	5	10	17	24	32	39	47	55	63	71	79	87	96	104	112	121
26	0	5	11	18	25	33	41	49	58	66	74	83	92	100	109	118	127
27	1	5	12	19	27	35	43	52	60	69	78	87	96	105	114	123	132
28	1	5	12	20	28	36	45	54	63	72	81	91	100	109	119	128	138
29	1	6	13	21	29	38	47	56	66	75	85	94	104	114	124	134	144
30	1	6	13	22	30	40	49	58	68	78	88	98	108	119	129	139	150

Tabulka 21.34: Kritické hodnoty $q_{m,\nu}(0,05)$ pro Tukeyovu metodu mnohonásobného porovnání, $X \sim q_{m,\nu}$, $\mathbb{P}(X \geq q_{m,\nu}(0,05)) = 0,05$

m/n	2	3	4	5	6	7	8	9	10	11	12	13	14
1	18,0	27,0	32,8	37,1	40,4	43,1	45,4	47,4	49,1	50,6	52,0	53,2	54
2	6,08	8,33	9,80	10,9	11,7	12,4	13,0	13,5	14,0	14,4	14,7	15,1	15
3	4,50	5,91	6,82	7,50	8,04	8,48	8,85	9,18	9,46	9,72	9,95	10,2	10
4	3,93	5,04	5,76	6,29	6,71	7,05	7,35	7,60	7,83	8,03	8,21	8,37	8,5
5	3,64	4,60	5,22	5,67	6,03	6,33	6,58	6,80	6,99	7,17	7,32	7,47	7,6
6	3,46	4,34	4,90	5,30	5,63	5,90	6,12	6,32	6,49	6,65	6,79	6,92	7,0
7	3,34	4,16	4,68	5,06	5,36	5,61	5,82	6,00	6,16	6,30	6,43	6,55	6,6
8	3,26	4,04	4,53	4,89	5,17	5,40	5,60	5,77	5,92	6,05	6,18	6,29	6,3
9	3,20	3,95	4,41	4,76	5,02	5,24	5,43	5,59	5,74	5,87	5,98	6,09	6,1
10	3,15	3,88	4,33	4,65	4,91	5,12	5,30	5,46	5,60	5,72	5,83	5,93	6,0
11	3,11	3,82	4,26	4,57	4,82	5,03	5,20	5,35	5,49	5,61	5,71	5,81	5,9
12	3,08	3,77	4,20	4,51	4,75	4,95	5,12	5,27	5,39	5,51	5,61	5,71	5,8
13	3,06	3,73	4,15	4,45	4,69	4,88	5,05	5,19	5,32	5,43	5,53	5,63	5,7
14	3,03	3,70	4,11	4,41	4,64	4,83	4,99	5,13	5,25	5,36	5,46	5,55	5,6
15	3,01	3,67	4,08	4,37	4,59	4,78	4,94	5,08	5,20	5,31	5,40	5,49	5,5
16	3,00	3,65	4,05	4,33	4,56	4,74	4,90	5,03	5,15	5,26	5,35	5,44	5,5
17	2,98	3,63	4,02	4,30	4,52	4,70	4,86	4,99	5,11	5,21	5,31	5,39	5,4
18	2,97	3,61	4,00	4,28	4,49	4,67	4,82	4,96	5,07	5,17	5,27	5,35	5,4
19	2,96	3,59	3,98	4,25	4,47	4,65	4,79	4,92	5,04	5,14	5,23	5,31	5,3
20	2,95	3,58	3,96	4,23	4,45	4,62	4,77	4,90	5,01	5,11	5,20	5,28	5,3
24	2,92	3,53	3,90	4,17	4,37	4,54	4,68	4,81	4,92	5,01	5,10	5,18	5,2
30	2,89	3,49	3,85	4,10	4,30	4,46	4,60	4,72	4,82	4,92	5,00	5,08	5,1
40	2,86	3,44	3,79	4,04	4,23	4,39	4,52	4,63	4,73	4,82	4,90	4,98	5,0
60	2,83	3,40	3,74	3,98	4,16	4,31	4,44	4,55	4,65	4,73	4,81	4,88	4,9
120	2,80	3,36	3,68	3,92	4,10	4,24	4,36	4,47	4,56	4,64	4,71	4,78	4,8
∞	2,77	3,31	3,63	3,86	4,03	4,17	4,29	4,39	4,47	4,55	4,62	4,68	4,7

