Masarykova univerzita v Brně
Ekonomicko–správní fakulta

Statistika

distanční studijní opora

Marie Budíková

Brno 2004

Tento projekt byl realizován za finanční podpory Evropské unie v rámci programu SOCRATES — Grundtvig.

Za obsah produktu odpovídá výlučně autor, produkt nereprezentuje názory Evropské komise a Evropská komise neodpovídá za použití informací, jež jsou obsahem produktu.

This project was realized with financial support of European Union in terms of program SOCRATES — Grundtvig.

Author is exclusively responsible for content of product, product does not represent opinions of European Union and European Commission is not responsible for any uses of informations, which are content of product

Statistika
Vydala Masarykova univerzita v Brně
Ekonomicko–správní fakulta

Vydání pilotní verze
Brno, 2004

RNDr. Marie Budíková, Dr.

Publikace neprošla jazykovou úpravou

Identifikace modulu

Znak

- KMSTAT

Název

- Statistika

Garant/autor

- RNDr. Marie Budíková, Dr.

Cíl

Vymezení cíle

Statistika jako metoda analýzy dat patří k vědním disciplínám, v nichž by měl být vzdělán každý ekonom. Její role v ekonomii je zcela nezastupitelná, neboť moderní řízení je založeno na nepřetržitém vyhodnocování informací o hospodářství jako celku i jeho subsystémech, a tyto informace poskytuje a následně zpracovává právě statistika.

Přiměřená znalost základních statistických pojmů je pro ekonoma důležitá také proto, že mu pomáhá porozumět odborné ekonomické literatuře, jejíž některé části statistiku v hojné míře využívají.

Význam statistiky v poslední době neustále roste, což úzce souvisí s rozvojem výpočetní techniky, která je používána jak při sběru a přenosu dat, tak při jejich zpracování a uklá-dání informací.

Dovednosti a znalosti získané po studiu textů

Předmět „Statistika“ vás má především naučit zpracovávat data, která se týkají ekonomických jevů, tj. data třídit, numericky vyhodnocovat a interpretovat. Velké množství příkladů, které jsou součástí učebního textu, vám pomůže při formulování vlastních úloh a výběru správné metody. Naučíte se rovněž využívat výpočetní techniku při řešení ekonomických problémů.

Časový plán

Časová náročnost

- prezenční část 22%
- samostudium 78%

Celkový studijní čas

- 14 týdnů

Harmonogram

- přednášky 24 hodin
- samostudium a práce s počítačem 85 hodin

Způsob studia

Studijní pomůcky

doporučená literatura:

[1] Anděl J.: *Matematická statistika*. SNTL/Alfa Praha 1978.
[2] Arltová M., Bílková D., Jarošová E., Pourová Z.: *Sbírka příkladů ze statistiky (Statistika A)*. VŠE Praha 1996.
[3] Budíková M., Mikoláš Š., Osecký P.: *Popisná statistika*. MU Brno 2001.
[4] Budíková M., Mikoláš Š., Osecký P.: *Teorie pravděpodobnosti a matematická statistika*. Sbírka příkladů. MU Brno 2001.
[5] Hebák P., Kahounová J.: *Počet pravděpodobnosti v příkladech*. SNTL Praha 1978.
[6] Karpíšek Z.: *Pravděpodobnostní metody*. VUT Brno 2000.
[7] Karpíšek Z., Drdla M.: *Statistické metody*. VUT Brno 1999.
[8] Novovičová J.: *Pravděpodobnost a matematická statistika*. ČVUT Praha 2002.
[9] Stuchlý J.: *Statistika I. Cvičení ze statistických metod pro managery*. VŠE Praha 1999.

Vybavení

- PC
- CD-ROM

Návod práce se studijními texty

Text je rozvržen do 13 kapitol a 2 příloh. 1. až 4. kapitola se zabývají popisnou statistikou. Popisná statistika je disciplína, která pomocí různých tabulek, grafů, funkcionálních a číselných charakteristik sumarizuje informace obsažené ve velkém množství dat. Používá jen základní matematické operace a lze ji snadno pochopit. Její důležitost spočívá jednak v tom, že se v praxi velmi často používá a jednak motivuje pojmy, které jsou potřeba v počtu pravděpodobnosti.

5. až 10. kapitola vás seznámí s počtem pravděpodobnosti, který se zabývá studiem zákonitostí v náhodných pokusech. Matematickými prostředky modeluje situace, v nichž hraje roli náhoda. Pod pojmem náhoda rozumíme působení faktorů, které se živelně mění při různých provedeních téhož pokusu a nepodléhají naší kontrole.

11. až 13. kapitola obsahují základní poznatky o matematické statistice. Matematická statistika je věda, která analyzuje a interpretuje data především za účelem získání předpovědi a zlepšení rozhodování v různých oblastech lidské činnosti. Při tom se řídí principem statistické indukce: na základě znalostí o náhodném výběru z určitého rozložení pravděpodobností se snaží odvodit vlastnosti tohoto rozložení pravděpodobností.

Příloha A je tvořena vybranými statistickými tabulkami, konkrétně obsahuje hodnoty distribuční funkce standardizovaného normálního rozložení, kvantily standardizovaného normálního rozložení, Pearsonova rozložení $\chi^2(n)$, Studentova rozložení $t(n)$ a Fisherova-Snedecorova rozložení $F(n_1, n_2)$. Příloha B pak obsahuje informace o programovém systému STATISTICA a podrobné návody na jeho použití.

V úvodu 1. až 13. kapitoly je vždy vymezen cíl kapitoly a je uvedena časová zátěž, která je potřebná ke zvládnutí příslušné kapitoly. Kapitoly jsou uzavřeny stručným shrnutím probrané látky a kontrolními otázkami a úkoly. Ty úkoly, jejichž řešení je nutné či alespoň vhodné provádět pomocí systému STATISTICA, jsou označeny (S). Výsledky úkolů můžete porovnat s výsledky, k nimž dospěla autorka učebního textu.

1. až 13. kapitola jsou uspořádány v logickém sledu. Do přílohy A budete nahlížet podle potřeby a příloha B vám poslouží rovněž průběžně.

Obsah

Obsah

1. Základní, výběrový a datový soubor 13
2. Bodové a intervalové rozložení četností 21
3. Číselné charakteristiky znaků 39
4. Regresní přímka 49
5. Jev a jeho pravděpodobnost 57
6. Stochasticky nezávislé jevy a podmíněná pravděpodobnost 65
7. Náhodná veličina a její distribuční funkce 71
8. Vybraná rozložení diskrétních a spojitých náhodných veličin 85
9. Číselné charakteristiky náhodných veličin 97
10. Zákon velkých čísel a centrální limitní věta 111
11. Základní pojmy matematické statistiky 117
12. Bodové a intervalové odhady parametrů a parametrických funkcí 123
13. Úvod do testování hypotéz a testy o parametrech normálního rozložení 137

Příloha A – Statistické tabulky 147

Příloha B – Základní informace o programu STATISTICA 6 163

Úvod

Proč se zabývat statistikou?

Statistika je metoda analýzy dat, která nachází široké uplatnění v celé řadě ekonomických, technických, přírodovědných a humanitních disciplín. Její význam v poslední době neustále roste, což úzce souvisí s rozvojem výpočetní techniky, která je používána jak při sběru a přenosu dat, tak při jejich zpracování a ukládání informací.

Role statistiky v ekonomii je zcela nezastupitelná, neboť moderní řízení je založeno na nepřetržitém vyhodnocování informací o hospodářství jako celku i jeho subsystémech, a tyto informace poskytuje a následně zpracovává právě statistika.

Přiměřená znalost základních statistických pojmů je pro ekonoma důležitá také proto, že mu pomáhá porozumět odborné ekonomické literatuře, jejíž některé části statistiku v hojné míře využívají.

Aplikovat statistiku znamená shromažďovat data o studovaných jevech a zpracovávat je, tj. třídit, numericky vyhodnocovat a interpretovat. Statistika se tak pro ekonoma ocitá v těsném sousedství informatiky a výpočetní techniky a je připravena řešit ekonomické problémy pomocí kvantitativní analýzy dat.

Způsob studia

Co lze očekávat od tohoto textu?

V předmětu „Statistika“ se budeme zabývat třemi oblastmi statistiky, a to popisnou statistikou, počtem pravděpodobnosti a matematickou statistikou.

Popisná statistika je disciplína, která pomocí různých tabulek, grafů, funkcionálních a číselných charakteristik sumarizuje informace obsažené ve velkém množství dat. Používá jen základní matematické operace a lze ji snadno pochopit. Její důležitost spočívá jednak v tom, že se v praxi velmi často používá a jednak motivuje pojmy, které jsou potřeba v počtu pravděpodobnosti.

Počet pravděpodobnosti se zabývá studiem zákonitostí v náhodných pokusech. Matematickými prostředky modeluje situace, v nichž hraje roli náhoda. Pod pojmem náhoda rozumíme působení faktorů, které se živelně mění při různých provedeních téhož pokusu a nepodléhají naší kontrole.

Matematická statistika je věda, která analyzuje a interpretuje data především za účelem získání předpovědi a zlepšení rozhodování v různých oblastech lidské činnosti. Při tom se řídí principem statistické indukce: na základě znalostí o náhodném výběru z určitého rozložení pravděpodobností se snaží odvodit vlastnosti tohoto rozložení pravděpodobností.

K úspěšnému zvládnutí předmětu „Statistika“ je zapotřebí ovládat kombinatoriku, základy diferenciálního a integrálního počtu jedné a dvou proměnných a znát základy práce s osobním počítačem.

Velmi účinným prostředkem pro řešení statistických úloh je programový systém STATISTICA, jehož instalační CD je součástí studijních materiálů. Informace o tomto systému a podrobné návody na jeho použití jsou uvedeny v příloze B studijních materiálů. Příklady či úkoly, jejichž řešení je nutné či alespoň vhodné provádět pomocí systému STATISTICA, jsou označeny (S).

Příloha A obsahuje vybrané statistické tabulky, konkrétně hodnoty distribuční funkce standardizovaného normálního rozložení, kvantily standardizovaného normálního rozložení, Pearsonova rozložení $\chi^2(n)$, Studentova rozložení $t(n)$ a Fisherova-Snedecorova rozložení $F(n_1, n_2)$. Všechny tyto tabelované hodnoty (a samozřejmě mnohé další) lze získat pomocí systému STATISTICA.

1. Základní, výběrový a datový soubor

Cíl kapitoly

Po prostudování této kapitoly budete umět:

- vymezit základní soubor a jeho objekty
- stanovit výběrový soubor
- spočítat absolutní a relativní četnosti množin ve výběrovém souboru a znát vlastnosti relativní četnosti a podmíněné relativní četnosti
- ověřit četnostní nezávislost dvou množin ve výběrovém souboru
- vytvořit datový soubor
- uspořádat jednorozměrný datový soubor a stanovit vektor variant
- vypočítat absolutní a relativní četnost jevu ve výběrovém souboru

Časová zátěž

Pro zvládnutí této kapitoly budete potřebovat 4 – 5 hodin studia.

Nejprve se seznámíme s definicí základního a výběrového souboru a pojmem absolutní a relativní četnosti množiny v daném výběrovém souboru. Uvedeme příklad, s jehož různými variantami se budeme setkávat ve všech kapitolách věnovaných popisné statistice. Rovněž shrneme vlastnosti relativní četnosti.

1.1. Definice
Základním souborem rozumíme libovolnou neprázdnou množinu E. Její prvky značíme ε a nazýváme je objekty. Libovolnou neprázdnou podmnožinu $\{\varepsilon_1, \dots, \varepsilon_n\}$ základního souboru E nazýváme *výběrový soubor rozsahu n*. Je-li $G \subseteq E$, pak symbolem $N(G)$ rozumíme *absolutní četnost* množiny G ve výběrovém souboru, tj. počet těch objektů množiny G, které patří do výběrového souboru. *Relativní četnost* množiny G ve výběrovém souboru zavedeme vztahem

$$p(G) = \frac{N(G)}{n}.$$

1.2. Příklad
Základním souborem E je množina všech ekonomicky zaměřených studentů 1. ročníku českých vysokých škol. Množina G_1 je tvořena těmi studenty, kteří uspěli v prvním zkušebním termínu z matematiky a množina G_2 obsahuje ty studenty, kteří uspěli v prvním zkušebním termínu z angličtiny. Ze základního souboru bylo náhodně vybráno 20 studentů, kteří tvoří výběrový soubor $\{\varepsilon_1, \dots, \varepsilon_{20}\}$. Z těchto 20 studentů 11 uspělo v matematice, 15 v angličtině a 11 v obou předmětech. Zapište absolutní a relativní četnosti úspěšných matematiků, angličtinářů a oboustranně úspěšných studentů.

Řešení:

$$N(G_1) = 12, \quad N(G_2) = 15, \quad N(G_1 \cap G_2) = 11, \quad n = 20$$
$$p(G_1) = \frac{12}{20} = 0{,}6, \quad p(G_2) = 0{,}75, \quad p(G_1 \cap G_2) = \frac{11}{20} = 0{,}55$$

Vidíme, že úspěšných matematiků je 60%, angličtinářů 75% a oboustranně úspěšných studentů jen 55%.

1.3. Věta
Relativní četnost má následujících 12 vlastností, které jsou obdobné vlastnostem procent.

- $p(\emptyset) = 0$
- $p(G) \geq 0$
- $p(G_1 \cup G_2) + p(G_1 \cap G_2) = p(G_1) + p(G_2)$
- $1 + p(G_1 \cap G_2) \geq p(G_1) + p(G_2)$
- $p(G_1 \cup G_2) \leq p(G_1) + p(G_2)$
- $G_1 \cap G_2 = \emptyset \;\Rightarrow\; p(G_1 \cup G_2) = p(G_1) + p(G_2)$
- $p(G_2 - G_1) = p(G_2) - p(G_1 \cap G_2)$
- $G_1 \subseteq G_2 \;\Rightarrow\; p(G_2 - G_1) = p(G_2) - p(G_1)$
- $G_1 \subseteq G_2 \;\Rightarrow\; p(G_1) \leq p(G_2)$
- $p(E) = 1$
- $p(G) + p(\overline{G}) = 1$
- $p(G) \leq 1$

Pokud se v daném základním souboru zajímáme o dvě podmnožiny, můžeme zavést pojem podmíněné relativní četnosti jedné podmnožiny v daném výběrovém souboru za předpokladu, že objekt pochází z druhé podmnožiny. V následujícím příkladu vypočteme podmíněné relativní četnosti úspěšných matematiků mezi úspěšnými angličtináři a naopak.

1.4. Definice
Nechť E je základní soubor, G_1, G_2 jeho podmnožiny, $\{\varepsilon_1, \ldots, \varepsilon_n\}$ výběrový soubor. Definujeme *podmíněnou relativní četnost* množiny G_1 ve výběrovém souboru za předpokladu G_2:

$$p(G_1|G_2) = \frac{N(G_1 \cap G_2)}{N(G_2)} = \frac{p(G_1 \cap G_2)}{p(G_2)}$$

a *podmíněnou relativní četnost* G_2 ve výběrovém souboru za předpokladu G_1:

$$p(G_2|G_1) = \frac{N(G_1 \cap G_2)}{N(G_1)} = \frac{p(G_1 \cap G_2)}{p(G_1)}.$$

1.5. Příklad
Pro údaje z příkladu 1.2 vypočtěte podmíněnou relativní četnost úspěšných matematiků mezi úspěšnými angličtináři a podmíněnou relativní četnost úspěšných angličtinářů mezi úspěšnými matematiky.

Řešení:
$p(G_1|G_2) = \frac{11}{15} = 0{,}73$ (tzn., že 73% těch studentů, kteří byli úspěšní v angličtině, uspělo i v matematice)

$p(G_2|G_1) = \frac{11}{12} = 0{,}92$ (tzn., že 92% těch studentů, kteří byli úspěšní v matematice, uspělo i v angličtině)

Nyní se naučíme, jak ověřovat četnostní nezávislost dvou množin v daném výběrovém souboru. Znamená to, že informace o původu objektu z jedné množiny nijak nemění šance, s nimiž soudíme na jeho původ i z druhé množiny. Ověříme, zda úspěch v matematice a angličtině jsou v daném výběrovém souboru četnostně nezávislé.

1.6. Definice
Řekneme, že množiny G_1, G_2 jsou *četnostně nezávislé* v daném výběrovém souboru, jestliže

$$p(G_1 \cap G_2) = p(G_1) \cdot p(G_2).$$

(V praxi jen zřídka dojde k tomu, že uvedený vztah platí přesně. Většinou je jen naznačena určitá tendence četnostní nezávislosti.)

1.7. Příklad
Pro údaje z příkladu 1.2 zjistěte, zda úspěchy v matematice a angličtině jsou v daném výběrovém souboru četnostně nezávislé.

Řešení:

$$p(G_1 \cap G_2) = 0{,}55, \quad p(G_1) \cdot p(G_2) = 0{,}6 \cdot 0{,}75 = 0{,}45,$$

tedy skutečná relativní četnost oboustranně úspěšných studentů je větší než by odpovídalo četnostní nezávislosti množin G_1, G_2 v daném výběrovém souboru.

Nyní každý objekt základního souboru ohodnotíme jedním nebo více čísly pomocí funkce, která se nazývá znak. Čísla, která se vztahují pouze k objektům výběrového souboru sestavíme do matice zvané datový soubor. Vystvětlíme si, co to je uspořádaný datový soubor a vektor variant. Uvedené pojmy objasníme na příkladu.

1.8. Definice
Nechť E je základní soubor. Potom funkce $X: E \to \mathbb{R}$, $Y: E \to \mathbb{R}, \dots,$ $Z: E \to \mathbb{R}$, které každému objektu přiřazují číslo, se nazývají *(skalární) znaky*. Uspořádaná p–tice $(X, Y, \dots, Z)$ se nazývá vektorový znak.

1.9. Definice
Nechť je dán výběrový soubor $\{\varepsilon_1, \dots, \varepsilon_n\} \subseteq E$. Hodnoty znaků $X, Y, \dots, Z$ pro i–tý objekt označíme $x_i = X(\varepsilon_i), y_i = Y(\varepsilon_i), \dots, z_i = Z(\varepsilon_i)$, $i = 1, \dots, n$. Matice

$$\begin{bmatrix} x_1 & y_1 & \dots & z_1 \\ x_2 & y_2 & \dots & z_2 \\ \vdots & \vdots & \vdots & \vdots \\ x_n & y_n & \dots & z_n \end{bmatrix}$$

typu $n \times p$ se nazývá datový soubor. Její řádky odpovídají jednotlivým objektům, sloupce znakům.
Libovolný sloupec této matice nazýváme *jednorozměrným datovým souborem*. Jestliže uspořádáme hodnoty některého znaku (např. znaku X) v jednorozměrném datovém souboru vzestupně podle velikosti, dostaneme *uspořádaný datový soubor*

$$\begin{bmatrix} x_{(1)} \\ \vdots \\ x_{(n)} \end{bmatrix},$$

kde $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. Vektor

$$\begin{bmatrix} x_{[1]} \\ \vdots \\ x_{[n]} \end{bmatrix},$$

kde $x_{[1]} < \cdots < x_{[r]}$ jsou navzájem různé hodnoty znaku X, se nazývá *vektor variant*.

1.10. Příklad
Pro studenty z výběrového souboru u vedeného v příkladu 1.2 byly zjišťovány hodnoty znaků X – známka z matematiky v prvním zkušebním termínu, Y – známka z angličtiny v prvním zkušebním termínu, Z – pohlaví studenta ($0 \ldots$ žena, $1 \ldots$ muž). Byl získán datový soubor

$$\begin{bmatrix} 2 & 2 & 0 \\ 1 & 3 & 1 \\ 4 & 3 & 1 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \\ 4 & 4 & 1 \\ 3 & 3 & 1 \\ 3 & 4 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 4 & 2 & 1 \\ 4 & 4 & 0 \\ 2 & 2 & 0 \\ 4 & 3 & 1 \\ 2 & 3 & 1 \\ 4 & 4 & 0 \\ 1 & 1 & 0 \\ 4 & 3 & 1 \\ 4 & 4 & 1 \\ 1 & 3 & 0 \end{bmatrix}$$

Utvořte jednorozměrný uspořádaný i neuspořádaný datový soubor pro známky z matematiky a vektory variant pro známky z matematiky.

Řešení:

$$\begin{bmatrix}2\\1\\4\\1\\1\\4\\3\\3\\1\\1\\4\\4\\2\\4\\2\\4\\1\\4\\4\\1\end{bmatrix}, \quad \begin{bmatrix}1\\1\\1\\1\\1\\1\\1\\2\\2\\2\\3\\3\\4\\4\\4\\4\\4\\4\\4\\4\end{bmatrix}, \quad \begin{bmatrix}1\\2\\3\\4\end{bmatrix}.$$

V závěrečné partii této kapitoly se seznámíme s pojmem jevu a jeho absolutní a relativní četnosti. V následujícím příkladu vypočítáme konkrétní absolutní a relativní četnosti několika jevů.

1.11. Definice

Nechť $\{\varepsilon_1, \ldots, \varepsilon_n\}$ je výběrový soubor, $X, Y, \ldots, Z$ jsou znaky, $B, B_1, \ldots, B_p$ jsou číselné množiny. Zápis $\{X \in B\}$ znamená jev „*znak X nabyl hodnoty z množiny B*“ a zápis $\{X \in B_1 \wedge Y \in B_2 \wedge \ldots Z \in B_p\}$ znamená jev „*znak X nabyl hodnoty z množiny B_1 a současně znak Y nabyl hodnoty z množiny B_2 atd. až znak Z nabyl hodnoty z množiny B_p*“. Symbol $N(X \in B)$ značí absolutní četnost jevu $X \in B$ ve výběrovém souboru, tj. počet těch objektů ve výběrovém souboru, pro něž $x_i \in B$. Symbol $p(X \in B)$ znamená relativní četnost jevu $\{X \in B\}$ ve výběrovém souboru, tj.

$$p(X \in B) = \frac{N(X \in B)}{n}.$$

Analogicky $N(X \in B_1 \wedge Y \in B_2 \wedge \cdots \wedge Z \in B_p)$ resp. $p(X \in B_1 \wedge Y \in B_2 \wedge \cdots \wedge Z \in B_p)$ znamená absolutní resp. relativní četnost jevu $\{X \in B_1 \wedge Y \in B_2 \wedge \cdots \wedge Z \in B_p\}$ ve výběrovém souboru.