Tabulka 21.35: Kritické hodnoty $q_{m,\nu}(0,01)$ pro Tukeyovu metodu mnoho-
násobného porovnání, $X \sim q_{m,\nu}$, $\mathbb{P}(X \geq q_{m,\nu}(0,01)) = 0,01$

ν/m	2	3	4	5	6	7	8	9	10	11	12	13	14
1	90,0	135	164	186	202	216	227	237	246	253	260	266	272
2	14,0	19,0	22,3	24,7	26,6	28,2	29,5	30,7	31,7	32,6	33,4	34,1	34,8
3	8,26	10,6	12,2	13,3	14,2	15,0	15,6	16,2	16,7	17,1	17,5	17,9	18,2
4	6,51	8,12	9,17	9,96	10,6	11,1	11,5	11,9	12,3	12,6	12,8	13,1	13,3
5	5,70	6,97	7,80	8,42	8,91	9,32	9,67	9,97	10,2	10,5	10,7	10,9	11,1
6	5,24	6,33	7,03	7,56	7,97	8,32	8,61	8,87	9,10	9,30	9,49	9,65	9,81
7	4,95	5,92	6,54	7,01	7,37	7,68	7,94	8,17	8,37	8,55	8,71	8,86	9,00
8	4,74	5,63	6,20	6,63	6,96	7,24	7,47	7,68	7,87	8,03	8,18	8,31	8,44
9	4,60	5,43	5,96	6,35	6,66	6,91	7,13	7,32	7,49	7,65	7,78	7,91	8,03
10	4,48	5,27	5,77	6,14	6,43	6,67	6,87	7,05	7,21	7,36	7,48	7,60	7,71
11	4,39	5,14	5,62	5,97	6,25	6,48	6,67	6,84	6,99	7,13	7,25	7,36	7,46
12	4,32	5,04	5,50	5,84	6,10	6,32	6,51	6,67	6,81	6,94	7,06	7,17	7,26
13	4,26	4,96	5,40	5,73	5,98	6,19	6,37	6,53	6,67	6,79	6,90	7,01	7,10
14	4,21	4,89	5,32	5,63	5,88	6,08	6,26	6,41	6,54	6,66	6,77	6,87	6,96
15	4,17	4,83	5,25	5,56	5,80	5,99	6,16	6,31	6,44	6,55	6,66	6,76	6,84
16	4,13	4,78	5,19	5,49	5,72	5,92	6,08	6,22	6,35	6,46	6,56	6,66	6,74
17	4,10	4,74	5,14	5,43	5,66	5,85	6,01	6,15	6,27	6,38	6,48	6,57	6,66
18	4,07	4,70	5,09	5,38	5,60	5,79	5,94	6,08	6,20	6,31	6,41	6,50	6,58
19	4,05	4,67	5,05	5,33	5,55	5,73	5,89	6,02	6,14	6,25	6,34	6,43	6,51
20	4,02	4,64	5,02	5,29	5,51	5,69	5,84	5,97	6,09	6,19	6,29	6,37	6,45
25	3,96	4,54	4,91	5,17	5,37	5,54	5,69	5,81	5,92	6,02	6,11	6,19	6,26
30	3,89	4,45	4,80	5,05	5,24	5,40	5,54	5,65	5,76	5,85	5,93	6,01	6,08
40	3,82	4,37	4,70	4,93	5,11	5,27	5,39	5,50	5,60	5,69	5,77	5,84	5,90
60	3,76	4,28	4,60	4,82	4,99	5,13	5,25	5,36	5,45	5,53	5,60	5,67	5,73
120	3,70	4,20	4,50	4,71	4,87	5,01	5,12	5,21	5,30	5,38	5,44	5,51	5,56
∞	3,64	4,12	4,40	4,60	4,76	4,88	4,99	5,08	5,16	5,23	5,29	5,35	5,40