1.12. Příklad

Pro datový soubor z příkladu 1.10 najděte relativní četnost

a) matematických jedničkářů,
b) úspěšných matematiků,

c) oboustranně neúspěšných studentů.

Řešení:

ad a) $p(X = 1) = \frac{7}{20} = 0{,}35$; ad b) $p(X \leq 3) = \frac{12}{20} = 0{,}60$;
ad c) $p(X = 4 \wedge Y = 4) = \frac{4}{20} = 0{,}20$.

Shrnutí kapitoly

Předmětem statistického zájmu není jednotlivý objekt, nýbrž soubor objektů, tzv. základní soubor. Zpravidla není možné vyšetřovat všechny objekty, ale jenom určitý počet objektů, které tvoří výběrový soubor. Ty prvky základního souboru, které vykazují určitou společnou vlastnost, tvoří množinu. Statistik zkoumá absolutní a relativní četnost množiny v daném výběrovém souboru. Zajímají-li nás ve výběrovém souboru dvě množiny, můžeme zkoumat výskyty objektů z jedné množiny mezi objekty pocházejícími z druhé množiny. Tím dospíváme k pojmu podmíněné relativní četnosti. Rovněž lze ověřovat četnostní nezávislost těchto dvou množin v daném výběrovém souboru. Četnostní nezávislost vlastně znamená, že informace o původu objektu z jedné množiny nijak nemění šance, s nimiž soudíme na jeho původ z druhé množiny. Každému objektu základního souboru lze pomocí funkce zvané znak přiřadit číslo (nebo i více čísel). Pokud hodnoty znaků pro objekty daného výběrového souboru uspořádáme do matice, dostáváme datový soubor. Libovolný sloupec této matice tvoří jednorozměrný datový soubor, který můžeme uspořádat podle velikosti a vytvořit tak uspořádaný datový soubor nebo z něj získat vektor variant. Jevem rozumíme skutečnost, že znak nabyl hodnoty z nějaké číselné množiny. Můžeme zkoumat absolutní a relativní četnost jevu v daném výběrovém souboru.

Kontrolní otázky a úkoly

1 Uveďte příklad základního souboru z ekonomické praxe.

2 Nechť množiny G_1, G_2 jsou neslučitelné, $p(G_1) = 0{,}27$, $p(G_1 \cup G_2) = 0{,}75$. Vypočtěte $p(G_2)$.

3 Nechť $G_1 \subseteq G_2$, $p(G_1) = 0{,}33$, $p(G_2 - G_1) = 0{,}15$. Vypočtěte $p(G_2)$.

4 Nechť $p(G_1 - G_2) = 0{,}36$, $p(G_1 \cap G_2) = 0{,}12$. Vypočtěte $p(G_2)$.

5 Je dán dvourozměrný datový soubor

$$\begin{bmatrix} 2 & 1 \\ 2 & 0 \\ 1 & 0 \\ 4 & 2 \\ 4 & 2 \\ 3 & 2 \\ 3 & 1 \\ 5 & 3 \\ 5 & 2 \\ 2 & 0 \end{bmatrix}$$

Znak X znamená počet členů domácnosti a znak Y počet dětí do 15 let v této domácnosti.

a) Utvořte uspořádané datové soubory pro znaky X a Y.
b) Najděte vektory variant znaků X a Y.
c) Vypočtěte relativní četnost tříčlenných domácností.
d) Vypočtěte relativní četnost nejvýše tříčlenných domácností.
e) Vypočtěte relativní četnost bezdětných domácností.
f) Vypočtěte relativní četnost dvoučlenných bezdětných domácností.
g) Vypočtěte podmíněnou relativní četnost dvoučlenných domácností, které jsou bezdětné.

2. Bodové a intervalové rozložení četností

Cíl kapitoly

Po prostudování této kapitoly budete umět:

- konstruovat diagramy znázorňující rozložení četností
- vytvářet tabulky četností
- sestrojit grafy četnostní funkce, empirické distribuční funkce, hustoty četnosti a empirické intervalové distribuční funkce

Časová zátěž

Pro zvládnutí této kapitoly budete potřebovat 7 – 8 hodin studia.

Nejprve se seznámíme s bodovým rozložením četností a ukážeme si, jak pomocí různých diagramů graficky znázornit bodové rozložení četností. Pro datový soubor známek z matematiky a angličtiny pak vytvoříme několik typů diagramů.

2.1. Definice
Nechť je dán jednorozměrný datový soubor. Jestliže počet variant znaku X není příliš velký, pak přiřazujeme četnosti jednotlivým variantám a hovoříme o *bodovém rozložení četností*.

2.2. Definice
Existuje několik způsobů, jak graficky znázornit bodové rozložení četností.

Tečkový diagram: na číselné ose vyznačíme jednotlivé varianty znaku X a nad každou variantu nakreslíme tolik teček, jaká je její absolutní četnost.

Polygon četnosti: je lomená čára spojující body, jejichž x–ová souřadnice je varianta znaku X a y–ová souřadnice je absolutní četnost této varianty.

Sloupkový diagram: je soustava na sebe nenavazujících obdélníků, kde střed základny je varianta znaku X a výška je absolutní četnost této varianty.

Výsečový graf: je kruh rozdělený na výseče, jejichž vnější obvod odpovídá absolutním četnostem variant znaku X.

Dvourozměrný tečkový diagram: na vodorovnou osu vyneseme varianty znaku X, na svislou varianty znaku Y a do příslušných průsečíků nakreslíme tolik teček, jaká je absolutní četnost dané dvojice.

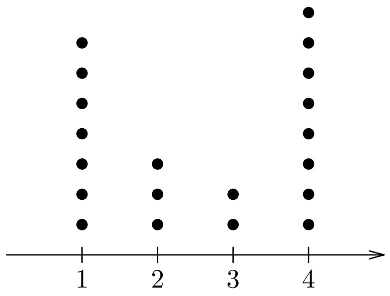
2.3. Příklad
Pro datový soubor z příkladu 1.10 sestrojte

a) jednorozměrné tečkové diagramy pro znak X a znak Y,
b) polygony četností pro znak X a znak Y,
c) sloupkové diagramy pro znak X a znak Y,
d) výsečové diagramy pro znak X a znak Y,
e) dvourozměrný tečkový diagram pro vektorový znak (X,Y),

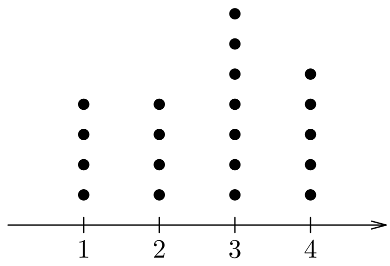
Řešení:

ad a)

Známka z matematiky

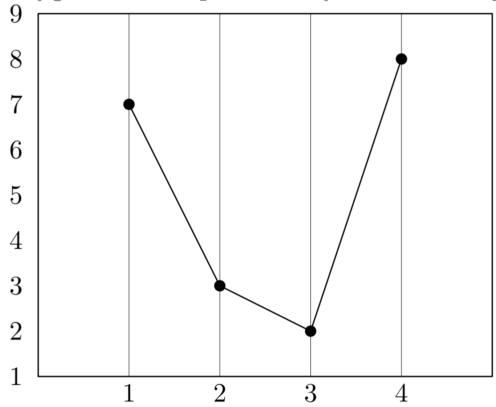


Známka z angličtiny

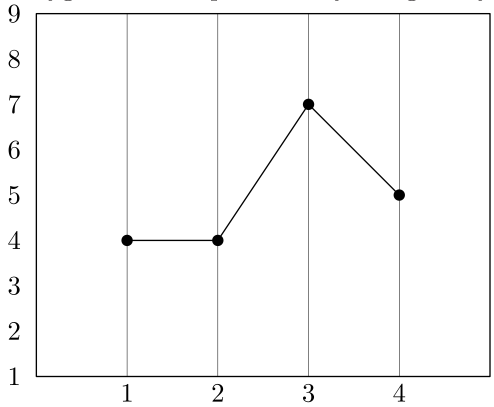


ad b)

Polygon četnosti pro známky z matematiky

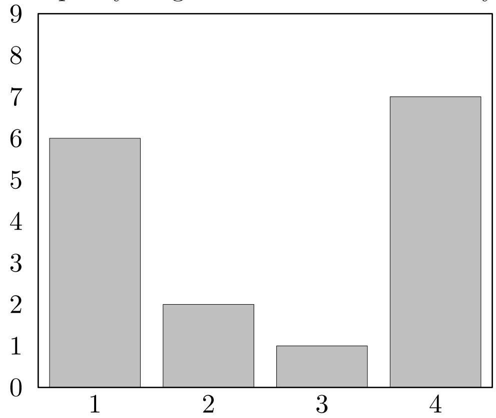


Polygon četnosti pro známky z angličtiny

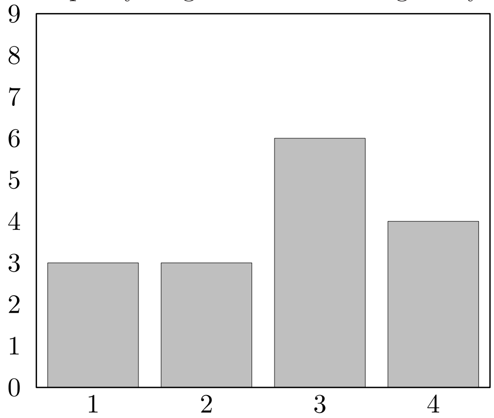


ad c)

Sloupkový diagram známek z matematiky

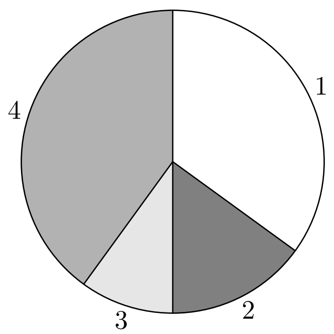


Sloupkový diagram známek z angličtiny

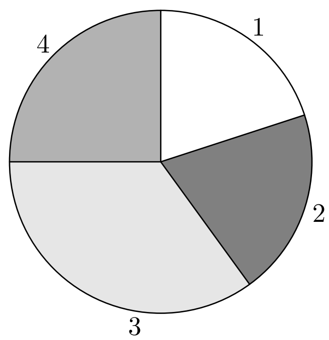


ad d)

Výsečový diagram známek z matematiky

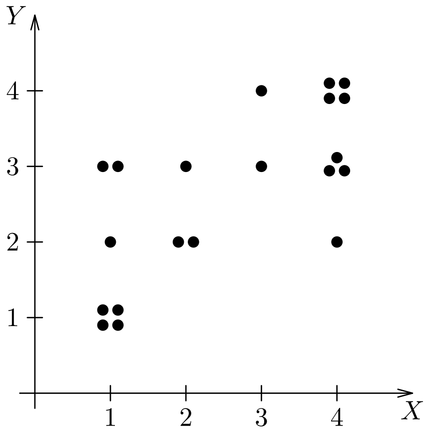


Výsečový diagram známek z angličtiny



Ze všech těchto diagramů je vidět odlišný přístup zkoušejících ke studentům. Matematik nešetří jedničkami, ale místo trojky raději rovnou dává čtyřku. Naproti tomu angličtinář považuje trojku za typickou studentskou známku.

ad e)



Dvourozměrný tečkový diagram svědčí o nepříliš výrazné tendenci k podobné klasifikaci v obou předmětech. Můžete si zkusit nakreslit dvourozměrné tečkové diagramy zvlášť pro muže a zvlášť pro ženy. Zjistíte, že u žen je tendence k podobným známkám daleko silnější než u mužů.

Bodové rozložení četností lze znázornit nejenom graficky, ale též tabulkou zvanou variační řada, která obsahuje absolutní a relativní četnosti jednotlivých variant znaku v daném výběrovém souboru a též absolutní a relativní kumulativní četnosti. Pomocí relativních četností se zavádí četnostní funkce, pomocí relativních kumulativních četností empirická distribuční funkce (je pro ni typické, že má schodovitý průběh). Tyto pojmy objasníme na příkladu známek z matematiky a uvedeme rovněž vlastnosti obou výše zmíněných funkcí.

2.4. Definice
Nechť je dán jednorozměrný datový soubor, v němž znak X nabývá r variant. Pro $j = 1, \dots, r$ definujeme:

absolutní četnost varianty $x_{[j]}$ ve výběrovém souboru

$$n_j = N(X = x_{[j]})$$

relativní četnost varianty $x_{[j]}$ ve výběrovém souboru

$$p_j = \frac{n_j}{n}$$

absolutní kumulativní četnost prvních j *variant* ve výběrovém souboru

$$N_j = N(X \leq x_{[j]}) = n_1 + \dots + n_j$$

relativní kumulativní četnost prvních j *variant* ve výběrovém souboru

$$F_j = \frac{N_j}{n} = p_1 + \dots + p_j$$

Tabulka typu

$x_{[j]}$	n_j	p_j	N_j	F_j
$x_{[1]}$	n_1	p_1	N_1	F_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_{[r]}$	n_r	p_r	N_r	F_r

se nazývá *variační řada.*

Funkce

$$p(x) = \begin{cases} p_j & \text{pro } x = x_{[j]},\ j = 1, \dots, r \\ 0 & \text{jinak} \end{cases}$$

se nazývá *četnostní funkce.*

Funkce

$$F(x) = \begin{cases} 0 & \text{pro } x < x_{[1]} \\ F_j & \text{pro } x_{[j]} \le x < x_{[j+1]},\ j = 1, \dots, r-1 \\ 1 & \text{pro } x \ge x_{[r]} \end{cases}$$

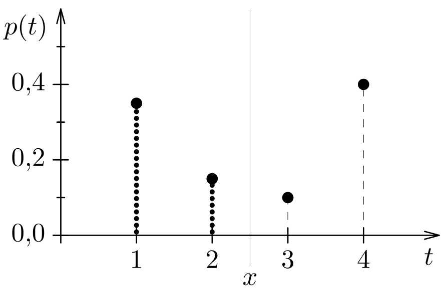
se nazývá *empirická distribuční funkce.*

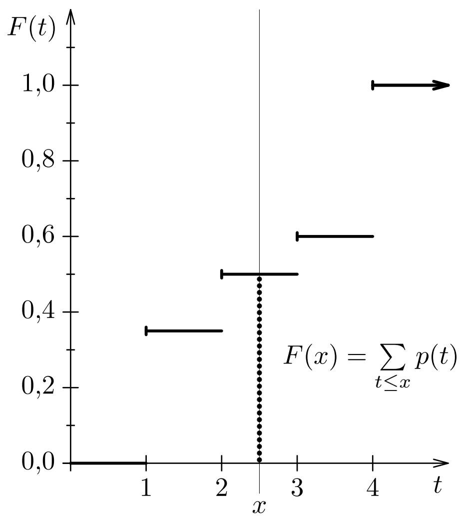
2.5. Příklad

Pro datový soubor z příkladu 1.10 sestavte variační řadu pro znak X. Nakreslete grafy četnostní funkce a empirické distribuční funkce.

Řešení:

$x_{[j]}$	n_j	p_j	N_j	F_j
1	7	0,35	7	0,35
2	3	0,15	10	0,50
3	2	0,10	12	0,60
4	8	0,40	20	1,00
–	20	1,00	–	–





2.6. Věta
Četnostní funkce je nezáporná ($\forall x \in \mathbb{R} : p(x) \leq 0$) a normovaná, tj.

$$\sum_{x=-\infty}^{\infty} p(x) = 1.$$

Empirická distribuční funkce je neklesající, tzn.

$$\forall x_1, x_2 \in \mathbb{R},\ x_1 < x_2 :\ F(x_1) \leq F(x_2),$$

zprava spojitá ($\forall x_0 \in \mathbb{R}$ libovolné, ale pevně dané: $\lim\limits_{x \to x_0} F(x) = F(x_0)$) a normovaná ($\lim\limits_{x \to -\infty} F(x) = 0,\ \lim\limits_{x \to \infty} F(x) = 1$).

Nyní se budeme zabývat dvourozměrných datovým souborem. Zavedeme simultánní absolutní a relativní četnosti pro dvojice variant znaků X a Y a ukážeme souvislost mezi simultánními a marginálními četnostmi. Budeme definovat podmíněné relativní četnosti. Vysvětlíme si, jak se uvedené četnosti zapisují do kontingenčních tabulek. Pomocí simultánních relativních četností zavedeme simultánní četnostní funkci, seznámíme se s jejími vlastnostmi a ukážeme vztah mezi simultánní četnostní funkcí a marginálními četnostními funkcemi. Zavedeme pojem četnostní nezávislosti znaků v daném výběrovém souboru. Se všemi uvedenými pojmy se naučíme pracovat v příkladu se známkami z matematiky a angličtiny.

2.7. Definice
Nechť je dán dvourozměrný datový soubor

$$\begin{bmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix},$$

kde znak X má r variant a znak Y má s variant. Pak definujeme:

simultánní absolutní četnost dvojice $(x_{[j]}, y_{[k]})$ ve výběrovém souboru

$$n_{jk} = N(X = x_{[j]} \wedge Y = y_{[k]}),$$

simultánní relativní četnost dvojice $(x_{[j]}, y_{[k]})$ ve výběrovém souboru

$$p_{jk} = \frac{n_{jk}}{n},$$

marginální absolutní četnost varianty $x_{[j]}$

$$n_{j.} = N(X = x_{[j]}) = n_{j1} + \cdots + n_{js},$$

marginální relativní četnost varianty $x_{[j]}$

$$p_{j.} = \frac{n_{j.}}{n} = p_{j1} + \cdots + p_{js},$$

marginální absolutní četnost varianty $y_{[k]}$

$$n_{.k} = N(Y = y_{[k]}) = n_{1k} + \cdots + n_{rk},$$

marginální relativní četnost varianty $y_{[k]}$

$$p_{.k} = \frac{n_{.k}}{n} = p_{1k} + \cdots + p_{rk},$$

sloupcově podmíněná relativní četnost varianty $x_{[j]}$ za předpokladu $y_{[k]}$

$$p_{j(k)} = \frac{n_{jk}}{n_{.k}},$$

řádkově podmíněná relativní četnost varianty $y_{[k]}$ za předpokladu $x_{[j]}$

$$p_{(j)k} = \frac{n_{jk}}{n_{j.}}.$$

Kteroukoliv ze simultánních četností či podmíněných relativních četností zapisujeme do *kontingenční tabulky*. Kontingenční tabulka simultánních absolutních četností má tvar:

	y	$y_{[1]}$	$\ldots$	$y_{[s]}$	$n_{j.}$
x	n_{jk}				
$x_{[1]}$		n_{11}	$\ldots$	n_{1s}	$n_{1.}$
$\vdots$		$\vdots$	$\ldots$	$\vdots$	$\vdots$
$x_{[r]}$		n_{r1}	$\ldots$	n_{rs}	$n_{r.}$
$n_{.k}$		$n_{.1}$	$\ldots$	$n_{.s}$	n

Funkce

$$p(x, y) = \begin{cases} p_{jk} & \text{pro } x = x_{[j]},\ y = y_{[k]},\ j = 1, \ldots, r,\ k = 1, \ldots s \\ 0 & \text{jinak} \end{cases}$$

se nazývá *simultánní četnostní funkce*. Četnostní funkce pro znaky X a Y odlišíme indexem takto:

$$p_1(x) = \begin{cases} p_{j.} & \text{pro } x = x_{[j]},\ j = 1, \ldots, r \\ 0 & \text{jinak} \end{cases}$$

$$p_2(y) = \begin{cases} p_{.k} & \text{pro } y = y_{[k]},\ k = 1, \ldots, s \\ 0 & \text{jinak} \end{cases}$$

Řekneme, že znaky X, Y jsou v daném výběrovém souboru četnostně nezávislé, právě když pro všechna $j = 1, \ldots, r$ a všechna $k = 1, \ldots, s$ platí multiplikativní vztah: $p_{jk} = p_{j.} \cdot p_{.k}$ neboli

$$\forall (x, y) \in \mathbb{R}^2 :\ p(x, y) = p_1(x) \cdot p_2(y).$$

2.8. Věta
Mezi simultánní četnostní funkcí a marginálními četnostními funkcemi platí vztahy:

$$p_1(x) = \sum_{y=-\infty}^{\infty} p(x,y),$$
$$p_2(y) = \sum_{x=-\infty}^{\infty} p(x,y).$$

2.9. Příklad
Pro datový soubor z příkladu 1.10

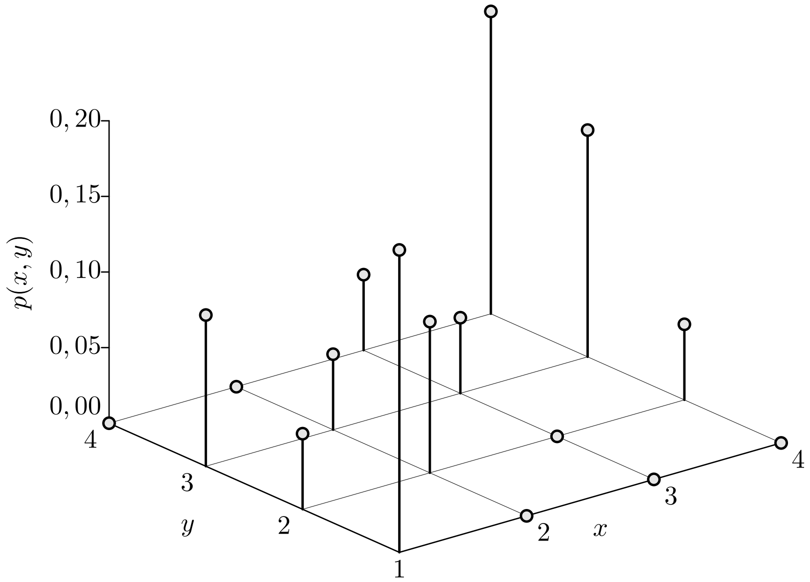
a) sestavte kontingenční tabulky simultánních absolutních a relativních četností,
b) nakreslete graf simultánní četnostní funkce $p(x,y)$,
c) sestavte kontingenční tabulky sloupcově a řádkově podmíněných relativních četností,
d) kolik procent těch studentů, kteří měli jedničku z angličtiny, mělo dvojku z matematiky,
e) kolik procent těch studentů, kteří měli jedničku z matematiky mělo dvojku z angličtiny,
f) zjistěte, zda znaky X, Y jsou v daném výběrovém souboru četnostně nezávislé.

Řešení:
ad a)

	y	1	2	3	4	$n_{j.}$
x	n_{jk}					
1		4	1	2	0	7
2		0	2	1	0	3
3		0	0	1	1	2
4		0	1	3	4	8
$n_{.k}$		4	4	7	5	$n = 20$

	y	1	2	3	4	$p_{j.}$
x	p_{jk}					
1		0,20	0,05	0,10	0,00	0,35
2		0,00	0,10	0,05	0,00	0,15
3		0,00	0,00	0,05	0,05	0,10
4		0,00	0,05	0,15	0,20	0,40
$p_{.k}$		0,20	0,20	0,35	0,25	1,00

ad b)



ad c)

y / x ($p_{j(k)}$)	1	2	3	4
1	1,00	0,25	0,29	0,00
2	0,00	0,50	0,14	0,00
3	0,00	0,00	0,14	0,20
4	0,00	0,25	0,43	0,80
$\sum$	1,00	1,00	1,00	1,00

y / x ($p_{(j)k}$)	1	2	3	4	$\sum$
1	0,57	0,14	0,29	0,00	1,00
2	0,00	0,67	0,33	0,00	1,00
3	0,00	0,00	0,50	0,50	1,00
4	0,00	0,12	0,38	0,50	1,00

ad d) Tento údaj najdeme ve druhém řádku prvního sloupce tabulky sloupcově podmíněných relativních četností: 0%.

ad e) Tento údaj najdeme v prvním řádku druhého sloupce tabulky řádkově podmíněných relativních četností: 14%.

ad f) Kdyby v daném výběrovém souboru byly oba znaky četnostně nezávislé, platil by pro všechna $j = 1, 2, 3, 4$ a všechna $k = 1, 2, 3, 4$ multiplikativní vztah: $p_{jk} = p_{j.} \cdot p_{.k}$, což splněno není. Tedy známky z matematiky a angličtiny nejsou četnostně nezávislé.

V některých datových souborech je počet variant znaku příliš veliký a použití bodového rozložení četností by vedlo k nepřehledným a roztříštěným výsledkům. V takových situacích používáme intervalové rozložení četností. Definujeme třídicí interval a jeho absolutní a relativní četnost, absolutní a relativní kumulativní četnost. Nově zavádíme četnostní hustotu třídícího intervalu. Uvedené četnosti zapisujeme do tabulky rozložení četností. Počet třídících intervalů stanovujeme např. podle Sturgesova pravidla. Intervalové rozložení četností požijeme v příkladu s datovým souborem obsahujícím údaje o mezích plasticity a pevnosti 60 vzorků oceli.

2.10. Definice
Nechť je dán jednorozměrný datový soubor. Jestliže počet variant znaku X je blízký rozsahu souboru, pak přiřazujeme četnosti nikoliv jednotlivým variantám, ale celým intervalům hodnot. Hovoříme pak o *intervalovém rozložení četnosti*.

2.11. Definice
Číselnou osu rozložíme na intervaly typu $(-\infty, u_1\rangle$, $(u_1, u_2\rangle$, $\ldots$, $(u_r, u_{r+1}\rangle$, (u_{r+1}, ∞) tak, aby okrajové intervaly neobsahovaly žádnou pozorovanou hodnotu znaku X. Užíváme označení:

j–tý třídicí interval znaku X, $j = 1, \ldots, r$:

$$(u_j, u_{j+1}\rangle,$$

délka j–tého třídicího intervalu znaku X:

$$d_j = u_{j+1} - u_j,$$

střed j–tého třídicího intervalu znaku X:

$$x_{[j]} = \frac{1}{2}(u_j + u_{j+1}).$$

Třídicí intervaly volíme nejčastěji stejně dlouhé. Jejich počet určíme např. pomocí *Sturgesova pravidla*: $r \approx 1 + 3{,}3 \cdot \log n$, kde n je počet variant znaku X.

2.12. Definice
Nechť je dán jednorozměrný datový soubor rozsahu n. Hodnoty znaku X roztřídíme do r třídících intervalů. Pro $j = 1, \ldots, r$ definujeme:

absolutní četnost j–tého třídicího intervalu ve výběrovém souboru

$$n_j = N(u_j < X \leq u_{j+1}),$$

relativní četnost j–tého třídicího intervalu ve výběrovém souboru

$$p_j = \frac{n_j}{n},$$

četnostní hustota j–tého třídicího intervalu ve výběrovém souboru

$$f_j = \frac{p_j}{d_j},$$

absolutní kumulativní četnost prvních j třídicích intervalů ve výběrovém souboru

$$N_j = N(X \leq u_{j+1}) = n_1 + \cdots + n_j,$$

relativní kumulativní četnost prvních j třídicích intervalů ve výběrovém souboru.

$$F_j = \frac{N_j}{n} = p_1 + \cdots + p_j.$$

Tabulka typu

$(u_j, u_{j+1}\rangle$	d_j	n_j	p_j	f_j	N_j	F_j
$(u_1, u_2\rangle$	d_1	n_1	p_1	f_1	N_1	F_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(u_r, u_{r+1}\rangle$	d_r	n_r	p_r	f_r	N_r	F_r
$\sum$		n	1			

se nazývá *tabulka rozložení četností.*

2.13. Příklad

Z fiktivního základního souboru všech vzorků oceli odpovídajících „všem myslitelným tavbám“ bylo do laboratoře dodáno 60 vzorků a zjištěny a hodnoty znaku X – mez plasticity a Y – mez pevnosti. Datový soubor má tvar:

$$\begin{bmatrix} 154 & 178 \\ 133 & 164 \\ 58 & 75 \\ 145 & 161 \\ 94 & 107 \\ 113 & 141 \\ 86 & 97 \\ 121 & 127 \\ 119 & 138 \\ 112 & 125 \\ 85 & 97 \\ 41 & 72 \\ 96 & 113 \\ 45 & 89 \\ 99 & 109 \end{bmatrix} \quad \begin{bmatrix} 51 & 95 \\ 101 & 114 \\ 160 & 169 \\ 87 & 101 \\ 88 & 139 \\ 83 & 98 \\ 106 & 111 \\ 92 & 104 \\ 85 & 103 \\ 112 & 118 \\ 98 & 102 \\ 103 & 108 \\ 99 & 119 \\ 104 & 128 \\ 107 & 118 \end{bmatrix} \quad \begin{bmatrix} 98 & 140 \\ 97 & 115 \\ 105 & 101 \\ 71 & 93 \\ 39 & 69 \\ 122 & 147 \\ 33 & 52 \\ 78 & 117 \\ 147 & 137 \\ 125 & 149 \\ 73 & 76 \\ 77 & 85 \\ 47 & 61 \\ 68 & 85 \\ 137 & 142 \end{bmatrix} \quad \begin{bmatrix} 44 & 68 \\ 92 & 116 \\ 141 & 157 \\ 155 & 189 \\ 136 & 155 \\ 82 & 81 \\ 136 & 163 \\ 72 & 79 \\ 66 & 81 \\ 42 & 61 \\ 113 & 123 \\ 42 & 85 \\ 133 & 147 \\ 153 & 179 \\ 85 & 91 \end{bmatrix}$$

a) Pro znak X stanovte optimální počet třídicích intervalů dle Sturgesova pravidla.
b) Sestavte tabulku rozložení četností.

Řešení:
ad a) Znak X má 50 variant, tedy podle Sturgesova pravidla je optimální počet třídicích intervalů $r = 7$. Budeme tedy volit 7 intervalů stejné délky tak, aby v nich byly obsaženy všechny pozorované hodnoty znaku X, z nichž nejmenší je 33, největší 160; volba $u_1 = 30, \ldots, u_8 = 170$ splňuje požadavky.

ad b)

$(u_j, u_{j+1}\rangle$	d_j	$x_{[j]}$	n_j	p_j	N_j	F_j	f_j
$(30, 50\rangle$	20	40	8	0,1333	8	0,1333	0,0066
$(50, 70\rangle$	20	60	4	0,0667	12	0,2000	0,0333
$(70, 90\rangle$	20	80	13	0,2166	25	0,4167	0,0108
$(90, 110\rangle$	20	100	15	0,2500	40	0,6667	0,0125
$(110, 130\rangle$	20	120	9	0,1500	49	0,8167	0,0075
$(130, 150\rangle$	20	140	7	0,1167	56	0,9333	0,0058
$(150, 170\rangle$	20	160	4	0,0667	60	1,0000	0,0033
Součet			60	1,0000			

Ke grafickému znázornění intervalového rozložení četností slouží histogram. S jeho pomocí lze dobře vysvětlit, co znamená hustota četnosti, což je funkce zavedená pomocí četnostních hustot jednotlivých třídicích intervalů. S hustotou četnosti úzce souvisí intervalová empirická distribuční funkce (je všude spojitá, protože je funkcí horní meze integrálu z hustoty četnosti). Pro údaje o mezi platicity oceli vytvoříme histogram a graf intervalové empirické distribuční funkce. Seznámíme se rovněž s vlastnostmi obou výše zmíněných funkcí.

2.14. Definice
Intervalové rozložení četností graficky znázorňujeme graficky pomocí *histogramu*. Je to graf skládající se z r obdélníků, sestrojených nad třídicími intervaly, přičemž obsah j–tého obdélníku je roven relativní četnosti p_j j–tého třídicího intervalu, $j = 1, \ldots, r$. Histogram je shora omezen schodovitou čarou, která je grafem funkce zvané hustota četnosti:

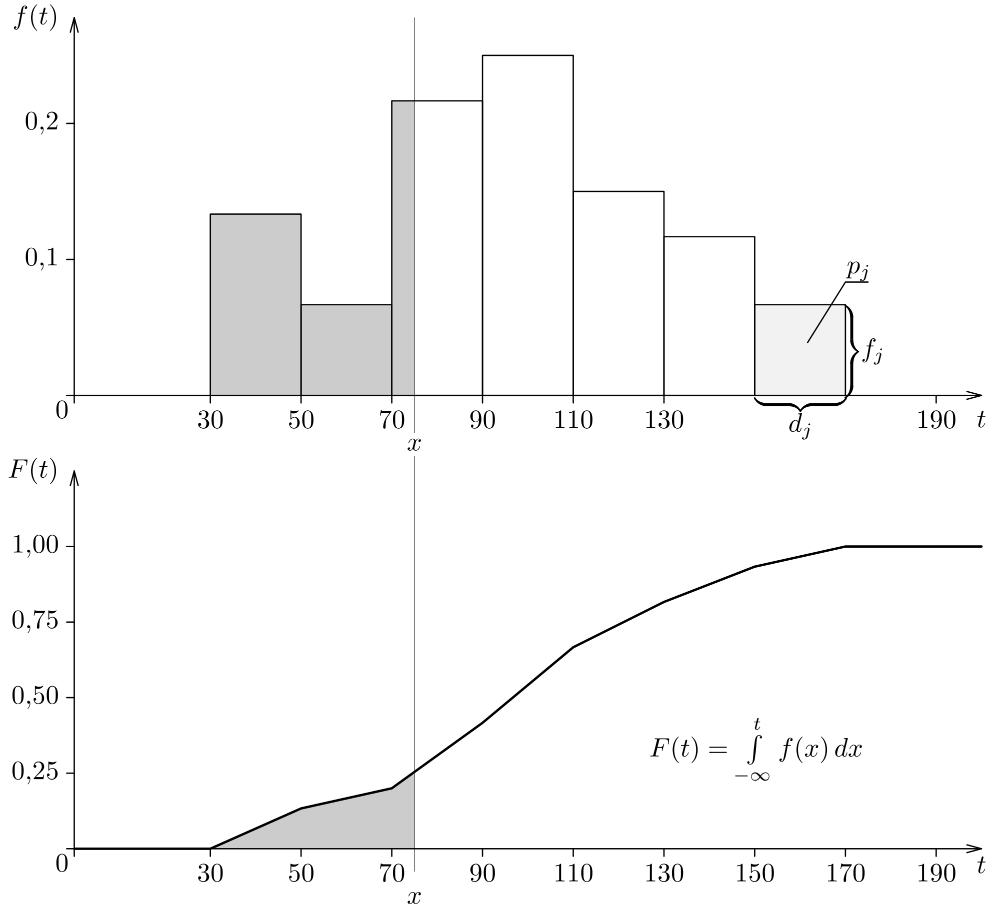
$$f(x) = \begin{cases} f_j & \text{pro } u_j < x \leq u_{j+1},\ j = 1, \ldots, r \\ 0 & \text{jinak} \end{cases}$$

Pomocí funkce hustoty četnosti zavedeme intervalovou empirickou distribuční funkci:

$$F(x) = \int_{-\infty}^{x} f(t)\, dt.$$

2.15. Příklad
Pro datový soubor z příkladu 2.13 nakreslete histogram pro znak X a pod histogram nakreslete graf intervalové empirické distribuční funkce.

Řešení:



2.16. Věta

Hustota četnosti je nezáporná ($\forall x \in \mathbb{R} : f(x) \geq 0$) a normovaná ($\int_{-\infty}^{\infty} f(x)\,dx$).

Intervalová empirická distribuční funkce je neklesající, spojitá a normovaná ($\lim\limits_{x \to -\infty} F(x) = 0$, $\lim\limits_{x \to \infty} F(x) = 1$).

V následujícím tématu se budeme věnovat dvourozměrnému intervalovému rozložení četnosti, tj. budeme pracovat s dvourozměrným datovým souborem. Zavedeme podobné pojmy jako u dvourozměrného bodového rozložení četnosti a jejich pochopení si ověříme na příkladě s datovým souborem obsahujícím údaje o mezi plasticity a mezi pevnosti oceli.

2.17. Definice

Nechť je dán dvourozměrný datový soubor

$$\begin{bmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix},$$

kde hodnoty znaku X roztřídíme do r třídicích intervalů $\langle u_j, u_{j+1}\rangle$, $j = 1, \dots, r$ s délkami $d_1, \dots, d_r$ a hodnoty znaku Y roztřídíme do s třídicích intervalů $\langle v_k, v_{k+1}\rangle$, $k = 1, \dots, s$ s délkami $h_1, \dots, h_s$. Pak definujeme:

simultánní absolutní četnost (j, k)*–tého třídicího intervalu*:

$$n_{jk} = N(u_j < X \leq u_{j+1} \wedge v_k < Y \leq v_{k+1}),$$

simultánní relativní četnost (j, k)*–tého třídicího intervalu*:

$$p_{jk} = \frac{n_{jk}}{n},$$

marginální absolutní četnost j*–tého třídicího intervalu pro znak* X:

$$n_{j.} = n_{j1} + \dots + n_{js},$$

marginální relativní četnost j*–tého třídicího intervalu pro znak* X:

$$p_{j.} = \frac{n_{j.}}{n},$$

marginální absolutní četnost k*–tého třídicího intervalu pro znak* Y:

$$n_{.k} = n_{1k} + \dots + n_{rk},$$

marginální relativní četnost k*–tého třídicího intervalu pro znak* Y:

$$p_{.k} = \frac{n_{.k}}{n},$$

simultánní četnostní hustota v (j, k)*–tém třídicím intervalu*:

$$f_{jk} = \frac{p_{jk}}{d_j h_k},$$

marginální četnostní hustota v j*–tém třídicím intervalu pro znak* X:

$$f_{j.} = \frac{p_{j.}}{d_j},$$

marginální četnostní hustota v k*–tém třídicím intervalu pro znak* Y:

$$f_{.k} = \frac{p_{.k}}{h_k}.$$

Kteroukoliv ze simultánních četností zapisujeme do kontingenční tabulky. Uveďme kontingenční tabulku simultánních absolutních četností:

	$\langle v_k, v_{k+1}\rangle$	$\langle v_1, v_2\rangle$	$\cdots$	$\langle v_s, v_{s+1}\rangle$	$n_{j.}$
$\langle u_j, u_{j+1}\rangle$	n_{jk}				
$\langle u_1, u_2\rangle$		n_{11}	$\dots$	n_{1s}	$n_{1.}$
$\vdots$		$\vdots$		$\vdots$	$\vdots$
$\langle u_r, u_{r+1}\rangle$		n_{r1}	$\dots$	n_{rs}	$n_{r.}$
$n_{.k}$		$n_{.1}$	$\dots$	$n_{.s}$	n

Funkce

$$f(x,y) = \begin{cases} f_{jk} & \text{pro } u_j < x \leq u_{j+1}, v_k < y \leq v_{k+1}, j = 1,\dots,r, k = 1,\dots,s \\ 0 & \text{jinak} \end{cases}$$

se nazývá *simultánní hustota četnosti*. Hustoty četnosti pro znaky X a Y odlišíme indexem takto:

$$f_1(x) = \begin{cases} f_{j.} & \text{pro } u_j < x \leq u_{j+1}, j = 1,\dots,r \\ 0 & \text{jinak} \end{cases}$$

$$f_2(y) = \begin{cases} f_{.k} & \text{pro } v_k < y \leq v_{k+1}, k = 1,\dots,s \\ 0 & \text{jinak} \end{cases}$$

Řekneme, že znaky X, Y jsou v daném výběrovém souboru četnostně nezávislé při intervalovém rozložení četností, jestliže pro všechna $j = 1,\dots,r$ a všechna $k = 1,\dots,s$ platí multiplikativní vztah: $f_{jk} = f_{j.} \cdot f_{.k}$ neboli pro

$$\forall (x,y) \in \mathbb{R}^2 : f(x,y) = f_1(x) f_2(y).$$

2.18. Věta

Mezi simultánní hustotou četnosti a marginálními hustotami četnosti platí vztahy:

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)\, dy, \qquad f_2(y) = \int_{-\infty}^{\infty} f(x,y)\, dx.$$

2.19. Příklad

Pro datový soubor z příkladu 2.13

a) stanovte dle Sturgesova pravidla optimální počet třídicích intervalů pro znak Y

b) sestavte kontingenční tabulku simultánních absolutních četností.

Řešení:

ad a) Počet variant znaku Y je 52. Podle Sturgesova pravidla je tedy optimální počet třídicích intervalů $s = 7$. Nejmenší hodnota je 52 a největší 189. Volíme $v_1 = 50$, $v_2 = 70, \dots, v_8 = 190$.

ad b)

$(u_j, u_{j+1}\rangle$	n_{jk} \ $(v_k, v_{k+1}\rangle$	$(50,70\rangle$	$(70,90\rangle$	$(90,110\rangle$	$(110,130\rangle$	$(130,150\rangle$	$(150,170\rangle$	$(170,190\rangle$	$n_{j.}$
$(30,50\rangle$		5	3	0	0	0	0	0	8
$(50,70\rangle$		0	3	1	0	0	0	0	4
$(70,90\rangle$		0	4	7	1	1	0	0	13
$(90,110\rangle$		0	0	6	8	1	0	0	15
$(110,130\rangle$		0	0	0	4	5	0	0	9
$(130,150\rangle$		0	0	0	0	2	5	0	7
$(150,170\rangle$		0	0	0	0	0	1	3	4
$n_{.k}$		5	10	14	13	9	6	3	60

Shrnutí kapitoly

Není-li v jednorozměrném souboru počet variant znaku příliš velký, pak přiřazujeme četnosti jednotlivým variantám znaku a hovoříme o **bodovém rozložení četnosti**. To lze znázornit graficky pomocí různých **diagramů** (např. tečkový diagram, sloupkový diagram atd.). Pokud zapíšeme četnosti do tabulky, dostaneme **variační řadu**. Pomocí relativních četností zavedeme **četnostní funkci**, pomocí kumulativních relativních četností **empirickou distribuční funkci**, která má schodovitý průběh.

Pracujeme-li s dvourozměrným datovým souborem, zavádíme **simultánní četnosti** a zapisujeme je do **kontingenční tabulky**. Na okrajích kontingenční tabulky jsou uvedeny **marginální četnosti**, které se vztahují jen k jednomu znaku. Pomocí simultánních kumulativních relativních četností zavádíme simultánní četnostní funkci. Simultánní a marginální četnosti či četnostní funkce nám snadno umožní ověřit **četnostní nezávislost** dvou znaků v daném výběrovém souboru.

Je-li počet variant znaku srovnatelný s rozsahem souboru, použijeme raději **intervalové rozložení četnosti**, při němž přiřazujeme četnosti nikoli jednotlivým variantám, ale třídicím intervalům. Jejich počet určíme např. pomocí **Sturgesova pravidla**. Četnosti třídicích intervalů zapisujeme do **tabulky rozložení četností.** Relativní četnosti třídicích intervalů znázorňujeme pomocí **histogramu**. Schodovitá čára shora omezující histogram je grafem **hustoty četnosti**. Spojitým protějškem schodovité empirické distribuční funkce je **intervalová empirická distribuční funkce** zavedená jako funkce horní meze integrálu z hustoty četnosti.

Při dvourozměrném intervalovém rozložení četností pracujeme s podobnými pojmy jako u dvourozměrného bodového rozložení četnosti. Místo simultánní a marginální četnostní funkce samozřejmě máme **simultánní** či **marginální hustotu četnosti**.

Kontrolní otázky a úkoly

1 Jaké grafy znázorňující rozložení četností znáte? Popište způsob jejich konstrukce.
2 Jak vzniká variační řada?
3 Jaké četnosti zapisujeme do kontingenční tabulky?
4 Kdy jsou v daném výběrovém souboru znaky četnostně nezávislé?
5 K čemu slouží Sturgesovo pravidlo?
6 Vyjmenujte funkcionální charakteristiky skalárního znaku a dvourozměrného vektorového znaku při bodovém a intervalovém rozložení četností.
7 (S) V rámci marketingového průzkumu trhu bylo dotázáno 25 náhodně vybraných zákazníků jisté pojišťovny a byl zjišťován jejich zájem o nový druh pojištění (znak X) a současně jejich rodinný stav (znak Y). Získané odpovědi byly zakódovány pro znak X takto: jednoznačný nezájem $= 1$, podprůměrný zájem $= 2$, průměrný zájem $= 3$, nadprů-

měrný zájem = 4, jednoznačný zájem = 5 a pro znak Y takto: svobodný = 1, rozvedený nebo ovdovělý = 2, ženatý = 3.

$$\begin{bmatrix} 5 & 1 \\ 3 & 2 \\ 4 & 2 \\ 4 & 1 \\ 5 & 2 \end{bmatrix} \quad \begin{bmatrix} 4 & 3 \\ 3 & 3 \\ 1 & 1 \\ 4 & 3 \\ 3 & 3 \end{bmatrix} \quad \begin{bmatrix} 5 & 2 \\ 3 & 2 \\ 4 & 1 \\ 5 & 1 \\ 1 & 3 \end{bmatrix} \quad \begin{bmatrix} 4 & 2 \\ 5 & 3 \\ 4 & 3 \\ 5 & 3 \\ 3 & 1 \end{bmatrix} \quad \begin{bmatrix} 4 & 1 \\ 4 & 3 \\ 4 & 3 \\ 2 & 3 \\ 2 & 2 \end{bmatrix}$$

a) Pro znak X sestrojte jednorozměrný tečkový diagram, sestavte variační řadu, sestrojte graf četnostní funkce a empirické distribuční funkce.

b) Pro vektorový znak (X, Y) sestavte kontingenční tabulku absolutních četností, absolutních kumulativních četností, dále kontingenční tabulky sloupcově a řádkově podmíněných četností a graf simultánní četnostní funkce.

c) Jsou znaky X, Y v daném výběrovém souboru četnostně nezávislé?