Tabulka 21.36: Kritické hodnoty pro Neményho metodu mnohonásobného porovnávání pořadí, $\alpha = 0,05$

m/I	3	4	5	6	7	8	9	10
1	3,3	4,7	6,1	7,5	9,0	10,5	12,0	13,5
2	8,8	12,6	16,5	20,5	24,7	28,9	33,1	37,4
3	15,7	22,7	29,9	37,3	44,8	52,5	60,3	68,2
4	23,9	34,6	45,6	57,0	68,6	80,4	92,4	104,6
5	33,1	48,1	63,5	79,3	95,5	112,0	128,8	145,8
6	43,3	62,9	83,2	104,0	125,3	147,0	169,1	191,4
7	54,4	79,1	104,6	130,8	157,6	184,9	212,8	240,9
8	66,3	96,4	127,6	159,6	192,4	225,7	259,7	294,1
9	78,9	114,8	152,0	190,2	229,3	269,1	309,6	350,6
10	92,3	134,3	177,8	222,6	268,4	315,0	362,4	410,5
11	106,3	154,8	205,0	256,6	309,4	363,2	417,9	473,3
12	120,9	176,2	233,4	292,2	352,4	413,6	476,0	539,1
13	136,2	198,5	263,0	329,3	397,1	466,2	536,5	607,7
14	152,1	221,7	293,8	367,8	443,6	520,8	599,4	679,0
15	168,6	245,7	325,7	407,8	491,9	577,4	664,6	752,8
16	185,6	270,6	358,6	449,1	541,7	635,9	732,0	829,2

Tabulka 21.37: Kritické hodnoty pro Neményho metodu mnohonásobného porovnávání pořadí, $\alpha = 0,01$

m/I	3	4	5	6	7	8	9	10
1	4,1	5,7	7,3	8,9	10,5	12,2	13,9	15,6
2	10,9	15,3	19,7	24,3	28,9	33,6	38,3	43,1
3	19,5	27,5	35,7	44,0	52,5	61,1	69,8	78,6
4	29,7	41,9	54,5	67,3	80,3	93,6	107,0	120,6
5	41,2	58,2	75,8	93,6	111,9	130,4	149,1	168,1
6	53,9	76,3	99,3	122,8	146,7	171,0	195,7	220,6
7	67,6	95,8	124,8	154,4	184,6	215,2	246,3	277,7
8	82,4	116,8	152,2	188,4	225,2	262,6	300,6	339,0
9	98,1	139,2	181,4	224,5	268,5	313,1	358,4	404,2
10	114,7	162,8	212,2	262,7	314,2	366,5	419,5	473,1
11	132,1	187,6	244,6	302,9	362,2	422,6	483,7	545,6
12	150,4	213,5	278,5	344,9	412,5	481,2	551,0	621,4
13	169,4	240,6	313,8	388,7	464,9	542,4	621,0	700,5
14	189,1	268,7	350,5	434,2	519,4	606,0	693,8	782,6
15	209,6	297,8	388,5	481,3	575,8	671,9	769,3	867,7
16	230,7	327,9	427,9	530,1	634,2	740,0	847,3	955,7