8 (S) U 50 náhodně vybraných posluchačů a posluchaček VŠE v Praze byla zjišťována jejich hmotnost v kg (znak X) a jejich výška v cm (znak Y).

$$\begin{bmatrix} 58 & 178 \\ 68 & 173 \\ 56 & 170 \\ 60 & 170 \\ 61 & 173 \\ 71 & 181 \\ 85 & 184 \\ 80 & 170 \\ 52 & 172 \\ 72 & 182 \end{bmatrix} \quad \begin{bmatrix} 65 & 170 \\ 57 & 169 \\ 65 & 169 \\ 60 & 170 \\ 54 & 162 \\ 52 & 169 \\ 83 & 182 \\ 60 & 168 \\ 68 & 173 \\ 63 & 171 \end{bmatrix} \quad \begin{bmatrix} 72 & 177 \\ 90 & 192 \\ 57 & 176 \\ 51 & 168 \\ 81 & 190 \\ 73 & 177 \\ 75 & 179 \\ 71 & 180 \\ 66 & 178 \\ 67 & 182 \end{bmatrix} \quad \begin{bmatrix} 72 & 191 \\ 57 & 174 \\ 57 & 160 \\ 56 & 170 \\ 56 & 172 \\ 52 & 165 \\ 72 & 185 \\ 75 & 170 \\ 52 & 163 \\ 63 & 184 \end{bmatrix} \quad \begin{bmatrix} 63 & 172 \\ 58 & 163 \\ 64 & 174 \\ 52 & 168 \\ 55 & 164 \\ 67 & 173 \\ 60 & 170 \\ 55 & 160 \\ 62 & 172 \\ 70 & 171 \end{bmatrix}$$

a) Pro znak X stanovte optimální počet třídicích intervalů podle Sturgesova pravidla, sestavte tabulku rozložení četnosti, nakreslete histogram a graf intervalové empirické distribuční funkce.

b) Pro znak Y rovněž stanovte optimální počet třídicích intervalů podle Sturgesova pravidla. Pro vektorový znak (X, Y) sestavte kontingenční tabulku absolutních četností a nakreslete dvourozměrný tečkový diagram.

c) Jsou znaky X, Y v daném výběrovém souboru četnostně nezávislé?

3. Číselné charakteristiky znaků

Cíl kapitoly

Po prostudování této kapitoly budete umět:

- rozlišovat různé typy znaků
- vypočítat různé charakteristiky, polohy a variability skalárního znaku
- vypočítat charakteristiky těsnosti lineární závislosti dvou znaků
- využít vlastností číselných charakteristik ke zjednodušení výpočtů
- vypočítat vážené číselné charakteristiky znaků.

Časová zátěž

Pro zvládnutí této kapitoly budete potřebovat 5 – 6 hodin studia.

Nejprve se naučíme rozlišovat různé typy znaků podle toho, jaký je jejich stupeň kvantifikace. Pro jednotlivé typy znaků pak zavedeme číselné charakteristiky popisující polohu hodnot znaku na číselné ose a jejich proměnlivost. Seznámíme se rovněž s důležitými vlastnostmi číselných charakteristik a naučíme se je počítat pro konkrétní datové soubory.

3.1. Motivace

Ve druhé kapitole jsme se seznámili s funkcionálními charakteristikami znaků, jako jsou $p(x, y)$, $p_1(x)$, $p_2(y)$, $F(x)$, $f(x, y)$, $f_1(x)$, $f_2(y)$, které nesou úplnou informaci o rozložení četností. V této kapitole zavedeme číselné charakteristiky, které nás informují o některých rysech tohoto rozložení četností: o poloze (úrovni) hodnot znaku, o jejich variabilitě (rozptýlení), o těsnosti závislosti dvou znaků a pod. Pro různé typy znaků se používají různé číselné charakteristiky, proto se nejdřív seznámíme s jednotlivými typy znaků.

3.2. Definice

Podle stupně kvantifikace znaky třídíme takto:

(n) *Nominální znaky* připouštějí obsahovou interpretaci jedině relace rovnosti $x_1 = x_2$ (popřípadě $x_1 \neq x_2$), tj. hodnoty znaku představují jen číselné kódy kvalitativních pojmenování. Např. městské tramvaje jsou očíslovány, ale např. č. 4 a 12 říkají jen to, že jde o různé tratě: nic jiného se z nich o vztahu obou tratí nedá vyčíst.

(o) *Ordinální znaky* připouštějí obsahovou interpretaci kromě relace rovnosti i v případě relace uspořádání $x_1 < x_2$ (popřípadě $x_1 > x_2$), tj. jejich uspořádání vyjadřuje větší nebo menší intenzitu zkoumané vlastnosti. Např. školní klasifikace vyjadřuje menší nebo větší znalosti zkoušených (jedničkář je lepší než dvojkař), ale intervaly mezi známkami nemají obsahové interpretace (netvrdíme, že rozdíl ve znalostech mezi jedničkářem a dvojkařem je stejný jako mezi trojkařem a čtyřkařem. Podobný charakter mají různá bodování ve sportovních, uměleckých a jiných soutěžích.

(i) *Intervalové znaky* připouštějí obsahovou interpretaci kromě relace rovnosti a uspořádání též u operace rozdílu $x_1 - x_2$ (popřípadě součtu $x_1 + x_2$), tj. stejný interval mezi jednou dvojicí hodnot a jinou dvojicí hodnot vyjadřuje i stejný rozdíl v extenzitě zkoumané vlastnosti. Např. teplota měřená ve stupních Celsia představuje intervalový znak. Naměříme-li ve čtyřech dnech polední teploty 0, 2, 4, 6, znamená to, že každým dnem stoupla teplota o 2 stupně Celsia. Bylo by však chybou interpretovat tyto údaje tvrzením, že ze druhého na třetí den vzrostla teplota dvakrát, kdežto ze třetího na čtvrtý pouze jedenapůlkrát.

(p) *Poměrové znaky* umožňují obsahovou interpretaci kromě relace rovnosti a uspořádání a operace rozdílu ještě u operace podílu x_1/x_2 (popřípadě součinu $x_1 \cdot x_2$), tj. stejný poměr mezi jednou dvojicí hodnot a druhou dvojicí hodnot znamená i stejný podíl v extenzitě zkoumané vlastnosti. Např. má-li jedna osoba hmotnost 150 kg a druhá 75 kg, má smysl prohlásit, že první je dvakrát hmotnější než druhá.

Zvláštní postavení mají:

(a) *Alternativní znaky*, které nabývají jen dvou hodnot, např. $0, 1$, což znamená absenci a prezenci nějakého jevu. Například 0 bude znamenat neúspěch, 1 úspěch při řešení určité úlohy. Alternativní znaky mohou být ztotožněny s kterýmkoliv z předcházejících typů.

3.3. Definice
Pro nominální znaky používáme jako charakteristiku polohy *modus*. U bodového rozložení četností je to nejčetnější varianta znaku, u intervalového střed nejčetnějšího třídicího intervalu.

3.4. Definice
Pro ordinální znaky používáme jako charakteristiku polohy α–kvantil. Je-li $\alpha \in (0,1)$, pak α–kvantil x_α je číslo, které rozděluje uspořádaný datový soubor na dolní úsek, obsahující aspoň podíl α všech dat a na horní úsek obsahující aspoň podíl $1-\alpha$ všech dat. Pro výpočet α–kvantilu slouží algoritmus:

$$n\alpha = \begin{cases} \text{celé číslo } c \;\Rightarrow\; x_\alpha = \dfrac{x_{(c)} + x_{(c+1)}}{2} \\ \text{necelé číslo } \Rightarrow \text{ zaokrouhlíme nahoru na nejbližší celé číslo } c \;\Rightarrow \\ \qquad \Rightarrow\; x_\alpha = x_{(c)} \end{cases}$$

Pro speciálně zvolená α užíváme názvů: $x_{0,50}$ – *medián*, $x_{0,25}$ – *dolní kvartil*, $x_{0,75}$ – *horní kvartil*, $x_{0,1}, \dots, x_{0,9}$ – *decily*, $x_{0,01}, \dots, x_{0,99}$ – *percentily*. Jako charakteristika variability slouží *kvartilová odchylka*:

$$q = x_{0,75} - x_{0,25}.$$

3.5. Příklad
Pro datový soubor známek z matematiky (viz příklad 1.10) vypočtěte medián, oba kvartily a kvartilovou odchylku.

Řešení:

α	$n\alpha$	c		x_α
0,25	5	5	$\frac{(1+1)}{2}$	1
0,50	10	10	$\frac{(2+3)}{2}$	2,5
0,75	15	15	$\frac{(4+4)}{2}$	4

$$q = 4 - 1 = 3$$

3.6. Definice
Pro intervalové a poměrové znaky slouží jako charakteristika polohy *aritmetický průměr*

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i$$

(lze ho interpretovat jako těžiště jednorozměrného tečkového digramu). Charakteristikou variability je *rozptyl*

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$$

či *směrodatná odchylka* $s = \sqrt{s^2}$. Pomocí průměru zavedeme *centrovanou hodnotu* $x_i - m$ (podle znaménka poznáme, zda i–tá hodnota je podprůměrná či nadprůměrná a pomocí směrodatné odchylky zavedeme *standardizovanou hodnotu* $\dfrac{x_i - m}{s}$ (vyjadřuje o kolik směrodatných odchylek se i–tá hodnota odchýlila od průměru).

3.7. Věta
Rozptyl je nulový, právě když $x_1 = x_2 = \cdots = x_n$.

3.8. Příklad
Vypočtěte průměr a rozptyl

a) centrovaných hodnot,
b) standardizovaných hodnot.

Řešení:
ad a) Průměr centrovaných hodnot:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - m) = m - \frac{1}{n}\cdot n \cdot m = 0.$$

Rozptyl centrovaných hodnot:

$$\frac{1}{n}\sum_{i=1}^{n}((x_i - m) - 0)^2 = s^2.$$

ad b) Průměr standardizovaných hodnot:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-m)}{s}=\frac{1}{s}\cdot 0=0.$$

Rozptyl standardizovaných hodnot:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-m}{s}-0\right)^2=\frac{s^2}{s^2}=1.$$

3.9. Poznámka

V předešlém příkladě jsme vypočítali, že průměr centrovaných hodnot je 0. Této skutečnosti lze využít k vysvětlení rozptylu: chceme získat číslo, které by charakterizovalo variabilitu jednotlivých hodnot kolem průměru. Průměr centrovaných hodnot nelze použít (vyjde 0), proto místo centrovaných hodnot vezmeme jejich kvadráty. Tím dospějeme ke vzorci pro rozptyl: $s^2=\frac{1}{n}\sum\limits_{i=1}^{n}(x_i-m)^2$. Rozptyl však vychází v kvadrátech jednotek, v nichž byl měřen znak X, proto raději používáme směrodatnou odchylku s. Definiční tvar vzorce pro rozptyl není příliš vhodný pro výpočty, v praxi se používá výpočetní tvar vzorce pro rozptyl:

$$s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-m)^2=\frac{1}{n}\sum_{i=1}^{n}(x_i^2-2mx_i+m^2)=\frac{1}{n}\sum_{i=1}^{n}x_i^2-$$
$$-\frac{1}{n}\cdot 2m\cdot\sum_{i=1}^{n}x_i+\frac{1}{n}\sum_{i=1}^{n}m^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2-2m^2+\frac{1}{n}\cdot n\cdot m^2=$$
$$=\frac{1}{n}\sum_{i=1}^{n}x_i^2-m^2.$$

3.10. Definice

Pro poměrové znaky používáme jako charakteristiku variability *koeficient variace* $\dfrac{s}{m}$. Je to bezrozměrné číslo, které se často vyjadřuje v procentech. Umožňuje porovnat variabilitu několika znaků. Jsou-li všechny hodnoty poměrového znaku kladné, pak jako charakteristiku polohy lze užít *geometrický průměr* $\sqrt[n]{x_1\cdot\ldots\cdot x_n}$.

3.11. Příklad

Vypočtěte koeficient variace meze plasticity a meze pevnosti oceli pro datový soubor z příkladu 2.13.

Řešení:

$$\frac{s_1}{m_1}=\frac{32{,}441}{95{,}88}=0{,}338,\qquad \frac{s_2}{m_2}=\frac{32{,}515}{114{,}40}=0{,}284.$$

Zjistili jsme, že koeficient variace meze plasticity je 33,8%, zatímco meze pevnosti jen 28,4%.

Nyní se budeme zabývat číselnými charakteristikami dvourozměrného datového souboru se znaky intervalového či poměrového typu. Společnou variabilitu těchto dvou znaků kolem jejich průměru měříme pomocí kovariance. Jako míra těsnosti lineární závislosti dvou znaků slouží koeficient korelace. Je velmi důležité porozumět vlastnostem koeficientu korelace, proto si pozorně prohlédněte obrázky ilustrující jeho význam. Pro praktické procvičení nám poslouží příklad na číselné charakteristiky mezí plasticity a pevnosti.

3.12. Definice
Pro dvourozměrný datový soubor

$$\begin{bmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix},$$

kde znaky X, Y jsou intervalového či poměrového typu, používáme jako charakteristiku společné variability znaků X, Y kolem jejich průměrů *kovarianci*

$$s_{12} = \frac{1}{n} \sum_{i=1}^{n} (x_i - m_1)(y_i - m_2).$$

3.13. Poznámka
Kovariance je průměrem součinů centrovaných hodnot. Pokud se nadprůměrné (podprůměrné) hodnoty znaku X sdružují s nadprůměrnými (podprůměrnými) hodnotami znaku Y, budou součiny centrovaných hodnot $x_i - m_1$ a $y_i - m_2$ vesměs kladné a jejich průměr (tj. kovariance) rovněž. Znamená to, že mezi znaky X, Y existuje určitý stupeň přímé lineární závislosti. Pokud se nadprůměrné (podprůměrné) hodnoty znaku X sdružují s podprůměrnými (nadprůměrnými) hodnotami znaku Y, budou součiny centrovaných hodnot vesměs záporné a jejich průměr rovněž. Znamená to, že mezi znaky X a Y existuje určitý stupeň nepřímé lineární závislosti. Je-li kovariance nulová, pak řekneme, že znaky X, Y jsou nekorelované a znamená to, že mezi nimi neexistuje žádná lineární závislost.

Pro výpočet kovariance používáme vzorec:

$$s_{12} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - m_1 m_2.$$

3.14. Definice
Jsou-li směrodatné odchylky s_1, s_2 nenulové, pak definujeme *koeficient korelace* znaků X, Y vzorcem

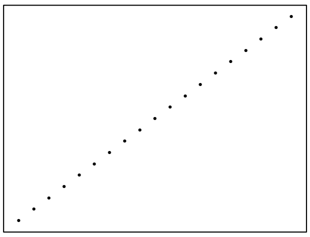
$$r_{12} = \frac{1}{n} \sum_{i=1}^{n} \frac{x_i - m_1}{s_1} \cdot \frac{y_i - m_2}{s_2}.$$

3.15. Věta

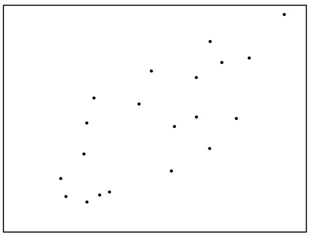
Pro koeficient korelace platí $-1 \le r_{12} \le 1$ a rovnosti je dosaženo právě když mezi hodnotami $x_1, \dots, x_n$ a $y_1, \dots, y_n$ existuje úplná lineární závislost, tj. existují konstanty a, b tak, že $y_i = a + bx_i$, $i = 1, \dots, n$, přičemž znaménko $+$ platí pro $b > 0$, znaménko $-$ pro $b < 0$. (Uvedená nerovnost se nazývá Cauchyova – Schwarzova – Buňakovského nerovnost.)

3.16. Poznámka

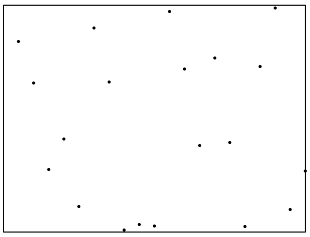
Koeficient korelace se počítá podle vzorce $r_{12} = \dfrac{s_{12}}{s_1 s_2}$. Představu o významu hodnot koeficientu korelace podávají následující dvourozměrné tečkové diagramy.



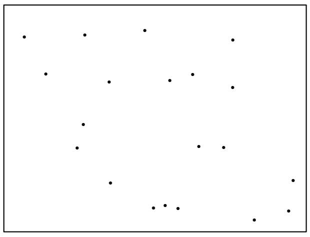
$r = 1{,}00$



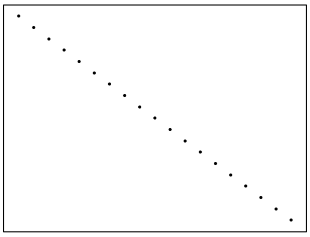
$r = 0{,}76$



$r = 0{,}00$



$r = -0{,}37$



$r = -1{,}00$

3.17. Příklad

Pro datový soubor z příkladu 2.13 vypočtěte

a) aritmetické průměry znaků X, Y,
b) rozptyly a směrodatné odchylky znaků X, Y,
c) kovarianci a koeficient korelace znaků X, Y.

Řešení:

ad a) $m_1 = 95{,}9$, $m_2 = 114{,}4$.

ad b) $s_1^2 = 1052{,}40$, $s_2^2 = 1057{,}21$, $s_1 = 32{,}4$, $s_2 = 32{,}5$.

ad c) $s_{12} = 985{,}76$, $r_{12} = 0{,}936$.

Koeficient korelace svědčí o tom, že mezi oběma znaky existuje velmi silná přímá lineární závislost – čím vyšší je mez plasticity, tím je vyšší mez pevnosti a čím je nižší mez plasticity, tím je nižší mez pevnosti.

Při výpočtu číselných charakteristik se v řadě situací uplatní věta shrnující některé jejich vlastnosti. Pro lepší pochopení uvedených vlastností slouží následující příklad.

3.18. Věta

Uveďme některé vlastnosti číselných charakteristik.

a) Nechť m_1 je aritmetický průměr a s_1^2 rozptyl znaku X. Pak znak $Y = a + bX$ má aritmetický průměr $m_2 = a + bm_1$ a rozptyl $s_2^2 = b^2 s_1^2$.
b) Nechť m_1, m_2 jsou aritmetické průměry, s_1^2, s_2^2 rozptyly a s_{12} kovariance znaků X, Y. Pak znak $U = X+Y$ má aritmetický průměr $m_3 = m_1+m_2$ a rozptyl $s_3^2 = s_1^2 + s_2^2 + 2s_{12}$.
c) Nechť s_{12} je kovariance znaků X, Y a m_1, m_2 jsou aritmetické průměry znaků X, Y. Pak znaky $U = a + bX$, $V = c + dY$ mají kovarianci $s_{34} = bds_{12}$.

3.19. Příklad

a) Znak X má aritmetický průměr 2 a rozptyl 3. Najděte aritmetický průměr a rozptyl znaku $Y = -1 + 3X$.
b) Znaky X a Y mají aritmetické průměry 3 a 2, rozptyly 2 a 3, kovarianci 1,5. Vypočtěte aritmetický průměr a rozptyl znaku $Z = 5X - 4Y$.
c) Součet rozptylů dvou znaků je 120, součin 1000 a rozptyl jejich součtů je 100. Vypočtěte koeficient korelace těchto znaků.

Řešení:

ad a) $m_2 = -1 + 3m_1 = -1 + 3 \cdot 2 = 5,\ s_2^2 = 3^2 \cdot s_1^2 = 9 \cdot 3 = 27.$

ad b) $m_3 = 5m_1 - 4m_2 = 5\cdot 3 - 4\cdot 2 = 7$, $s_3^2 = 5^2 \cdot s_1^2 + (-4)^2 \cdot s_2^2 + 2\cdot 5\cdot(-4)\cdot s_{12} = 25 \cdot 2 + 16 \cdot 9 - 40 \cdot 1{,}5 = 134.$

ad c) $s_1^2 + s_2^2 = 150,\ s_1^2 \cdot s_2^2 = 1000,\ s_{1+2}^2 = 100 = s_1^2 + s_2^2 + 2s_{12} \Rightarrow s_{12} = \frac{s_{1+2}^2 - s_1^2 - s_2^2}{2} = \frac{100-120}{2} = -10,\ r_{12} = \frac{s_{12}}{s_1 \cdot s_2} = \frac{-10}{\sqrt{1000}} = -0{,}316.$

Pokud nemáme k dispozici původní datový soubor, ale jenom variační řadu nebo tabulku rozložení četností (resp. kontingenční tabulku), můžeme vypočítat tzv. vážené číselné charakteristiky. Pro datový soubor obsahující údaje o mezi plasticity a mezi pevnosti oceli je zajímavé porovnat původní číselné charakteristiky a vážené číselné charakteristiky.

3.20. Definice

a) Vážené číselné charakteristiky u bodového rozložení četností:

Vážený aritmetický průměr

$$m = \frac{1}{n} \sum_{i=1}^{n} n_j x_{[j]}.$$

Vážený rozptyl

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} n_j (x_{[j]} - m)^2.$$

Vážená kovariance

$$s_{12} = \frac{1}{n} \sum_{j=1}^{n} \sum_{k=1}^{n} n_{jk} (x_{[j]} - m_1)(y_{[k]} - m_2).$$

b) Vážené číselné charakteristiky u intervalového rozložení četnosti:
Vzorce jsou formálně shodné s předešlými. Je však zapotřebí uvést, že výpočty jsou přesné jen tehdy, souhlasí-li průměry v jednotlivých třídicích intervalech se středy těchto intervalů, resp. vykompenzují-li se vzájemně chyby vzniklé v důsledku odchylek středů intervalů od průměru v těchto intervalech. Oba tyto případy jsou však vzácné a většinou se dopustíme určité chyby.

3.21. Příklad
Pro intervalové rozložení četností uvedené v příkladu 2.13 spočtěte vážené číselné charakteristiky a porovnejte je s číselnými charakteristikami uvedenými v příkladu 3.17.

Řešení:

	bodové rozložení	intervalové rozložení
m_1	95,88	96,67
m_2	114,40	113,67
s_1^2	1052,40	1148,89
s_2^2	1057,21	1019,89
s_1	32,441	33,895
s_2	32,515	31,936
s_{12}	985,76	998,89
r_{12}	0,939	0,923

Shrnutí kapitoly

Podle stupně kvantifikace znaky třídíme na **nominální**, **ordinální**, **intervalové**, **poměrové** a **alternativní**. Jako charakteristika polohy nominálních znaků slouží **modus**. Charakteristikou polohy ordinálních znaků je kterýkoliv **α–kvantil**, často se používá **medián**, **dolní** a **horní kvartil**, **decily**, **percentily**. Rozdíl horního a dolního kvartilu je **kvartilová odchylka**, kterou používáme jako charakteristiku variability. U intervalových znaků slouží jako charakteristika polohy **aritmetický průměr** a jako charakteristika variability **rozptyl** či **směrodatná odchylka**. Odečteme-li od libovolné hodnoty průměr, dostaneme **centrovanou hodnotu**, a podělíme-li centrovanou hodnotu směrodatnou odchylkou, získáme **standardizovanou hodnotu**. Pro poměrové znaky používáme **koeficient variace**. Mají-li kladné hodnoty, pak jejich polohu charakterizujeme **geometrickým průměrem**.

Máme-li dvourozměrný datový soubor, pak jako charakteristiku společné variability zavedeme kovarianci a jako míru těsnosti lineární závislosti **koeficient korelace**. Podle **Cauchy – Schwarzovy – Buňakovského nerovnosti** nabývá koeficient korelace hodnot mezi -1 a 1.

Je-li k dispozici variační řada u bodového rozložení četností nebo tabulka rozložení četností u intervalového rozložení četností (resp. kontingenční tabulka), můžeme vypočítat vážené číselné charakteristiky: **vážený aritmetický průměr**, **vážený rozptyl** a **váženou kovarianci**.

Kontrolní otázky a úkoly

1 Udejte příklad nominálního, ordinálního, intervalového, poměrového a alternativního znaku.

2 Jaké charakteristiky polohy a variability užíváme pro uvedené typy znaků?

3 Kdy se shodují číselné charakteristiky s váženými číselnými charakteristikami?

4 Jaký význam má koeficient korelace?

5 V akciové společnosti je průměrná mzda 13 500 Kč. Přitom 30% pracovníků s nejnižší mzdou má průměrně 9 000 Kč. Na začátku roku dostal každý z těchto pracovníků přidáno 500 Kč. O kolik % vzrostla průměrná mzda v celé akciové společnosti?

6 (S) Při statistickém šetření pojištěnců byly získány tyto výše pojistek v Kč:

výše pojistky	390	410	430	450	470	490	510	530	550	570
abs. četnost	7	10	14	22	25	12	3	3	2	2

Určete aritmetický průměr, medián, modus, rozptyl, směrodatnou odchylku a koeficient variace výše pojistky.

7 V datovém souboru, z něhož byl vypočten průměr 110 a rozptyl 800, byly zjištěny 2 chyby: místo 85 má být 95 a místo 120 má být 150. Ostatních 18 údajů je správných. Opravte průměr a rozptyl.

8 Vážený aritmetický průměr činil 1500 a vážený rozptyl 90000. Varianty $x_{[j]}$ byly transformovány vztahem:

$$y_{[j]} = \frac{x_{[j]} - a}{h},$$

$j = 1, \ldots, r$. Po této transformaci byl vážený aritmetický průměr 5 a vážený rozptyl 9. Určete konstanty a a h.

9 (S) Pro dvourozměrný datový soubor

2	4	4	5	6	8	10	10	10	10
1	2	3	4	4	4	5	5	5	6

vypočtěte koeficient korelace.

10 Rozptyl součtů hodnot dvou znaků je 350, rozptyl rozdílů je 700. Vypočtěte koeficient korelace, víte-li, že oba znaky mají stejné rozptyly.

4. Regresní přímka

Cíl kapitoly

Po prostudování této kapitoly budete umět:

- stanovit odhady parametrů regresní přímky a znát jejich význam
- posoudit kvalitu proložení regresní přímky dvourozměrným tečkovým diagramem
- vypočítat regresní odhady závisle proměnného znaku
- stanovit odhady parametrů druhé regresní přímky
- znát vztahy mezi parametry první a druhé regresní přímky.

Časová zátěž

Pro zvládnutí této kapitoly budete potřebovat 3 – 4 hodiny studia.

Budeme se zabývat speciálním případem, kdy hodnoty znaku Y závisejí na hodnotách znaku X přibližně lineárně. Ukážeme si, jak tuto závislost popsat regresní přímkou, jak odhadnout její parametry metodou nejmenších čtverců na základě znalosti dvourozměrného datového souboru a jak posoudit kvalitu regresní přímky pomocí indexu determinace. Vysvětlíme si význam regresních parametrů a v příkladu se budeme zabývat regresní přímkou meze pevnosti na mez plasticity.

4.1. Motivace

Cílem regresní analýzy je vystižení závislosti hodnot znaku Y na hodnotách znaku X. Při tom je nutné vyřešit dva problémy: jaký typ funkce použít k vystižení dané závislosti a jak stanovit konkrétní parametry zvoleného typu funkce? Typ funkce určíme buď logickým rozborem zkoumané závislosti nebo se snažíme ho odhadnout pomocí dvourozměrného tečkového diagramu. Zde se omezíme na lineární závislost $y = \beta_0 + \beta_1 x$. Odhady b_0 a b_1 neznámých parametrů β_0, β_1 získáme na základě dvourozměrného datového souboru

$$\begin{bmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix}$$

metodou nejmenších čtverců. Požadujeme, aby průměr součtu čtverců odchylek skutečných a odhadnutých hodnot byl minimální, tj. aby výraz

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

nabýval svého minima vzhledem k β_0 a β_1. Tento výraz je minimální, jsou-li jeho první derivace podle β_0 a β_1 nulové. Stačí tyto derivace spočítat, položit je rovny 0 a řešit systém dvou rovnic o dvou neznámých, tzv. systém normálních rovnic.

4.2. Definice
Nechť je dán dvourozměrný datový soubor

$$\begin{bmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix}$$

a přímka $y = \beta_0 + \beta_1 x$. Výraz

$$q(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

se nazývá rozptyl hodnot znaku Y kolem přímky $y = \beta_0 + \beta_1 x$. Přímka $y = \beta_0 + \beta_1 x$, jejíž parametry minimalizují rozptyl $q(\beta_0, \beta_1)$ v celém dvourozměrném prostoru, se nazývá *regresní přímka znaku Y na znak X*. *Regresní odhad* i–té hodnoty znaku Y značíme $\hat{y}_i = b_0 + b_1 x_i$, $i = 1, \ldots, n$. Kvadrát koeficientu korelace znaků X, Y se nazývá *index determinace* a značí se ID^2. (Index determinace udává, jakou část variability hodnot znaku Y vystihuje regresní přímka. Nabývá hodnot z intervalu $\langle 0, 1 \rangle$. Čím je bližší 1, tím lépe vystihuje regresní přímka závislost Y na X.)

4.3. Věta
Nechť $y = b_0 + b_1 x$ je regresní přímka znaku Y na znak X. Pak použitím metody nejmenších čtverců dostaneme:

$$b_1 = \frac{s_{12}}{s_1^2}, \qquad b_0 = m_2 - \frac{s_{12}}{s_1^2} \cdot m_1,$$

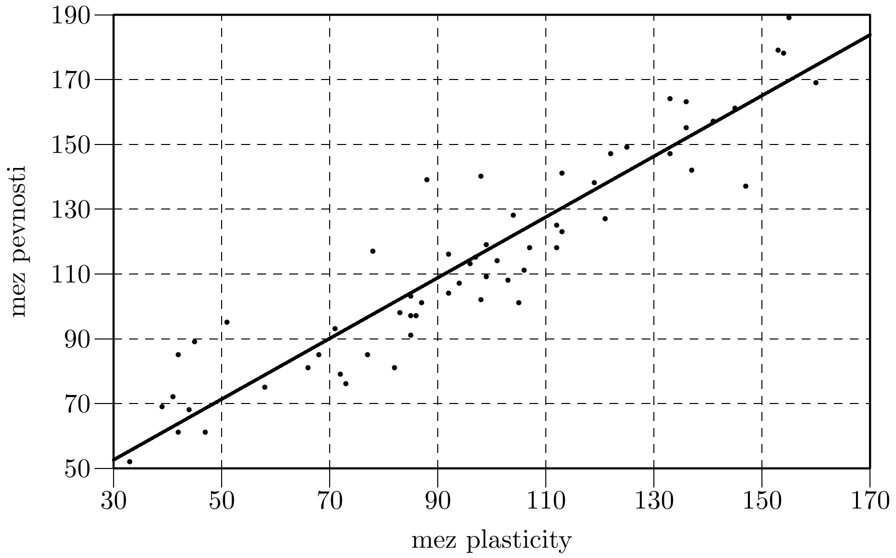
tedy $y = m_2 + \frac{s_{12}}{s_1^2}(x - m_1)$. Přitom úsek b_0 regresní přímky udává velikost jejího posunutí na svislé ose (tj. udává, jaký je regresní odhad hodnoty znaku Y, nabývá-li znak X hodnoty 0) a směrnice b_1 udává, o kolik jednotek se změní hodnota znaku Y, změní-li se hodnota znaku X o jednotku. Jestliže je $b_1 > 0$, dochází s růstem X k růstu Y a hovoříme o přímé závislosti hodnot znaku Y na hodnotách znaku X. Je-li $b_1 < 0$, dochází s růstem X k poklesu Y a hovoříme o nepřímé závislosti hodnot znaku Y na hodnotách znaku X.

4.4. Příklad
Pro datový soubor z příkladu 2.13

a) určete regresní přímku meze pevnosti na mez plasticity.
b) Zakreslete regresní přímku do dvourozměrného tečkového diagramu.
c) Jak se změní mez pevnosti, vzroste-li mez plasticity o jednotku?
d) Najděte regresní odhad meze pevnosti pro mez plasticity = 60.
e) Vypočtěte index determinace a interpretujte ho.

Řešení:
ad a) Na základě výsledků příkladu 3.17 dostáváme: $b_1 = \frac{s_{12}}{s_1^2} = \frac{985{,}76}{1052{,}4}$; $b_0 = m_2 - b_1 m_1 = 114{,}4 - 0{,}937 \cdot 95{,}9 = 24{,}5$; $y = 24{,}5 + 0{,}937x$.

ad b)



Povšimněte si, že koeficient korelace znaků X, Y vypočtený v příkladě 3.17 činil 0,936. Tato hodnota je blízká 1, což svědčí o silné přímé lineární závislosti mezi znaky X a Y. Tečky v dvourozměrném tečkovém diagramu nejsou příliš rozptýleny kolem regresní přímky.

ad c) Mez pevnosti vzroste o $0{,}937\,\mathrm{kp\,cm^{-2}}$.

ad d) $= 24{,}5 + 0{,}937 \cdot 60 = 80{,}72$.

ad e) $\mathrm{ID}^2 = r_{12}^2 = 0{,}936^2 = 0{,}876$. Znamená to, že 87,6% variability hodnot meze pevnosti je vysvětleno regresní přímkou.

4.5. Definice
Regresní přímkou znaku X na znak Y nazveme tu přímku $x = \overline{b_0} + \overline{b_1}y$, jejíž parametry minimalizují rozptyl

$$q(\overline{\beta_0}, \overline{\beta_1}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{\beta_0} - \overline{\beta_1}y_i)^2$$

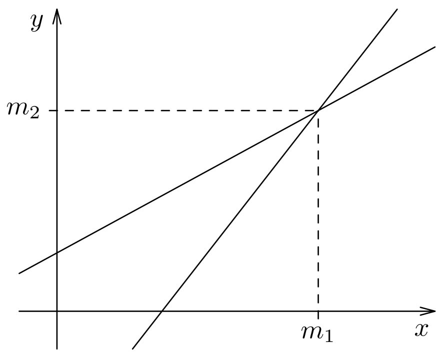
v celé rovině. Nazývá se též *druhá regresní přímka*. Regresní přímka znaku Y na znak X a regresní přímka znaku X na znak Y se nazývají *sdružené regresní přímky*.

4.6. Věta
Rovnice regresní přímky znaku X na znak Y má tvar $x = m_1 + \frac{s_{12}}{s_2^2}(y - m_2)$. Sdružené regresní přímky se protínají v bodě (m_1, m_2). Pro regresní parametry b_1, $\overline{b_1}$ platí: $b_1\overline{b_1} = r_{12}^2$. Rovnice sdružených regresních přímek můžeme psát ve tvaru

$$y = m_2 + r_{12}\frac{s_2}{s_1}(x - m_1), \qquad y = m_2 + \frac{1}{r_{12}}\frac{s_2}{s_1}(x - m_1), \qquad (\text{je-li } r_{12} \neq 0).$$

Regresní přímky svírají tím menší úhel, čím méně se od sebe liší r_{12} a $\frac{1}{r_{12}}$. Regresní přímky splynou, je-li $r_{12}^2 = 1$. K tomu dojde právě tehdy, existuje-li mezi X a Y úplná lineární závislost. Všechny body (x_i, y_i), $i = 1, \dots, n$ leží na jedné přímce, tedy ze znalosti x_i můžeme přesně vypočítat y_i, $i = 1, \dots, n$. Jsou-li znaky X, Y nekorelované, pak mají sdružené regresní přímky rovnice $y = m_2$, $x = m_1$ a jsou na sebe kolmé. Označíme-li α úhel, který svírají sdružené regresní přímky, pak platí:

- $\cos\alpha = 0$, právě když mezi X a Y neexistuje žádná lineární závislost,
- $\cos\alpha = 1$, právě když mezi X a Y existuje úplná přímá lineární závislost,
- $\cos\alpha = -1$, právě když mezi X a Y existuje úplná nepřímá lineární závislost.



4.7. Příklad

Pro datový soubor z příkladu 2.13

a) Určete regresní přímku meze plasticity na mez pevnosti.
b) Zakreslete regresní přímku do dvourozměrného tečkového diagramu.

Řešení:

ad a) S využitím výsledků příkladu 3.17 dostáváme:

$$\overline{b_1} = \frac{s_{12}}{s_2^2} = \frac{985{,}76}{1057{,}21} = 0{,}932,$$
$$\overline{b_0} = m_1 - \overline{b_1} m_2 = 95{,}9 - 0{,}932 \cdot 114{,}4 = -10{,}7,$$

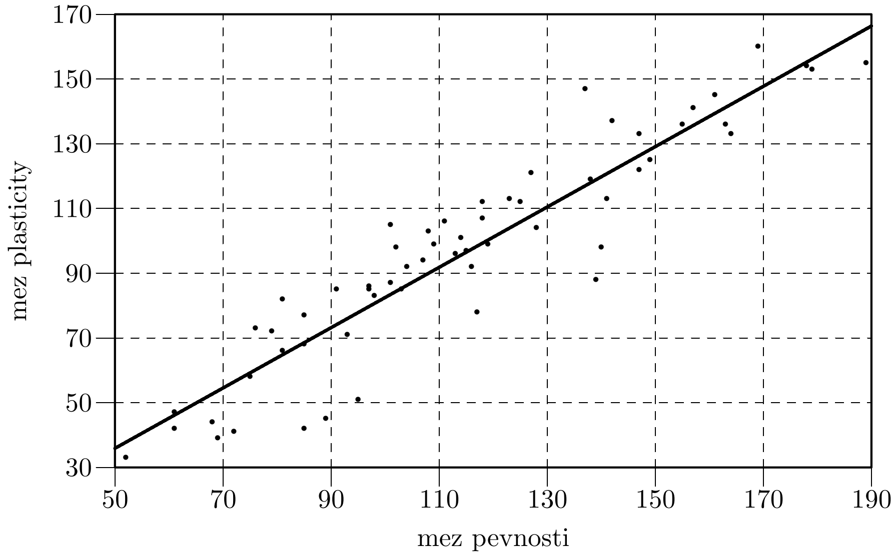
tedy

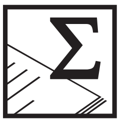
$$x = -10{,}7 + 0{,}932y.$$

ad b) Uvědomte si, že součin směrnic sdružených regresních přímek je

$$0{,}937 \cdot 0{,}932 = 0{,}87,$$

což je index derminace naboli kvadrát indexu korelace.

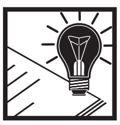




Shrnutí kapitoly

Pokud vzhled dvourozměrného tečkového diagramu svědčí o existenci určitého stupně lineární závislosti znaku Y na znaku X, můžeme diagramem proložit **regresní přímku** znaku Y na znak X. (Pozor – nelze se spokojit pouze s výpočtem korelačního koeficientu, je nutné grafické posouzení závislosti.) Její parametry (tj. posunutí a směrnici) odhadujeme **metodou nejmenších čtverců**. Kvalitu proložení posuzujeme pomocí **indexu determinace** – čím je tento index bližší 1, tím je regresní přímka výstižnější a čím je bližší 0, tím je regresní přímka nevhodnější pro vystižení závislosti Y na X. Dosadíme-li danou hodnotu znaku X do rovnice regresní přímky, získáme **regresní odhad** příslušné hodnoty znaku Y.

Má-li smysl zkoumat též opačný směr závislosti, tj. X na Y, hledáme **druhou regresní přímku**. 1. a 2. regresní přímka se označují jako **sdružené regresní přímky**.



Kontrolní otázky a úkoly

1 V čem spočívá princip metody nejmenších čtverců?
2 Uveďte příklad dvourozměrného datového souboru z ekonomické praxe vhodný pro použití regresní přímky.
3 Co vyjadřuje index determinace a jak se počítá?
4 Jaký je vztah mezi směrnicemi sdružených regresních přímek
5 Jsou-li sdružené regresní přímky kolmé, co lze říct o znacích X a Y?
6 Rozhodněte, zda přímky $y = 13 - 2x$, $x = 8 - y$ mohou být sdruženými regresními přímkami.
7 Je dána rovnice regresní přímky $y = 87 + 0{,}3(x - 25)$ a koeficient korelace $r_{12} = 0{,}77$. Najděte rovnici sdružené regresní přímky.

8 (S) U osmi náhodně vybraných studentů byly zjišťovány jejich matematické a verbální schopnosti. Výsledky matematického testu udává znak X, výsledky verbálního Y.

X	80	50	36	58	72	60	56	68
Y	65	60	35	39	48	44	48	61

a) Vypočtěte koeficient korelace a interpretujte ho.
b) Najděte rovnice sdružených regresních přímek.
c) Zlepší-li se výsledek v matematickém testu o 10 bodů, o kolik bodů se zlepší výsledek ve verbálním testu?
d) Zlepší-li se výsledek ve verbálním testu o 10 bodů, o kolik bodů se zlepší výsledek v matematickém testu?

9 Jak se změní úsek a směrnice regresní přímky, když každou hodnotu závisle proměnného znaku zvětšíme o 10%?

10 Závislost mezi vnější teplotou a teplotou ve skladišti je popsána regresní přímkou $y = 8 + 0{,}6x$. Při jaké vnější teplotě klesne teplota ve skladišti pod bod mrazu?

5. Jev a jeho pravděpodobnost

Cíl kapitoly

Po prostudování této kapitoly budete umět

- rozlišit náhodný a deterministický pokus
- stanovit základní prostor
- popsat vztahy mezi jevy pomocí množinových operací
- vypočítat pravděpodobnost jevu a znát vlastnosti pravděpodobnosti

Časová zátěž

Na prostudování této kapitoly budete potřebovat asi 6 hodin.

Nejprve se seznámíme s pojmem pokusu, a to deterministického a náhodného pokusu. Nadále se budeme zabývat náhodnými pokusy. Množinu možných výsledků pokusu považujeme za základní prostor. Na základním prostoru vybudujeme jevové pole jako systém podmnožin, který je uzavřený vzhledem k množinovým operacím. Základní prostor spolu s jevovým polem tvoří tzv. měřitelný prostor. Libovolná podmnožina možných výsledků náhodného pokusu, která patří do jevového pole, je jev. Naučíme se vyjadřovat vztahy mezi jevy pomocí množinových operací a uvedeme vlastnosti těchto operací.

5.1. Definice
Pokusem rozumíme jednorázové uskutečnění konstantně vymezeného souboru definičních podmínek. Předpokládáme, že pokus můžeme mnohonásobně nezávisle opakovat za dodržení definičních podmínek (ostatní podmínky se mohou měnit, proto různá opakování pokusu mohou vést k různým výsledkům). Dále předpokládáme, že opakováním pokusu vzniká opět pokus.

Deterministickým pokusem nazýváme takový pokus, jehož každé opakování vede k jedinému možnému výsledku. (Např. zahřívání vody na 100 °C při atmosférickém tlaku 1015 hPa vede k varu vody.)

Náhodným pokusem nazýváme takový pokus, jehož každé opakování vede k právě jednomu z více možných výsledků, které jsou vzájemně neslučitelné. (Např. hod kostkou vede k právě jednomu ze šesti možných výsledků.)

5.2. Definice
Neprázdnou množinu možných výsledků náhodného pokusu značíme Ω a nazýváme ji *základní prostor*. Možné výsledky značíme $\omega_1, \omega_2, \dots$. Na základním prostoru Ω vytvoříme *jevové pole* $\mathcal{A}$ jako systém podmnožin, který s každými dvěma množinami obsahuje i jejich rozdíl, obsahuje celý základní prostor a obsahuje-li každou ze spočetné posloupnosti množin, obsahuje i jejich spočetné sjednocení (znamená to, že systém $\mathcal{A}$ je uzavřený vzhledem k množinovým operacím). Jestliže $A \in \mathcal{A}$, pak řekneme, že A je jev. Dvojice $(\Omega, \mathcal{A})$ se nazývá *měřitelný prostor*. Ω se nazývá *jistý jev*, $\emptyset$ *nemožný jev*.