Tabulka 21.38: Kritické hodnoty Friedmanova testu, $\alpha = 0,05$

I/J	3	4	5	6	7	8	9	10	11	12
3	6,000	7,4	8,53	9,86	11,24	12,57	13,88	15,19	16,48	17,76
4	6,500	7,8	8,8	10,24	11,63	12,99	14,34	15,67	16,98	18,3
5	6,400	7,8	8,99	10,43	11,84	13,23	14,59	15,93	17,27	18,6
6	7,000	7,6	9,08	10,54	11,97	13,38	14,76	16,12	17,4	18,8
7	7,143	7,8	9,11	10,62	12,07	13,48	14,87	16,23	17,6	18,9
8	6,250	7,65	9,19	10,68	12,14	13,56	14,95	16,32	17,7	19,0
9	6,222	7,66	9,22	10,73	12,19	13,61	15,02	16,40	17,7	19,1
10	6,200	7,67	9,25	10,76	12,23	13,66	15,07	16,44	17,8	19,2
11	6,545	7,68	9,27	10,79	12,27	13,70	15,11	16,48	17,9	19,2
12	6,167	7,70	9,29	10,81	12,29	13,73	15,15	16,53	17,9	19,3
13	6,000	7,70	9,30	10,83	12,32	13,76	15,17	16,56	17,9	19,3
14	6,143	7,71	9,32	10,85	12,34	13,78	15,19	16,58	17,9	19,3
15	6,400	7,72	9,33	10,87	12,35	13,80	15,20	16,6	18,0	19,3
16	5,99	7,73	9,34	10,88	12,37	13,81	15,23	16,6	18,0	19,3
20	5,99	7,74	9,37	10,92	12,41	13,8	15,3	16,7	18,0	19,4
∞	5,99	7,82	9,49	11,07	12,59	14,07	15,51	16,92	18,31	19,68

Tabulka 21.39: Kritické hodnoty Friedmanova testu, $\alpha = 0,01$

I/J	3	4	5	6	7	8	9	10	11	12
3	-	9,000	10,13	11,76	13,26	14,78	16,28	17,74	19,19	20,61
4	8,000	9,600	11,20	12,59	14,19	15,75	17,28	18,77	20,24	21,7
5	8,400	9,96	11,43	13,11	14,74	16,32	17,86	19,37	20,86	22,3
6	9,000	10,200	11,75	13,45	15,10	16,69	18,25	19,77	21,3	22,7
7	8,857	10,371	11,97	13,69	15,35	16,95	18,51	20,04	21,5	23,0
8	9,000	10,35	12,14	13,87	15,53	17,15	18,71	20,24	21,8	23,2
9	8,667	10,44	12,27	14,01	15,68	17,29	18,87	20,42	21,9	23,4
10	9,600	10,53	12,38	14,12	15,79	17,41	19,00	20,53	22,0	23,5
11	9,455	10,60	12,46	14,21	15,89	17,52	19,10	20,64	22,1	23,6
12	9,500	10,68	12,53	14,28	15,96	17,59	19,19	20,73	22,2	23,7
13	9,385	10,72	12,58	14,34	16,03	17,67	19,25	20,80	22,3	23,8
14	9,000	10,76	12,64	14,40	16,09	17,72	19,31	20,86	22,4	23,9
15	8,933	10,80	12,68	14,44	16,14	17,78	19,35	20,9	22,4	23,9
16	8,79	10,84	12,72	14,48	16,18	17,81	19,40	20,9	22,5	24,0
20	8,87	10,94	12,83	14,60	16,30	18,00	19,5	21,1	22,6	24,1
∞	9,21	11,35	13,28	15,09	16,81	18,48	20,09	21,67	23,21	24,73

Tabulka 21.40: Kritické hodnoty pro mnohonásobná porovnání u Friedmanova testu, $\alpha = 0,05$