5.3. Poznámka

Vztahy mezi jevy vyjadřujeme pomocí množinových inkluzí a operace s jevy popisujeme pomocí množinových operací.

a) $A \subseteq B$ znamená, že jev A má za důsledek jev B.
b) $A \cup B$ znamená nastoupení aspoň jednoho z jevů A, B.
c) $A \cap B$ znamená společné nastoupení jevů A, B.
d) $A - B$ znamená nastoupení jevu A za nenastoupení jevu B.
e) $\overline{A} = \Omega - A$ znamená jev opačný k jevu A.
f) $A \cap B = \emptyset$ znamená, že jevy A, B jsou neslučitelné.
g) $\omega \in A$ znamená, že možný výsledek ω je příznivý nastoupení jevu A.

5.4. Věta

Uveďme některé vlastnosti, které mají operace s jevy:

a) Pro sjednocení a průnik jevů platí komutativní zákon, který pro dva jevy A, B má tvar:

$$A \cup B = B \cup A, \qquad A \cap B = B \cap A.$$

b) Pro sjednocení a průnik tří jevů A, B, C platí zákon asociativní:

$$A \cup (B \cup C) = (A \cup B) \cup C, \qquad A \cap (B \cap C) = (A \cap B) \cap C,$$

a zákon distributivní:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C), \qquad A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$$

c) Pro sjednocení a průnik jevů opačných platí de Morganovy zákony, které pro dva jevy A, B zapíšeme takto:

$$\overline{A} \cup \overline{B} = \overline{A \cap B}, \qquad \overline{A} \cap \overline{B} = \overline{A \cup B}.$$

5.5. Příklad

Náhodný pokus spočívá v hodu kostkou. Jev A znamená, že padne sudé číslo a jev B znamená, že padne číslo větší než 4.

a) Určete základní prostor Ω.
b) Vypište možné výsledky příznivé nastoupení jevů A, B.
c) Pomocí operací s jevy vyjádřete následující jevy: padne liché číslo; nepadne číslo 1 ani 3, padne číslo 6; padne číslo 2 nebo 4.

Řešení:

ad a) $\Omega = \{\omega_1, \dots, \omega_6\}$, kde možný výsledek ω_i znamená, že padne číslo i, $i = 1, \dots, 6$.

ad b) $A = \{\omega_2, \omega_4, \omega_6\}$, $B = \{\omega_5, \omega_6\}$.

ad c) $\overline{A} = \{\omega_1, \omega_3, \omega_5\}$; $A \cup B = \{\omega_2, \omega_4, \omega_5, \omega_6\}$; $A \cap B = \{\omega_6\}$; $A - B = \{\omega_2, \omega_4\}$

Na měřitelném prostoru zavedeme pravděpodobnost jako funkci, která splňuje určité axiomy a každému jevu přiřazuje číslo mezi 0 a 1. Měřitelný prostor spolu s pravděpodobností tvoří pravděpodobnostní prostor. Seznámíme se s vlastnostmi pravděpodobnosti a uvidíme, že téměř všechny jsou obdobné vlastnostem relativní četnosti jak jsme je poznali v první kapitole. Zavedeme speciální případ pravděpodobnosti – klasickou pravděpodobnost a vypočítáme několik příkladů.

5.6. Definice
Nechť $(\Omega, \mathcal{A})$ je měřitelný prostor. *Pravděpodobností* rozumíme reálnou množinovou funkci $P : \mathcal{A} \to \mathbb{R}$, která splňuje následující tři axiómy: každému jevu přiřazuje nezáporné číslo, jistému jevu přiřazuje číslo 1, sjednocení neslučitelných jevů přiřazuje součet pravděpodobností těchto jevů. Trojice $(\Omega, \mathcal{A}, P)$ se nazývá *pravděpodobnostní prostor.*

(Axiómy pravděpodobnosti jsou zvoleny tak, aby pravděpodobnost byla „zidealizovaným“ protějškem relativní četnosti zavedené v definici 1.1. Znamená to, že pro velký počet opakování pokusu, v němž sledujeme nastoupení jevu A, se relativní četnost jevu A blíží pravděpodobnosti jevu A. Tento poznatek je znám jako *empirický zákon velkých čísel.* Zdálo by se přirozené definovat pravděpodobnost jako limitu relativní četnosti pro $n \to \infty$. Tento postup by však nebyl korektní, protože počet pokusů n je vždy konečný a nelze se tedy přesvědčit o existenci uvedené limity.)

5.7. Věta
Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor. Pak pro libovolné jevy $A, A_1, A_2, \dots \in \mathcal{A}$ platí následujících 14 vlastností:

P1: $P(\emptyset) = 0$
P2: $P(A) \geq 0$ (nezápornost – axióm)
P3: $P(A_1 \cup A_2) + P(A_1 \cap A_2) = P(A_1) + P(A_2)$
P4: $1 + P(A_1 \cap A_2) \geq P(A_1) + P(A_2)$
P5: $P(A_1 \cup A_2) \leq P(A_1) + P(A_2)$ (subaditivita)
P6: $A_1 \cap A_2 = \emptyset \;\Rightarrow\; P(A_1 \cup A_2) = P(A_1) + P(A_2)$ (aditivita)
P7: $P(A_2 - A_1) = P(A_2) - P(A_1 \cap A_2)$
P8: $A_1 \subseteq A_2 \;\Rightarrow\; P(A_2 - A_1) = P(A_2) - P(A_2)$ (subtraktivita)
P9: $A_1 \subseteq A_2 \;\Rightarrow\; P(A_2) \leq P(A_2)$ (monotonie)
P10: $P(\Omega) = 1$ (normovanost – axióm)
P11: $P(A) + P(\overline{A}) = 1$ (komplementarita)
P12: $P(A) \leq 1$
P13: $A_i \cap A_j = \emptyset$ pro $i \neq j \;\Rightarrow\; P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots$
(spočetná aditivita – axióm)
P14:

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} P(A_i \cap A_j) +$$
$$+ \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^{n} P(A_i \cap A_j \cap A_k) - \dots + (-1)^{n-1} P(A_1 \cap A_2 \cap \dots \cap A_n)$$

Pro neslučitelné jevy $A_1, \ldots, A_n$ dostáváme

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

(Vlastnosti P1,…, P12 odpovídají vlastnostem relativní četnosti z věty 1.3, vlastnost P14 je známa jako věta o sčítání pravděpodobností.)

5.8. Definice
Nechť Ω je konečný základní prostor a nechť všechny možné výsledky mají stejnou šanci nastat. Klasická pravděpodobnost je funkce, která jevu A přiřazuje číslo $P(A) = \dfrac{m(A)}{m(\Omega)}$, kde $m(A)$ je počet možných výsledků příznivých nastoupení jevu A a $m(\Omega)$ je počet všech možných výsledků.

5.9. Příklad
Vypočítejte pravděpodobnosti jevů A, B, $\overline{A}$, $A \cup B$, $A \cap B$, $A - B$ z příkladu 5.5.

Řešení:
$m(\Omega) = 6$, $\quad P(A) = \frac{3}{6} = \frac{1}{2}$, $\quad P(B) = \frac{2}{6} = \frac{1}{3}$, $\quad P(\overline{A}) = \frac{3}{6} = \frac{1}{2}$,
$P(A \cup B) = \frac{4}{6} = \frac{2}{3}$, $\quad P(A \cap B) = \frac{1}{6}$, $\quad P(A - B) = \frac{2}{6} = \frac{1}{3}$.

5.10. Příklad
V dodávce 100 kusů výrobků nemá požadovaný průměr 10 kusů, požadovanou délku 20 kusů a současně nemá požadovaný průměr i délku 5 kusů. Jaká je pravděpodobnost, že náhodně vybraný výrobek z této dodávky má požadovaný průměr i délku?

Řešení:
Jev A spočívá v tom, že výrobek má požadovaný průměr a jev B v tom, že výrobek má požadovanou délku. Počítáme

$$\begin{aligned} P(A \cap B) &= P(\overline{\overline{A} \cup \overline{B}}) = 1 - P(\overline{A} \cup \overline{B}) = \\ &= 1 - [P(\overline{A}) + P(\overline{B}) - P(\overline{A} \cap \overline{B})] = 1 - \left(\frac{10}{100} + \frac{20}{100} - \frac{5}{100}\right) = 0{,}75. \end{aligned}$$

5.11. Příklad
Mezi N výrobky je M zmetků. Náhodně bez vracení vybereme n výrobků. Jaká je pravděpodobnost, že vybereme právě k zmetků?

Řešení:
Základní prostor Ω je tvořen všemi neuspořádanými n–ticemi vytvořenými z N prvků. Tedy $m(\Omega) = \binom{N}{n}$. Jev A spočívá v tom, že vybereme právě k zmetků z M zmetků (ty lze vybrat $\binom{M}{k}$ způsoby) a výběr doplníme $n - k$

kvalitními výrobky vybranými z $N-M$ kvalitních výrobků (tento výběr lze provést $\binom{N-M}{n-k}$ způsoby). Podle kombinatorického pravidla součinu dostáváme

$$m(A) = \binom{M}{k}\binom{N-M}{n-k}, \quad \text{tedy} \quad P(A) = \frac{m(A)}{m(\Omega)} = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}.$$

Shrnutí kapitoly

Deterministický pokus vede při každém opakování k jedinému možnému výsledku, zatímco **náhodný pokus** vede při každém opakování právě k jednomu z více možných výsledků. Množina možných výsledků náhodného pokusu tvoří **základní prostor**. Systém podmnožin základního prostoru, který je uzavřený vzhledem k množinovým operacím, se nazývá **jevové pole**. Základní prostor spolu s jevovým polem označujeme jako **měřitelný prostor**. Podmnožina, která patří do jevového pole, je jev. Celý základní prostor je **jevem jistým**, prázdná množina **jevem nemožným**.

Šanci jevu na uskutečnění vyjadřujeme pomocí **pravděpodobnosti**, což je funkce, která každému jevu přiřazuje číslo mezi 0 a 1 a splňuje určité axiomy, které stanovil ruský matematik A. N. Kolmogorov tak, aby pravděpodobnost byla „zidealizovaným" protějškem relativní četnosti. Při mnohonásobném nezávislém opakování téhož náhodného pokusu totiž platí **empirický zákon velkých čísel**: relativní četnost jevu se ustaluje kolem nějaké konstanty, kterou považujeme za pravděpodobnost tohoto jevu. Měřitelný prostor spolu s pravděpodobností tvoří **pravděpodobnostní prostor**. V praxi se nejčastěji používá **klasická pravděpodobnost** zavedená jako podíl počtu těch výsledků, které jsou příznivé nastoupení daného jevu, a počtu všech možných výsledků.

Kontrolní otázky a úkoly

1. Uveďte příklad deterministického pokusu a náhodného pokusu.
2. Náhodný pokus spočívá v hodu dvěma kostkami. Určete základní prostor.
3. Pro zkoušku provozní spolehlivosti určitého zařízení je předepsán tento postup: zařízení je uvedeno v činnost pětkrát při maximálním zatížení. Jakmile při některém z těchto pěti pokusů zařízení selže, nesplnilo podmínky zkoušky. Označme A_i jev: „při i–tém pokusu zařízení selhalo" pro $i = 1, \dots, 5$. Pomocí jevů A_i vyjádřete jevy:
   a) Zařízení neprošlo úspěšně zkouškou.
   b) První tři pokusy byly úspěšné, ve 4. a 5. pokusu zařízení selhalo.
   c) 1. a 5. pokus byly úspěšné, ale zkouška byla neúspěšná.
4. Formulujte emprický zákon velkých čísel.
5. Uveďte příklad situace, v níž nelze použít klasickou pravděpodobnost.
6. Z karetní hry o 32 kartách vybereme náhodně bez vracení 4 karty. Jaká je pravděpodobnost, že aspoň jedna z nich je eso?

7 Dva hráči házejí střídavě mincí. Vyhrává ten, komu padne dřív líc. Stanovte pravděpodobnost výhry 1. hráče a pravděpodobnost výhry 2. hráče.

8 Chevalier de Méré pozoroval, že při házení třemi kostkami padá součet 11 častěji než součet 12, i když podle jeho názoru (nesprávného) mají oba součty stejnou pravděpodobnost. Stanovte pravděpodobnost obou jevů.

9 Student se ke zkoušce připravil na 15 otázek z 20 zadaných. Při zkoušce si vybere náhodně dvě otázky. Jaká je pravděpodobnost, že aspoň na jednu zná odpověď?

10 Mezi následujícími tvrzeními vyberte ta, která jsou pravdivá:

a) $P(A \cap B) \leq P(B)$,
b) $P(A \cup B) < P(B)$,
c) $P(A \cup B) \leq P(A) + P(B)$,
d) $P(A) < 0$.

6. Stochasticky nezávislé jevy a podmíněná pravděpodobnost

Cíl kapitoly

Po prostudování této kapitoly budete umět

- ověřit stochastickou nezávislost posloupnosti jevů
- řešit příklady využívající stochastickou nezávislost jevů
- počítat podmíněnou pravděpodobnost
- použít větu o násobení pravděpodobností, vzorec pro úplnou pravděpodobnost a Bayesův vzorec

Časová zátěž

Pro zvládnutí této kapitoly budete potřebovat asi 6 hodin studia.

Z předešlé kapitoly víme, že pravděpodobnost je „zidealizovaným“ protějškem relativní četnosti. Lze tedy očekávat, že stochasticky nezávislé jevy zavedeme podobně jako četnostně nezávislé množiny: pomocí multiplikativního vztahu. Uvedeme vlastnosti stochasticky nezávislých jevů a s jejich pomocí odvodíme dvě důležitá rozložení pravděpodobnosti – geometrické a binomické, která mají, jak uvidíme později, časté využití v praxi.

6.1. Definice
Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor. Jevy $A_1, A_2 \in \mathcal{A}$ jsou *stochasticky nezávislé*, jestliže $P(A_1 \cap A_2) = P(A_1)P(A_2)$. (Tento vztah znamená, že informace o nastoupení jednoho jevu neovlivní šance, s nimiž očekáváme nastoupení druhého jevu. Stochastická nezávislost jevů A_1, A_2 je motivována četnostní nezávislostí množin G_1, G_2 ve výběrovém souboru – viz definice 1.6.) Jevy $A_1, \dots, A_n \in \mathcal{A}$ jsou *stochasticky nezávislé*, jestliže platí systém multiplikativních vztahů:

$$\begin{gathered}
\forall 1 \le i < j \le n: \quad P(A_i \cap A_j) = P(A_i)P(A_j), \\
\forall 1 \le i < j < k \le n: \quad P(A_i \cap A_j \cap A_k) = P(A_i)P(A_j)P(A_k), \\
\vdots \\
P(A_1 \cap \dots \cap A_n) = P(A_1) \dots P(A_n).
\end{gathered}$$

Jevy $A_1, A_2, \dots \in \mathcal{A}$ jsou *stochasticky nezávislé*, jestliže pro všechna přirozená n jsou stochasticky nezávislé jevy $A_1, \dots, A_n \in \mathcal{A}$.

(Upozornění: při ověřování stochastické nezávislosti jevů musíme prozkoumat platnost všech multiplikativních vztahů.)

6.2. Věta

a) Nemožný jev je stochasticky nezávislý s každým jevem.
b) Jistý jev je stochasticky nezávislý s každým jevem.
c) Stochastická nezávislost se neporuší, jestliže některé (nebo i všechny) jevy nahradíme jevy opačnými.
d) Neslučitelné jevy nemohou být stochasticky nezávislé (pokud nemají všechny nulovou pravděpodobnost).

6.3. Příklad

Nezávisle opakujeme týž náhodný pokus. Nechť jev A_i znamená úspěch v i–tém pokusu, přičemž $P(A_i) = \nu$, $i = 1, 2, \ldots$ Vypočítejte pravděpodobnost, že

a) prvnímu úspěchu předchází z neúspěchů, $z = 0, 1, 2, \ldots$,
b) v prvních n pokusech nastane právě y úspěchů, $y = 0, 1, \ldots, n$.

Řešení:

ad a) $P(\overline{A_1} \cap \cdots \cap \overline{A_z} \cap A_{z+1}) = P(\overline{A_1}) \ldots P(\overline{A_z}) P(A_{z+1}) = (1-\nu)^z \nu$ (geometrické rozložení pravděpodobností)

ad b)

$$\begin{aligned} P((A_1 \cap \cdots \cap A_y \cap \overline{A_{y+1}} \cap \cdots \cap \overline{A_n}) \cup \cdots \cup (\overline{A_1} \cap \cdots \cap \overline{A_{n-y}} \cap A_{n-y+1} \cap \cdots \cap A_n)) = \\ = P(A_1) \ldots P(A_y) P(\overline{A_{y+1}}) \ldots P(\overline{A_n}) + \cdots + \\ + P(\overline{A_1}) \ldots P(\overline{A_{n-y}}) P(A_{n-y+1}) \ldots P(A_n) = \\ = \nu^y (1-\nu)^{n-y} + \cdots + (1-\nu)^{n-y} \nu^y = \binom{n}{y} \nu^y (1-\nu)^{n-y} \end{aligned}$$

(binomické rozložení pravděpodobností)

Nyní zavedeme podmíněnou pravděpodobnost na základě analogie s podmíněnou relativní četností. Shrneme vlastnosti podmíněné pravděpodobnosti a naučíme se používat vzorec pro výpočet úplné pravděpodobnosti a Bayesův vzorec.

6.4. Definice

Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor a dále $H \in \mathcal{A}$ jev s nenulovou pravděpodobností. *Podmíněnou pravděpodobností* za podmínky H rozumíme funkci $P(.|H): \mathcal{A} \to \mathbb{R}$ danou vzorcem:

$$A \in \mathcal{A}: \; P(A|H) = \frac{P(A \cap H)}{P(H)}.$$

(Vysvětlení: Opakovaně nezávisle provádíme týž náhodný pokus a sledujeme nastoupení jevu A v těch pokusech, v nichž nastoupil jev H. Podmíněnou relativní četnost A za podmínky H jsme v definici 1.4 zavedli vztahem $p(A|H) = \frac{p(A \cap H)}{p(H)}$. Tato podmíněná relativní četnost se s rostoucím počtem pokusů ustaluje kolem konstanty $P(A|H)$, kterou považujeme za podmíněnou pravděpodobnost jevu A za podmínky H.)

6.5. Věta

Pro podmíněnou pravděpodobnost platí:

a) $P(A_1 \cap A_2) = P(A_1) P(A_2|A_1)$ pro $P(A_1) \neq 0$.
b) $P(A_1 \cap A_2) = P(A_2) P(A_1|A_2)$ pro $P(A_2) \neq 0$.
c) $P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1) P(A_2|A_1) P(A_3|A_1 \cap A_2) \ldots P(A_n|A_1 \cap \cdots \cap A_{n-1})$ pro $P(A_1 \cap \cdots \cap A_{n-1}) \neq 0$. (Věta o násobení pravděpodobností)

d) Jevy A_1, A_2 jsou stochasticky nezávislé, právě když $P(A_1|A_2) = P(A_1)$ nebo $P(A_2) = 0$ a právě když $P(A_2|A_1) = P(A_2)$ nebo $P(A_1) = 0$.

6.6. Příklad

Ze skupiny 100 výrobků, která obsahuje 10 zmetků, vybereme náhodně bez vracení 3 výrobky. Vypočtěte pravděpodobnost jevu, že první dva výrobky budou kvalitní a třetí bude zmetek.

Řešení:

Jev A_i znamená, že i–tý vybraný výrobek je kvalitní, $i = 1, 2, 3$. Počítáme $P(A_1 \cap A_2 \cap \overline{A_3}) = P(A_1)P(A_2|A_1)P(\overline{A_3}|A_1 \cap A_2) = \frac{90}{100} \cdot \frac{89}{99} \cdot \frac{10}{98} = 0{,}083$.

6.7. Věta

Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor, $H_1, \ldots, H_n \in \mathcal{A}$ takové jevy, že $P(H_i) > 0$, $\bigcup\limits_{i=1}^{n} H_i = \Omega$, $H_i \cap H_j = \emptyset$ pro $i \neq j$ (říkáme, že jevy $H_1, \ldots, H_n$ tvoří úplný systém hypotéz).

a) Pro libovolný jev $A \in \mathcal{A}$ platí vzorec úplné pravděpodobnosti:

$$P(A) = \sum_{i=1}^{n} P(H_i)P(A|H_i).$$

b) Pro libovolnou hypotézu H_k, $k = 1, \ldots, n$ a jev $A \in \mathcal{A}$ s nenulovou pravděpodobností platí Bayesův vzorec:

$$P(H_k|A) = \frac{P(H_k)P(A|H_k)}{P(A)}.$$

($P(H_k|A)$ se nazývá *aposteriorní pravděpodobnost* hypotézy H_k, $P(H_k)$ je *apriorní pravděpodobnost.*)

6.8. Příklad

Je známo, že 90% výrobků odpovídá standardu. Byla vypracována zjednodušená kontrolní zkouška, která u standardního výrobku dá kladný výsledek s pravděpodobností 0,95, zatímco u výrobku nestandardního s pravděpodobností 0,2. Jaká je pravděpodobnost, že

a) zkouška u náhodně vybraného výrobku dopadla kladně,
b) výrobek, u něhož zkouška dopadla kladně, je standardní?

Řešení:

Jev A znamená, že zkouška u náhodně vybraného výrobku dopadla kladně, jev H_1 znamená, že výrobek je standardní, jev H_2 znamená, že výrobek není standardní, $P(H_1) = 0{,}9$, $P(H_2) = 0{,}1$, $P(A|H_1) = 0{,}95$, $P(A|H_2) = 0{,}2$.

ad a) $P(A) = P(H_1)P(A|H_1) + P(H_2)P(A|H_2) = 0{,}9 \cdot 0{,}95 + 0{,}1 \cdot 0{,}2 = 0{,}875$

ad b) $P(H_1|A) = \frac{P(H_1)P(A|H_1)}{P(A)} = \frac{0{,}9 \cdot 0{,}95}{0{,}875} = 0{,}98$.