I/J	3	4	5	6	7	8	9	10
1	3,3	4,7	6,1	7,5	9,0	10,5	12,0	13,5
2	4,7	6,6	8,6	10,7	12,7	14,8	17,0	19,2
3	5,7	8,1	10,6	13,1	15,6	18,2	20,8	23,5
4	6,6	9,4	12,2	15,1	18,0	21,0	24,0	27,1
5	7,4	10,5	13,6	16,9	20,1	23,5	26,9	30,3
6	8,1	11,5	14,9	18,5	22,1	25,7	29,4	33,2
7	8,8	12,4	16,1	19,9	23,9	27,8	31,8	35,8
8	9,4	13,3	17,3	21,3	25,5	29,7	34,0	38,3
9	9,9	14,1	18,3	22,6	27,0	31,5	36,0	40,6
10	10,5	14,8	19,3	23,8	28,5	33,2	38,0	42,8
11	11,0	15,6	20,2	25,0	29,9	34,8	39,8	44,9
12	11,5	16,2	21,1	26,1	31,2	36,4	41,6	46,9
13	11,9	16,9	22,0	27,2	32,5	37,9	43,3	48,8
14	12,4	17,5	22,8	28,2	33,7	39,3	45,0	50,7
15	12,8	18,2	23,6	29,2	34,9	40,7	46,5	52,5
16	13,3	18,8	24,4	30,2	36,0	42,0	48,1	54,2

Tabulka 21.41: Kritické hodnoty pro mnohonásobná porovnání u Friedmanova testu, $\alpha = 0,01$

I/J	3	4	5	6	7	8	9	10
1	4,1	5,7	7,3	8,9	10,5	12,2	13,9	15,6
2	5,8	8,0	10,3	12,6	14,9	17,3	19,7	22,1
3	7,1	9,8	12,6	15,4	18,3	21,2	24,1	27,0
4	8,2	11,4	14,6	17,8	21,1	24,4	27,8	31,2
5	9,2	12,7	16,3	19,9	23,6	27,3	31,1	34,9
6	10,1	13,9	17,8	21,8	25,8	29,9	34,1	38,2
7	10,9	15,0	19,3	23,5	27,9	32,3	36,8	41,3
8	11,7	16,1	20,6	25,2	29,8	34,6	39,3	44,2
9	12,4	17,1	21,8	26,7	31,6	36,6	41,7	46,8
10	13,0	18,0	23,0	28,1	33,4	38,6	44,0	49,4
11	13,7	18,9	24,1	29,5	35,0	40,5	46,1	51,8
12	14,3	19,7	25,2	30,8	36,5	42,3	48,2	54,1
13	14,9	20,5	26,2	32,1	38,0	44,0	50,1	56,3
14	15,4	21,3	27,2	33,3	39,5	45,7	52,0	58,4
15	16,0	22,0	28,2	34,5	40,8	47,3	53,9	60,5
16	16,5	22,7	29,1	35,6	42,2	48,9	55,6	62,5

Tabulka 21.42: Kritické hodnoty $D_n(\alpha)$ pro jednovýběrový Kolmogorovův-Smirnovův test