Shrnutí kapitoly

Stochasticky nezávislé jevy jsou protipólem deterministicky závislých jevů: informace o nastoupení jednoho jevu nijak nemění šance, s nimiž očekáváme nastoupení druhého jevu. Formálně zavádíme stochastickou nezávislost jevů pomocí multiplikativních vztahů na základě analogie s četnostní nezávislostí množin. Pomocí stochasticky nezávislých jevů lze odvodit **geometrické** a **binomické rozložení pravděpodobností**. Obě tato rozložení se často používají v praxi.

Podmíněná relativní četnost motivuje zavedení **podmníněné pravděpodobnosti** – zkoumáme pravděpodobnost nastoupení nějakého jevu za podmínky, že nastal jiný jev. Podmíněná pravděpodobnost se vyskytuje v několika důležitých vzorcích, které umožňují řešit řadu příkladů. Jedná se o **větu o násobení pravděpodobností**, **vzorec pro výpočet úplné pravděpodobnosti** a **Bayesův vzorec**.

Kontrolní otázky a úkoly

1 Uveďte příklad stochasticky nezávislých jevů

2 Nechť $P(A) = p$, $P(B) = q$. Pomocí čísel p, q vyjádřete pravděpodobnost nastoupení aspoň jednoho z jevů A, B, jsou-li tyto jevy
  a) stochasticky nezávislé,
  b) neslučitelné.

3 Co lze říci o jevech A, B, které nejsou nemožné a platí pro ně:

$$P(A \cup B) = 1 - [1 - P(A)][1 - P(B)]?$$

4 Je pravděpodobnější vyhrát se stejně silným soupeřem tři partie ze čtyř nebo pět z osmi, když nerozhodný výsledek je vyloučen a výsledky jsou nezávislé?

5 První dělník vyrobí denně 60 výrobků, z toho 10% zmetků. Druhý dělník vyrobí denně 40 výrobků, z toho 5% zmetků. Jaká je pravděpodobnost, že náhodně vybraný výrobek z denní produkce je zmetek a pochází od prvního dělníka?

6 Ze šesti vajec jsou dvě prasklá. Náhodně vybereme dvě vejce. Jaká je pravděpodobnost, že budou
  a) obě prasklá,
  b) právě jedno prasklé,
  c) obě dobrá?

7 Doplňte chybějící člen x v rovnici $P(B) = P(B|A)P(A) + xP(\overline{A})$.

8 Pro jaké jevy A, B, $B \neq \emptyset$ platí $P(A|B) = P(A)$?

9 Co lze říci o jevech $A_1, \ldots, A_n$ s nenulovými pravděpodobnostmi, které jsou neslučitelné a jejich sjednocením je celý základní prostor?

10 Pojišťovací společnost rozlišuje při pojišťování tři skupiny řidičů – A, B a C. Pravděpodobnost toho, že řidič patřící do skupiny A bude mít během roku nehodu, je 0,03, zatímco u řidiče skupiny B je to 0,06 a u řidiče skupiny C 0,1. Podle dlouhodobých záznamů společnosti je 70% pojistných smluv uzavřeno s řidiči skupiny A, 20% s řidiči skupiny B a 10% s řidiči skupiny C. Jestliže došlo k nehodě řidiče pojištěného u této společnosti, jaká je pravděpodobnost, že patřil do skupiny C?

11 U jistého druhu elektrického spotřebiče se s pravděpodobností 0,01 vyskytuje výrobní vada. U spotřebiče s touto výrobní vadou dochází v záruční lhůtě k poruše s pravděpodobností 0,5. Výrobky, které tuto vadu nemají, se v záruční lhůtě porouchají s pravděpodobností 0,01. Jaká je pravděpodobnost, že

a) u náhodně vybraného výrobku nastane v záruční lhůtě porucha,
b) výrobek, který se v záruční lhůtě porouchá, bude mít dotyčnou výrobní vadu?

7. Náhodná veličina a její distribuční funkce

Cíl kapitoly

Po prostudování této kapitoly budete umět:

- číselně popsat výsledky náhodného pokusu pomocí náhodných veličina a náhodných vektorů,
- najít distribuční funkci náhodné veličiny či náhodného vektoru,
- rozlišit diskrétní a spojité náhodné veličiny a náhodné vektory a najít jejich funkcionální charakteristiky,
- ověřit stochastickou nezávislost náhodných veličin.

Časová zátěž

Na prostudování této kapitoly budete potřebovat asi 8 hodin studia.

Naučíme se, jak popisovat výsledky náhodného pokusu pomocí náhodné veličiny, tj. zobrazení, které možnému výsledku přiřadí číslo či několik čísel. Existuje zřetelná analogie mezi znakem, který známe z 1. kapitoly, a náhodnou veličinou. V některých situacích potřebujeme náhodnou veličinu transformovat. Získáme složenou funkci zvanou transformovaná náhodná veličina.

Statistika často zajímá pravděpodobnost jevu, že hodnota náhodné veličiny nepřesáhne nějakou mez. Pomocí této pravděpodobnosti zavedeme distribuční funkci, která je „zidealizovaným“ protějškem empirické distribuční funkce, s níž jsme se setkali ve 2. kapitole. Seznámíme se s vlastnostmi distribuční funkce a vyřešíme několik příkladů.

7.1. Definice
Libovolná funkce $X : \Omega \to \mathbb{R}$, která každému možnému výsledku $\omega \in \Omega$ přiřazuje reálné číslo $X(\omega)$, se nazývá *náhodná veličina* a číslo $X(\omega)$ je *číselná realizace náhodné veličiny* X *příslušná možnému výsledku* ω. Uspořádaná posloupnost náhodných veličin $(X_1, \dots, X_n)$ se nazývá *náhodný vektor* a značí se $\boldsymbol{X}$. Je-li $g : \mathbb{R} \to \mathbb{R}$ (resp. $(g_1, \dots, g_m) : \mathbb{R}^n \to \mathbb{R}^m$) funkce, pak složená funkce $Y = g(X)$ (resp. $\boldsymbol{Y} = (Y_1, \dots, Y_m) = (g_1(x_1, \dots, x_n), \dots, g_m(x_1, \dots, x_n)))$ se nazývá *transformovaná náhodná veličina* (resp. *transformovaný náhodný vektor*).

Vysvětlení: Náhodná veličina i náhodný vektor popisují výsledky náhodného pokusu pomocí reálných čísel. Musí přitom splňovat podmínku tzv. měřitelnosti, kterou se zde nebudeme zabývat. Náhodná veličina v počtu pravděpodobnosti a znak v popisné statistice – viz definice 1.8 – jsou sice pojmy blízké, nikoli však totožné. Znak lze považovat za náhodnou veličinu, pokud jeho hodnotu zjišťujeme na objektu, který byl vybrán ze základního souboru náhodně.

Upozornění: V dalším textu se omezíme na dvourozměrné náhodné vektory. Poznatky lze jednoduše zobecnit i na n-rozměrné náhodné vektory.

7.2. Označení
Nechť $B \subseteq \mathbb{R}$. Jev $\{\omega \in \Omega; X(\omega) \in B\}$ zkráceně zapisujeme $\{X \in B\}$ a čteme: náhodná veličina X se realizovala v množině B.

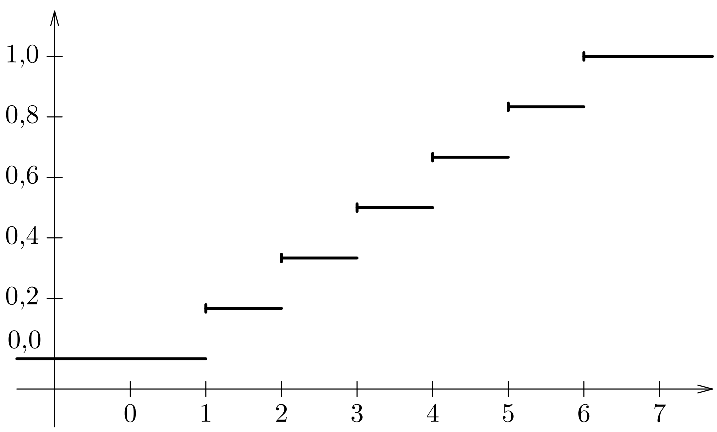
7.3. Definice
Pravděpodobnostní chování náhodné veličiny X (resp. náhodného vektoru $\boldsymbol{X} = (X_1, X_2)$) popisujeme *distribuční funkcí* $\Phi : \mathbb{R} \to \mathbb{R}$, která je dána vztahem: $\forall x \in \mathbb{R} : \Phi(x) = P(X \leq x)$ (resp. *simultánní distribuční funkcí* $\Phi : \mathbb{R}^2 \to \mathbb{R}$, která je definována vztahem: $\forall (x_1, x_2) \in \mathbb{R}^2 : \Phi(x_1, x_2) = P(X_1 \leq x_1, X_2 \leq x_2)$).

Vysvětlení: Distribuční funkce $\Phi(x)$ je zidealizovaným protějškem empirické distribuční funkce $F(x)$ zavedené v definici 2.4 či 2.14: $\forall x \in \mathbb{R} : F(x) = \frac{N(X \leq x)}{n}$. S rostoucím rozsahem výběrového souboru se budou hodnoty $F(x)$ ustalovat kolem hodnot $\Phi(x)$.

7.4. Příklad
Najděte distribuční funkci náhodné veličiny X, která udává, jaké číslo padlo při hodu kostkou a nakreslete graf této distribuční funkce.

Řešení:
Náhodná veličina X může nabývat hodnot $1, 2, 3, 4, 5, 6$. Číselnou osu tedy rozdělíme na 7 intervalů.

$$
\begin{aligned}
x \in (-\infty, 1) &: \Phi(x) = P(X \leq x) = 0 \\
x \in \langle 1, 2) &: \Phi(x) = P(X \leq x) = \frac{1}{6} \\
x \in \langle 2, 3) &: \Phi(x) = P(X \leq x) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} \\
x \in \langle 3, 4) &: \Phi(x) = P(X \leq x) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} \\
x \in \langle 4, 5) &: \Phi(x) = P(X \leq x) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{4}{6} \\
x \in \langle 5, 6) &: \Phi(x) = P(X \leq x) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{5}{6} \\
x \in \langle 6, \infty) &: \Phi(x) = P(X \leq x) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{6}{6} = 1
\end{aligned}
$$



7.5. Věta

a) Skalární případ: Distribuční funkce $\Phi(x)$ skalární náhodné veličiny X má následující vlastnosti:

- $\Phi(x)$ je neklesající,
- $\Phi(x)$ je zprava spojitá,
- $\Phi(x)$ je normovaná v tom smyslu, že $\lim\limits_{x\to-\infty}\Phi(x)=0$, $\lim\limits_{x\to\infty}\Phi(x)=1$,
- $\forall a,b\in\mathbb{R}, a<b$ platí: $P(a<x\le b)=\Phi(b)-\Phi(a)$,
- pro libovolné, ale pevně dané $x_0\in\mathbb{R}:\ P(X=x_0)=\Phi(x_0)-\lim\limits_{x\to x_0^-}\Phi(x)$.

b) Vektorový případ: Simultánní distribuční funkce $\Phi(x_1,x_2)$ náhodného vektoru $X=(X_1,X_2)$ má následující vlastnosti:

- $\Phi(x_1,x_2)$ je neklesající vzhledem ke každé jednotlivé proměnné,
- $\Phi(x_1,x_2)$ je zprava spojitá vzhledem ke každé jednotlivé proměnné,
- $\Phi(x_1,x_2)$ je normovaná v tom smyslu, že $\lim\limits_{x_1\to\infty,x_2\to\infty}\Phi(x_1,x_2)=1$, $\lim\limits_{x_1\to-\infty}\Phi(x_1,x_2)=\lim\limits_{x_2\to-\infty}\Phi(x_1,x_2)=0$,
- $\forall(x_1,x_2)\in\mathbb{R}^2, h_1>0, h_2>0:\ P(x_1<X_1\le x_1+h_1\wedge x_2<X_2\le x_2+h_2)=\Phi(x_1+h_1,x_2+h_2)-\Phi(x_1+h_1,x_2)-\Phi(x_1,x_2+h_2)+\Phi(x_1,x_2)$ (tato vlastnost vyjadřuje pravděpodobnost, že náhodný vektor se realizuje v obdélníku $(x_1,x_1+h_1\rangle\times(x_2,x_2+h_2\rangle)$,
- $\lim\limits_{x_2\to\infty}\Phi(x_1,x_2)=\Phi_1(x_1)$, $\lim\limits_{x_1\to\infty}\Phi(x_1,x_2)=\Phi_2(x_2)$, kde $\Phi_1(x_1)$, $\Phi_2(x_2)$ jsou distribuční funkce náhodných veličin X_1, X_2. Nazývají se marginální distribuční funkce.

7.6. Příklad

Náhodný vektor (X_1,X_2) má distribuční funkci

$$\Phi(x_1,x_2)=\frac{1}{\pi^2}\left(\operatorname{arctg}x_1+\frac{\pi}{2}\right)\left(\operatorname{arctg}x_2+\frac{\pi}{2}\right).$$

Vypočtěte pravděpodobnost, že náhodný vektor (X_1,X_2) se bude realizovat v jednotkovém čtverci $(0,1\rangle\times(0,1\rangle$. Najděte obě marginální distribuční funkce $\Phi_1(x_1)$, $\Phi_2(x_2)$.

Řešení:

Podle 4. vlastnosti v věty 7.5(b), kde $x_1=0$, $x_2=0$, $h_1=1$, $h_2=0$ dostáváme

$$\begin{aligned}P(0<X_1\le 1\wedge 0<X_2\le 1)&=\Phi(1,1)-\Phi(1,0)-\Phi(0,1)+\Phi(0,0)=\\&=\frac{1}{\pi^2}\left(\frac{\pi}{4}+\frac{\pi}{2}\right)\left(\frac{\pi}{4}+\frac{\pi}{2}\right)-\frac{1}{\pi^2}\left(\frac{\pi}{4}+\frac{\pi}{2}\right)\left(0+\frac{\pi}{2}\right)-\\&\quad-\frac{1}{\pi^2}\left(0+\frac{\pi}{2}\right)\left(\frac{\pi}{4}+\frac{\pi}{2}\right)+\frac{1}{\pi^2}\left(0+\frac{\pi}{2}\right)\left(0+\frac{\pi}{2}\right)=\frac{1}{16}.\end{aligned}$$

$$\Phi_1(x_1)=\lim_{x_2\to\infty}\frac{1}{\pi^2}\left(\operatorname{arctg}x_1+\frac{\pi}{2}\right)\left(\operatorname{arctg}x_2+\frac{\pi}{2}\right)=\frac{1}{\pi}\left(\operatorname{arctg}x_1+\frac{\pi}{2}\right)$$

$$\Phi_2(x_2)=\lim_{x_1\to\infty}\frac{1}{\pi^2}\left(\operatorname{arctg}x_1+\frac{\pi}{2}\right)\left(\operatorname{arctg}x_2+\frac{\pi}{2}\right)=\frac{1}{\pi}\left(\operatorname{arctg}x_2+\frac{\pi}{2}\right)$$

Nyní se budeme zabývat dvěma speciálními typy náhodných veličin, a to diskrétními a spojitými náhodnými veličinami. Diskrétní náhodná veličina nabývá nejvýše spočetně mnoha izolovaných hodnot, zatímco spojitá veličina nabývá všech hodnot z nějakého intervalu. Pravděpodobnostní chování diskrétní (resp. spojité) náhodné veličiny popíšeme pomocí pravděpodobnostní funkce (resp. pomocí hustoty pravděpodobnosti). Uvidíme, že vlastnosti pravděpodobnostní funkce jsou podobné jako vlastnosti četnostní funkce a vlastnosti hustoty pravděpodobnosti jsou analogické vlastnostem hustoty četnosti.

7.7. Definice

a) Skalární případ: Náhodná veličina X se nazývá *diskrétní*, jestliže její distribuční funkci lze vyjádřit pomocí nezáporné funkce $\pi(x)$ v součtovém tvaru:

$$\forall x \in \mathbb{R} : \Phi(x) = \sum_{t \leq x} \pi(x).$$

Funkce $\pi(x)$ se nazývá *pravděpodobnostní funkce diskrétní náhodné veličiny* X.

b) Vektorový případ: Náhodný vektor (X_1, X_2) se nazývá *diskrétní*, jestliže jeho simultánní distribuční funkci lze vyjádřit pomocí nezáporné funkce $\pi(x_1, x_2)$ v součtovém tvaru:

$$\forall (x_1, x_2)\mathbb{R}^2 : \Phi(x_1, x_2) = \sum_{t_1 \leq x_1} \sum_{t_2 \leq x_2} \pi(t_1, t_2).$$

Funkce $\pi(x_1, x_2)$ se nazývá *simultánní pravděpodobnostní funkce diskrétního náhodného vektoru* (X_1, X_2).

Vysvětlení: Pravděpodobnostní funkce $\pi(x)$ je zidealizovaným protějškem četnostní funkce $p(x)$ zavedené v definici 2.4: $\forall x \in \mathbb{R} : p(x) = \frac{N(X=x)}{n}$. S rostoucím rozsahem výběrového souboru se hodnoty četnostní funkce ustalují kolem hodnot pravděpodobnostní funkce. Diskrétní náhodná veličina nabývá nejvýše spočetně mnoha hodnot. Její distribuční funkce má schodovitý průběh – viz graf v příkladu 7.4.

Simultánní pravděpodobnostní funkce $\pi(x_1, x_2)$ je zidealizovaným protějškem simultánní četnostní funkce z definice 2.7: $\forall (x_1, x_2) \in \mathbb{R}^2 : p(x_1, x_2) = \frac{N((X_1=x_1) \wedge (X_2=x_2))}{n}$. S rostoucím rozsahem výběrového souboru se hodnoty simultánní pravděpodobnostní funkce ustalují kolem hodnot simultánní pravděpodobnostní funkce.

7.8. Věta

a) Skalární případ: Je-li $\pi(x)$ pravděpodobnostní funkce diskrétní náhodné veličiny X, pak platí:

- $\forall x \in \mathbb{R} : \pi(x) \geq 0$ (nezápornost),
- $\sum_{x=-\infty}^{\infty} \pi(x) = 1$ (normovanost),

- $\forall x \in \mathbb{R}: \ \pi(x) = P(X = x)$,
- $\forall B \subseteq \mathbb{R}: \ P(X \in B) = \sum\limits_{x \in B} \pi(x)$.

b) Vektorový případ: Je-li $\pi(x_1, x_2)$ simultánní pravděpodobnostní funkce diskrétního náhodného vektoru (X_1, X_2), pak platí:

- $\forall (x_1, x_2) \in \mathbb{R}^2: \ \pi(x_1, x_2) \geq 0$ (nezápornost),
- $\sum\limits_{x_1=-\infty}^{\infty} \sum\limits_{x_2=-\infty}^{\infty} \pi(x_1, x_2) = 1$ (normovanost),
- $\forall (x_1, x_2) \in \mathbb{R}^2: \ \pi(x_1, x_2) = P(X_1 = x_1 \wedge X_2 = x_2)$,
- $\forall B \subseteq \mathbb{R}^2: \ P((X_1, X_2) \in B) = \sum\limits_{(x_1,x_2) \in B} \pi(x_1, x_2)$,
- $\sum\limits_{x_2=-\infty}^{\infty} \pi(x_1, x_2) = \pi_1(x_1)$, $\sum\limits_{x_1=-\infty}^{\infty} \pi(x_1, x_2) = \pi_2(x_2)$, přičemž $\pi_1(x_1)$, $\pi_2(x_2)$ jsou marginální pravděpodobnostní funkce náhodných veličin X_1, X_2.

7.9. Příklad

Pravděpodobnost poruchy každé ze tří nezávisle pracujících výrobních linek je 0,5. Náhodná veličina X udává počet výrobních linek, které mají poruchu. Najděte pravděpodobnostní funkci náhodné veličiny X.

Řešení:

Náhodná veličina X nabývá hodnot $0, 1, 2, 3$.

$$\pi(0) = P(X = 0) = 0{,}5^3 = 0{,}125,$$
$$\pi(1) = P(X = 1) = 3 \cdot 0{,}5^3 = 0{,}375,$$
$$\pi(2) = P(X = 2) = 3 \cdot 0{,}5^3 = 0{,}375,$$
$$\pi(3) = P(X = 3) = 0{,}5^3 = 0{,}125,$$
$$\pi(x) = 0 \quad \text{jinak.}$$

7.10. Příklad

Je dán systém složený ze dvou bloků. Pravděpodobnost, že i-tý blok správně funguje, je ν_i, $i = 1, 2$, a pravděpodobnost, že správně fungují oba bloky, je ν_{12}. Nechť náhodná veličina X_i je ukazatel fungování i-tého bloku, tj.

$$X_i = \begin{cases} 1, & \text{pokud } i\text{-tý blok funguje,} \\ 0, & \text{pokud } i\text{-tý blok nefunguje,} \end{cases} \qquad i = 1, 2.$$

Najděte simultánní pravděpodobnostní funkci $\pi(x_1, x_2)$ náhodného vektoru (X_1, X_2) a obě marginální pravděpodobnostní funkce $\pi_1(x_1)$ a $\pi_2(x_2)$.