n	$\alpha = 0,05$	$\alpha = 0,01$	n	$\alpha = 0,05$	$\alpha = 0,01$	n	$\alpha = 0,05$	$\alpha = 0,01$
1	,975 00	,995 00	31	,237 88	,285 30	61	,170 91	,205 06
2	,841 89	,929 29	32	,234 24	,280 94	62	,169 56	,203 43
3	,707 60	,829 00	33	,230 76	,276 77	63	,168 23	,201 84
4	,623 94	,734 24	34	,227 43	,272 79	64	,166 93	,200 29
5	,563 28	,668 53	35	,224 25	,268 97	65	,165 67	,198 77
6	,519 26	,616 61	36	,221 19	,265 32	66	,164 43	,197 29
7	,483 42	,575 81	37	,218 26	,261 80	67	,163 22	,195 84
8	,454 27	,541 79	38	,215 44	,258 43	68	,162 04	,194 42
9	,430 01	,513 32	39	,212 73	,252 05	69	,160 88	,193 03
10	,409 25	,488 93	40	,210 12	,252 05	70	,159 75	,191 67
11	,391 22	,467 70	41	,207 60	,249 04	71	,158 64	,190 34
12	,375 43	,449 05	42	,205 17	,246 13	72	,157 55	,189 03
13	,361 43	,432 47	43	,202 83	,243 32	73	,156 49	,187 76
14	,348 90	,417 62	44	,200 56	,240 60	74	,155 44	,186 50
15	,337 60	,404 20	45	,198 37	,237 98	75	,154 42	,185 28
16	,327 33	,392 01	46	,196 25	,235 44	76	,153 42	,184 08
17	,317 96	,380 86	47	,194 20	,232 98	77	,152 44	,182 90
18	,309 36	,370 62	48	,192 21	,230 59	78	,151 47	,181 74
19	,301 43	,361 17	49	,190 28	,228 28	79	,150 52	,180 60
20	,294 08	,352 41	50	,188 41	,226 04	80	,149 60	,179 49
21	,287 24	,344 27	51	,186 59	,223 86	81	,148 68	,178 40
22	,280 87	,336 66	52	,184 82	,221 74	82	,147 79	,177 32
23	,274 90	,329 54	53	,183 11	,219 68	83	,146 91	,176 27
24	,269 31	,322 86	54	,181 44	,217 68	84	,146 05	,175 23
25	,264 04	,316 57	55	,179 81	,215 74	85	,145 20	,174 21
26	,259 07	,310 64	56	,178 23	,213 84	86	,144 37	,173 21
27	,254 38	,305 02	57	,176 69	,211 99	87	,143 55	,172 23
28	,249 93	,299 71	58	,175 19	,210 19	90	,141 17	,169 38
29	,245 71	,294 66	59	,173 73	,208 44	95	,137 46	,164 93
30	,241 70	,289 87	60	,172 31	,206 73	100	,134 03	,160 81

Tabulka 21.43: Kritické hodnoty pro korelační koeficient r

n	$\alpha = 0,05$	$\alpha = 0,01$	n	$\alpha = 0,05$	$\alpha = 0,01$	n	$\alpha = 0,05$	$\alpha = 0,01$
3	0,996 9	0,999 9	14	0,532 4	0,661 4	25	0,396 1	0,505 2
4	0,950 0	0,990 0	15	0,514 0	0,641 1	30	0,361 0	0,462 9
5	0,878 3	0,958 7	16	0,497 3	0,622 6	35	0,333 8	0,429 6
6	0,811 4	0,917 2	17	0,482 2	0,605 5	40	0,312 0	0,402 6
7	0,754 5	0,874 5	18	0,468 3	0,589 7	45	0,294 0	0,380 1
8	0,706 7	0,834 3	19	0,455 5	0,575 1	50	0,278 7	0,361 0
9	0,666 4	0,797 7	20	0,443 8	0,561 4	60	0,254 2	0,330v1
10	0,631 9	0,764 6	21	0,432 9	0,548 7	70	0,235 2	0,306 0
11	0,602 1	0,734 8	22	0,422 7	0,536 8	80	0,235 2	0,286 4
12	0,576 0	0,707 9	23	0,412 3	0,525 6	90	0,207 2	0,270 2
13	0,552 9	0,683 5	24	0,404 4	0,515 1	100	0,196 6	0,256 5

Tabulka 21.44: Kritické hodnoty pro Spearmanův korelační koeficient

n	$\alpha = 0,05$	$\alpha = 0,01$	n	$\alpha = 0,05$	$\alpha = 0,01$	n	$\alpha = 0,05$	$\alpha = 0,01$
5	0,900 0	-	15	0,517 9	0,653 6	25	0,397 7	0,510 0
6	0,828 6	0,942 9	16	0,500 0	0,632 4	26	0,389 4	0,500 2
7	0,745 0	0,892 9	17	0,485 3	0,615 2	27	0,382 2	0,491 5
8	0,690 5	0,757 1	18	0,471 6	0,597 5	28	0,374 9	0,482 8
9	0,683 3	0,816 7	19	0,457 9	0,582 5	29	0,368 5	0,474 4
10	0,636 4	0,781 8	20	0,445 1	0,568 4	30	0,362 0	0,466 5
11	0,609 1	0,754 5	21	0,435 1	0,554 5			
12	0,580 4	0,727 3	22	0,424 1	0,542 6			
13	0,554 9	0,697 8	23	0,415 0	0,530 6			
14	0,534 1	0,674 7	24	0,406 1	0,520 0			

Literatura

- [1] Anděl, J. (1978). *Matematická statistika*. SNTL/ALFA.
- [2] Anděl, J. (2003). *Statistické metody*. Matfyzpress.
- [3] Anděl, J. (2005). *Základy matematické statistiky*. Matfyzpress.
- [4] Anděl, J. (2011). *Statistické metody*. Matfyzpress.
- [5] Budíková, M., Mikkoláš, Š., & Osecký, P. (2007). *Teorie pravděpodobnosti a matematická statistika* (3. vyd.). Masarykova univerzita, Přírodovědecká fakulta.
- [6] Büchter, A., & Henn, H.-W. (2007). *Elementare Stochastik*. Eine Einführung in die Mathematik der Daten und des Zufalls. Springer.
- [7] Denny, M., & Gaines, S. (2000). *Chance in Biology, Using probability to Explore Nature*. Princeton University Press.
- [8] Dupač, V., & Hušková, M. (1999). *Pravděpodobnost a matematická statistika*. Karolinum.
- [9] Hindls, R., Hronová, S., & Seger, J. (2004). *Statistika pro ekonomy*. Professional Publishing.
- [10] Hendl, J. (2015). *Přehled statistických metod*. Portál.
- [11] Hindls, R., Arltová, M., Hronová, S., Malá, I., Marek, L., Pecáková, I., & Řezanková, H. (2018). *Statistika v ekonomii*. Professional publishing.
- [12] Kahounová, J., & Hebák, P. (2014). *Počet pravděpodobnosti v příkladech*. Informatorium.
- [13] Mrkvička, T., & Petrášková, V. (2006). *Úvod do statistiky*. Jihočeská univerzita v Českých Budějovicích.

- [14] Mrkvička, T., & Petrášková, V. (2008). *Úvod do teorie pravděpodobnosti*. Jihočeská univerzita v Českých Budějovicích.
- [15] Neubauer, J., Sedlačík, M., & Kříž, O. (2021). *Základy statistiky*. Grada.
- [16] Płocki, A., & Tlustý, P. (2007). *Pravděpodobnost a statistika pro začátečníky a mírně pokročilé*. Prometheus.
- [17] Potocký, R. (1986). *Zbierka úloh z pravdepodobnosti a matematickej štatistiky*. Alfa, vydavateľstvo technickej a ekonomickej literatúry.
- [18] Prášilová, M., & Svatošová, L. (1997). *Cvičení ze statistiky*. Česká zemědělská univerzita v Praze, Provozně ekonomická fakulta.
- [19] Prášilová, M., & Zeipelt, R. (2001). *Cvičení ze statistiky I*. Česká zemědělská univerzita v Praze, Provozně ekonomická fakulta.
- [20] Stocker, T. C., & Steinke, I. (2017). *Statistik. Übungsbuch*, DE GRUYTER Oldenbourg.

**Pravděpodobnost a statistika s řešenými příklady v prostředí
MS Excel**

Prof. RNDr. Tomáš Mrkvička, Ph.D.
Doc. RNDr. Vladimíra Petrášková, Ph.D.
Mgr. Michal Houda, Ph.D.
RNDr. PhDr. Jan Fiala, Ph.D.

Vytištěno v České republice





Pedagogická
fakulta
Faculty
of Education

Jihočeská univerzita
v Českých Budějovicích
University of South Bohemia
in České Budějovice

ISBN 978-80-7694-135-9



9 788076 941359