Řešení:

Hodnoty pravděpodobnostních funkcí zapíšeme do kontingenční tabulky.

x_i		x_2		$\pi_1(x_1)$
		0	1	
x_1	0	$1 - \nu_1 - \nu_2 + \nu_{12}$	$\nu_2 - \nu_{12}$	$1 - \nu_1$
	1	$\nu_1 - \nu_{12}$	ν_{12}	ν_1
$\pi_2(x_2)$		$1 - \nu_2$	ν_2	1

$$
\begin{aligned}
\pi(0,0) &= P(X_1 = 0 \wedge X_2 = 0) = 1 - P(X_1 = 1 \vee X_2 = 1) = \\
&\qquad = 1 - (\nu_1 + \nu_2 - \nu_{12}) = 1 - \nu_1 - \nu_2 + \nu_{12}, \\
\pi(0,1) &= P(X_1 = 0 \wedge X_2 = 1) = P(X_2 = 1) - P(X_1 = 1 \wedge X_2 = 1) = \\
&\qquad = \nu_2 - \nu_{12}, \\
\pi(1,0) &= P(X_1 = 1 \wedge X_2 = 0) = P(X_1 = 1) - P(X_1 = 1 \wedge X_2 = 1) = \\
&\qquad = \nu_1 - \nu_{12}, \\
\pi(1,1) &= P(X_1 = 1 \wedge X_2 = 1) = \nu_{12}, \\
\pi(x_1, x_2) &= 0 \qquad \text{jinak.}
\end{aligned}
$$

7.11. Definice

a) Skalární případ: Náhodná veličina X se nazývá *spojitá*, jestliže její distribuční funkci lze vyjádřit pomocí nezáporné funkce $\varphi(x)$ v integrálním tvaru :

$$\forall x \in \mathbb{R} : \ \Phi(x) = \int_{-\infty}^{x} \varphi(t)\, dt.$$

Funkce $\varphi(x)$ se nazývá *hustota pravděpodobnosti spojité náhodné veličiny* X.

b) Vektorový případ: Náhodný vektor (X_1, X_2) se nazývá *spojitý*, jestliže jeho simultánní distribuční funkci je možné vyjádřit pomocí nezáporné funkce $\varphi(x_1, x_2)$ v integrálním tvaru:

$$\forall (x_1, x_2) \in \mathbb{R}^2 : \ \Phi(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \varphi(t_1, t_2)\, dt_1 dt_2.$$

Funkce $\varphi(x_1, x_2)$ se nazývá *simultánní hustota pravděpodobnosti spojitého náhodného vektoru* (X_1, X_2).

Vysvětlení: Hustota pravděpodobnosti $\varphi(x)$ je zidealizovaným protějškem hustoty četnosti $f(x)$ zavedené v definici 2.14. S rostoucím rozsahem výběrového souboru a klesající šířkou třídicích intervalů se hodnoty hustoty četnosti ustalují kolem hodnot hustoty pravděpodobnosti. Spojitá náhodná veličina nabývá všech hodnot z nějakého intervalu. Její distribuční funkce je všude spojitá.

Simultánní hustota pravděpodobnosti je zidealizovaným protějškem simultánní hustoty četnosti zavedené v definici 2.17. S rostoucím rozsahem výběrového souboru a klesající plochou dvourozměrných třídicích intervalů se hodnoty simultánní hustoty pravděpodobnosti a ustalují kolem hodnot simultánní hustoty četnosti.

7.12. Věta

a) Skalární případ: Je-li $\varphi(x)$ hustota pravděpodobnosti spojité náhodné veličiny X, pak platí:

- $\forall x \in \mathbb{R}: \ \varphi(x) \geq 0$ (nezápornost)
- $\int\limits_{-\infty}^{\infty} \varphi(x)\, dx = 1$ (normovanost)
- $\forall x \in \mathbb{R}: \ P(X = x) = 0$
- $\forall B \subseteq \mathbb{R}: \ P(X \in B) = \int\limits_{x \in B} \varphi(x)\, dx$
- $\varphi(x) = \frac{d\Phi(x)}{dx}$ ve všech bodech spojitosti funkce $\varphi(x)$

b) Vektorový případ: Je-li $\varphi(x_1, x_2)$ simultánní hustota pravděpodobnosti spojitého náhodného vektoru (X_1, X_2), pak platí:

- $\forall (x_1, x_2) \in \mathbb{R}^2: \ \varphi(x_1, x_2) \geq 0$ (nezápornost)
- $\int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} \varphi(x_1, x_2)\, dx_1 dx_2 = 1$ (normovanost)
- $\forall (x_1, x_2) \in \mathbb{R}^2: P((X_1 = x_1) \wedge (X_2 = x_2)) = 0$
- $B \in \mathbb{R}^2: \ P((X_1, X_2) \in B) = \int\int\limits_{(x_1, x_2) \in B} \varphi(x_1, x_2)\, dx_1 dx_2$
- $\int\limits_{-\infty}^{\infty} \varphi(x_1, x_2)\, dx2 = \varphi_1(x_1)$, $\int\limits_{-\infty}^{\infty} \varphi(x_1, x_2)\, dx_1 = \varphi_2(x_2)$, přičemž $\varphi_1(x_1)$, $\varphi_2(x_2)$ jsou marginální hustoty pravděpodobnosti náhodných veličin X_1, X_2.

7.13. Příklad

Na automatické lince se plní láhve mlékem. Každá láhev má obsahovat přesně 1000 ml mléka, ale v důsledku působení náhodných vlivů množství mléka kolísá v intervalu (980 ml, 1020 ml). Každé množství mléka v tomto intervalu považujeme za stejně možné. Náhodná veličina X udává množství mléka v náhodně vybrané lahvi. Najděte její hustotu pravděpodobnosti $\varphi(x)$ a distribuční funkci $\Phi(x)$.

Řešení:

$$\varphi(x) = \begin{cases} k & \text{pro } x \in (980, 1020), \\ 0 & \text{jinak.} \end{cases}$$

Z normovanosti hustoty plyne: $1 = \int\limits_{980}^{1020} k\, dx = 40k$, tedy $k = \frac{1}{40}$. Pro distribuční funkci platí:

$$\Phi(x) = \begin{cases} 0 & \text{pro } x \leq 980, \\ \int\limits_{980}^{x} \frac{1}{40}\, dt = \frac{x-980}{40} & \text{pro } 980 < x < 1020, \\ 1 & \text{pro } x \geq 1020. \end{cases}$$

7.14. Příklad

Spojitý náhodný vektor (X_1, X_2) má simultánní hustotu pravděpodobnosti

$$\varphi(x_1, x_2) = \frac{1}{\pi^2 (1 + x_1^2)(1 + x_2)^2}.$$

Najděte obě marginální distribuční funkce $\varphi_1(x_1)$, $\varphi_2(x_2)$.

Řešení:

$$\varphi_1(x_1) = \int\limits_{-\infty}^{\infty} \frac{1}{\pi^2(1+x_1^2)(1+x_2)^2}\,dx_2 = \frac{1}{\pi^2(1+x_1^2)} \int\limits_{-\infty}^{\infty} \frac{1}{1+x_2^2}\,dx_2 =$$
$$= \frac{1}{\pi^2(1+x_1^2)}[\operatorname{arctg} x_2]_{-\infty}^{\infty} = \frac{1}{\pi^2(1+x_1^2)}\left(\frac{\pi}{2} - \left(-\frac{\pi}{2}\right)\right) = \frac{1}{\pi(1+x_1^2)}.$$

Analogicky dostáváme

$$\varphi_2(x_2) = \frac{1}{\pi(1+x_2^2)}.$$

V popisné statistice, konkrétně ve 2. kapitole, jsme se setkali s četnostní nezávislostí znaků v daném výběrovém souboru. V počtu pravděpodobnosti má tento pojem svou analogii ve stochastické nezávislosti náhodných veličin. Spočítáme několik příkladů, v nichž se vyskytují stochasticky nezávislé veličiny, a ukážeme si, že transformováním se stochastická nezávislost náhodných veličin neporuší.

7.15. Definice

a) Obecný případ: Řekneme, že náhodné veličiny $X_1, \dots, X_n$ s marginálními distribučními funkcemi $\Phi_1(x_1), \dots, \Phi_n(x_n)$ a simultánní distribuční funkcí $\Phi(x_1, \dots, x_n)$ jsou *stochasticky nezávislé*, jestliže pro $\forall(x_1, \dots, x_n) \in \mathbb{R}^n : \Phi(x_1, \dots, x_n) = \Phi_1(x_1) \cdot \dots \cdot \Phi_n(x_n)$.

b) Diskrétní případ: Řekneme, že diskrétní náhodné veličiny $X_1, \dots, X_n$ s marginálními pravděpodobnostními funkcemi $\pi_1(x_1), \dots, \pi_n(x_n)$ a simultánní pravděpodobnostní funkcí $\pi(x_1, \dots, x_n)$ jsou stochasticky nezávislé, jestliže pro $\forall(x_1, \dots, x_n) \in \mathbb{R}^n : \pi(x_1, \dots, x_n) = \pi_1(x_1) \cdot \dots \cdot \pi_n(x_n)$.

c) Spojitý případ: Řekneme, že spojité náhodné veličiny $X_1, \dots, X_n$ s marginálními hustotami pravděpodobnosti $\varphi_1(x_1), \dots, \varphi_n(x_n)$ a simultánní pravděpodobnostní funkcí $\varphi(x_1, \dots, x_n)$ jsou stochasticky nezávislé, jestliže pro $\forall(x_1, \dots, x_n) \in \mathbb{R}^n : \varphi(x_1, \dots, x_n) = \varphi_1(x_1) \cdot \dots \cdot \varphi_n(x_n)$ s případnou výjimkou na množině bodů neovlivňujících integraci.

Řekneme, že posloupnost $\{X_n\}_{n=1}^{\infty}$ je posloupností stochasticky nezávislých náhodných veličin, jestliže pro všechna přirozená n jsou stochasticky nezávislé náhodné veličiny $X_1, \dots, X_n$.

Vysvětlení: Jsou-li náhodné veličiny $X_1, \dots, X_n$ stochasticky nezávislé, pak to znamená, že informace o realizaci jedné náhodné veličiny nijak neovlivní šance, s nimiž očekáváme realizace ostatních náhodných veličin. Stochastická nezávislost náhodných veličin je zidealizovaným protějškem četnostní nezávislosti znaků v daném výběrovém souboru — viz definice 2.7 a 2.17.

7.16. Příklad
Na výrobcích měříme délku s přesností $\pm 0{,}5$ mm a šířku s přesností $\pm 0{,}2$ mm. Náhodná veličina X_1 udává chybu při měření délky a náhodná veličina X_2 udává chybu při měření šířky. Předpokládáme, že simultánní hustota pravděpodobnosti $\varphi(x_1, x_2)$ je uvnitř mezí chyb konstantní, tj.

$$\varphi(x_1, x_2) = \begin{cases} k & \text{pro } -0{,}5 < x_1 < 0{,}5; -0{,}2 < x_2 < 0{,}2, \\ 0 & \text{jinak.} \end{cases}$$

Určete konstantu k, najděte marginální hustoty pravděpodobnosti $\varphi_1(x_1)$, $\varphi_2(x_2)$, simultánní distribuční funkci $\Phi(x_1, x_2)$, obě marginální distribuční funkce $\Phi_1(x_1)$, $\Phi_2(x_2)$, vypočítejte pravděpodobnost $P((-0{,}1 < X_1 < 0{,}1) \wedge (-0{,}1 < X_2 < 0{,}1))$ a zjistěte, zda náhodné veličiny X_1, X_2 jsou stochasticky nezávislé.

Řešení:
Z normovanosti simultánní hustoty pravděpodobnosti plyne:

$$1 = \int\limits_{-0{,}5}^{0{,}5} \int\limits_{-0{,}2}^{0{,}2} k\, dx_1 dx_2 = k[x_1]_{-0{,}5}^{0{,}5} [x_2]_{-0{,}2}^{0{,}2} = k \cdot 1 \cdot 0{,}4 \quad \Rightarrow \quad k = 2{,}5.$$

Marginální hustoty pravděpodobnosti pomocí věty 7.12 (b):

$$\varphi_1(x_1) = \int\limits_{-0{,}2}^{0{,}2} 2{,}5\, dx_2 = 2{,}5[x_2]_{-0{,}2}^{0{,}2} = 1 \text{ pro } -0{,}5 < x_1 < 0{,}5,$$
$$\varphi_1(x_1) = 0 \text{ jinak.}$$

Podobně

$$\varphi_2(x_2) = \int\limits_{-0{,}5}^{0{,}5} 2{,}5\, dx_1 = 2{,}5[x_1]_{-0{,}5}^{0{,}5} = 2{,}5 \text{ pro } -0{,}2 < x_2 < 0{,}2,$$
$$\varphi_2(x_2) = 0 \text{ jinak.}$$

Z definice 7.11 (vektorový případ) plyne:

$$\Phi(x_1, x_2) = \int\limits_{-0{,}5}^{x_1} \int\limits_{-0{,}2}^{x_2} 2{,}5\, dt_1 dt_2 = 2{,}5[t_1]_{-0{,}5}^{x_1} [t_2]_{-0{,}2}^{x_2} = 2{,}5(x_1 + 0{,}5)(x_2 + 0{,}2)$$

pro $-0{,}5 < x_1 < 0{,}5, -0{,}2 < x_2 < 0{,}2$, $\Phi(x_1, x_2) = 0$ pro $x_1 < -0{,}5$ nebo $x_2 < -0{,}2$, $\Phi(x_1, x_2) = 1$ pro $x_1 > 0{,}5$ a $x_2 > 0{,}2$. Z definice 7.11 (skalární případ) dostaneme:

$$\Phi_1(x_1) = \int\limits_{-0{,}5}^{x_1} 1\, dt_1 = [t_1]_{-0{,}5}^{x_1} = x_1 + 0{,}5$$

pro $-0{,}5 < x_1 < 0{,}5$, $\Phi_1(x_1) = 1$ pro $x_1 \geq 0{,}5$, $\Phi_1(x_1) = 0$ pro $x_1 \leq -0{,}5$. Dále

$$\Phi_2(x_2) = \int_{-0{,}2}^{x_2} 1\, dt_2 = [t_2]_{-0{,}2}^{x_2} = 2{,}5(x_2 + 0{,}2)$$

pro $-0{,}2 < x_2 < 0{,}2$, $\Phi_2(x_2) = 1$ pro $x_2 \geq 0{,}2$, $\Phi_2(x_2) = 0$ pro $x_2 \leq -0{,}2$. Stochastickou nezávislost náhodných veličin X_1, X_2 ověříme pomocí definice 7.15 (c): $\forall (x_1, x_2) \in \mathbb{R}^2 : \varphi(x_1, x_2) = \varphi_1(x_1)\varphi_2(x_2)$, tedy náhodné veličiny X_1, X_2 jsou stochasticky nezávislé.

7.17. Příklad

Diskrétní náhodný vektor (X_1, X_2) má simultánní pravděpodobnostní funkci $\pi(x_1, x_2)$ danou hodnotami: $\pi(-1, 2) = \pi(-1, 3) = \pi(0, 3) = \pi(1, 0) = \pi(1, 1) = 0$, $\pi(-1, 0) = \pi(0, 1) = \pi(1, 2) = 2c$, $\pi(-1, 1) = \pi(0, 0) = \pi(0, 2) = \pi(1, 3) = c$. Určete konstantu c, hodnotu simultánní distribuční funkce $\Phi(0, 2)$, obě marginální pravděpodobnostní funkce $\pi_1(x_1)$, $\pi_2(x_2)$ a hodnotu marginální distribuční funkce $\Phi_1(1)$. Zjistěte, zda náhodné veličiny X_1, X_2 jsou stochasticky nezávislé.

Řešení:

Hodnoty simultánní pravděpodobnostní funkce $\pi(x_1, x_2)$ uspořádáme do kontingenční tabulky, kterou ještě doplníme o sloupec s hodnotami $\pi_1(x_1)$ a řádek s hodnotami $\pi_2(x_2)$. Tyto hodnoty získáme pomocí věty 7.8 (vektorový případ).

		x_2				$\pi_1(x_1)$
		0	1	2	3	
x_1	-1	$2c$	c	0	0	$3c$
	0	c	$2c$	c	0	$4c$
	1	0	0	$2c$	c	$3c$
$\pi_2(x_2)$		$3c$	$3c$	$3c$	c	1

Z normovanosti pravděpodobnostní funkce diskrétního náhodného vektoru (viz věta 7.8, vektorový případ) dostáváme $10c = 1$, tedy $c = 0{,}1$. Z definice diskrétního náhodného vektoru (definice 7.7, vektorový případ) plyne

$$\begin{aligned}\Phi(0, 2) = \pi(-1, 0) + \pi(-1, 1) + \pi(-1, 2) + \pi(-1, 3) + \pi(0, 0) + \\ + \pi(0, 1) + \pi(0, 2) = 0{,}2 + 0{,}1 + 0 + 0{,}1 + 0{,}2 + 0{,}1 = 0{,}6.\end{aligned}$$

Z definice diskrétní náhodné veličiny (definice 7.7, skalární případ) plyne

$$\Phi_1(1) = \pi_1(-1) + \pi_1(0) + \pi_1(1) = 0{,}3 + 0{,}4 + 0{,}3 = 1.$$

Pokud by náhodné veličiny X_1, X_2 byly stochasticky nezávislé, musel by pro všechna $\forall (x_1, x_2) \in \mathbb{R}^2$ platit multiplikativní vztah: $\pi(x_1, x_2) = \pi_1(x_1)\pi_2(x_2)$ (viz definice 7.15 (b)). Avšak již pro $x_1 = -1$, $x_2 = 0$ dostáváme $\pi(-1, 0) = 0{,}2$, $\pi_1(-1) = 0{,}3$, $\pi_2(0) = 0{,}3$. Vidíme tedy, že multiplikativní vztah splněn není a náhodné veličiny X_1, X_2 nejsou stochasticky nezávislé.

7.18. Věta
Jsou-li náhodné veličiny $X_1, \dots, X_n$ stochasticky nezávislé, pak jsou stochasticky nezávislé také transformované náhodné veličiny $Y_1 = g_1(X_1), \dots, Y_n = g_n(X_n)$.

Shrnutí kapitoly

Náhodná veličina se zavádí jako zobrazení, které každému výsledku náhodného pokusu přiřazuje číslo (pak se jedná o **skalární náhodnou veličinu**) nebo více čísel (v tomto případě jde o **náhodný vektor**). Náhodnou veličinu lze pomocí libovolné funkce transformovat a získat tak **transformovanou náhodnou veličinu**. Pravděpodobnostní chování náhodné veličiny popisuje **distribuční funkce**, jejíž zavedení je motivováno empirickou distribuční funkcí známou z popisné statistiky. Vlastnosti těchto dvou funkcí jsou analogické.

Praktický význam mají dva speciální druhy náhodných veličin. **Diskrétní náhodná veličina** může nabývat pouze spočetně mnoha hodnot a její pravděpodobnostní chování je popsáno **pravděpodobnostní funkcí**, což je „zidealizovaný“ protějšek četnostní funkce. **Diskrétní náhodný vektor** je tvořen diskrétními náhodnými veličinami. Zabývali jsme se náhodnými vektory se dvěma složkami. V souvislosti s diskrétním náhodným vektorem zavádíme **simultánní pravděpodobnostní funkci**. **Marginální pravděpodobnostní funkce** se vztahují k jednotlivým složkám náhodného vektoru.

Spojitá náhodná veličina nabývá všech hodnot z nějakého intervalu. Její pravděpodobnostní chování je popsáno **hustotou pravděpodobnosti**, což je „zidealizovaný“ protějšek hustoty četnosti. **Spojitý náhodný vektor** je tvořen spojitými náhodnými veličinami. Jeho pravděpodobnostní chování je popsáno **simultánní hustotou pravděpodobnosti. Marginální hustoty pravděpodobnosti** se vztahují k jednotlivým složkám náhodného vektoru.

Pomocí multiplikativního vztahu, v němž vystupují simultánní a marginální distribuční funkce (resp. pravděpodobnostní funkce v diskrétním případě resp. hustoty pravděpodobnosti ve spojitém případě), zavedeme pojem **stochastické nezávislosti náhodných veličin.**

Kontrolní otázky a úkoly

1 Uveďte příklad náhodné veličiny a náhodného vektoru z ekonomické praxe.

2 Najděte distribuční funkci náhodné veličiny, která udává počet líců při hodu třemi mince-mi a nakreslete její graf.

3 Rozhodněte, které z uvedených náhodných veličin jsou diskrétní a které jsou spojité:
- a) počet členů domácnosti
- b) věk člověka v letech
- c) náhodně vybrané reálné číslo
- d) počet zákazníků ve frontě

e) cena výrobku
f) počet zmetků z celkové denní produkce
g) délka určitého předmětu
h) životnost televizoru v letech

4 Které funkcionální charakteristiky popisují pravděpodobnostní chování diskrétní náhodné veličiny a které diskrétního náhodného vektoru?

5 Které funkcionální charakteristiky popisují pravděpodobnostní chování spojité náhodné veličiny a které spojitého náhodného vektoru?

6 Je-li X diskrétní náhodná veličina s pravděpodobnostní funkcí $\pi(x)$, může být $\pi(x) > 1$?

7 Je-li X spojitá náhodná veličina s hustotou pravděpodobnosti $\varphi(x)$, může být $\varphi(x) > 1$?

8 Náhodná veličina udává průměrný počet ok při hodu dvěma kostkami. Nakreslete graf její pravděpodobnostní funkce.

9 Diskrétní náhodný vektor (X_1, X_2) má simultánní pravděpodobnostní funkci $\pi(x_1, x_2)$ danou hodnotami:

$$\begin{aligned}&\pi(0,0) = \pi(0,2) = \pi(1,1) = \pi(2,0) = \pi(2,2) = 0,\\ &\pi(0,1) = \pi(1,2) = \pi(2,1) = 0{,}25.\end{aligned}$$

Jsou náhodné veličiny X_1, X_2 stochasticky nezávislé?

10 Nechť spojitý vektor (X_1, X_2) má simultánní hustotu pravděpodobnosti

$$\varphi(x_1, x_2) = \begin{cases} 24x_1^2 x_2(1 - x_1) & \text{pro } 0 \le x_1 < 1, 0 \le x_2 < 1,\\ 0 & \text{jinak.}\end{cases}$$

Dokažte, že náhodné veličiny X_1, X_2 jsou stochasticky nezávislé.

8. Vybraná rozložení diskrétních a spojitých náhodných veličin

Cíl kapitoly

Po prostudování této kapitoly budete umět:

- rozlišovat důležité typy diskrétních a spojitých rozložení
- využívat vlastností těchto rozložení při výpočtu pravděpodobností různých jevů
- hledat v tabulkách hodnot distribuční funkce standardizovaného normálního rozložení

Časová zátěž

Na prostudování této kapitoly budete potřebovat asi 5 hodin studia.

Nyní se seznámíme s přehledem důležitých pravděpodobnostních funkcí a hustot pravděpodobnosti. Uvedeme nejenom analytické vyjádření těchto funkcí, ale též grafy. Vysvětlíme rovněž, v jakých situacích se lze s uvedenými rozloženími pravděpodobnosti setkat. Zvláštním pozornost budeme věnovat normálnímu rozložení, které hraje velkou roli v celé řadě praktických aplikací počtu pravděpodobnosti a, jak uvidíme později, i v matematické statistice.

8.1. Označení

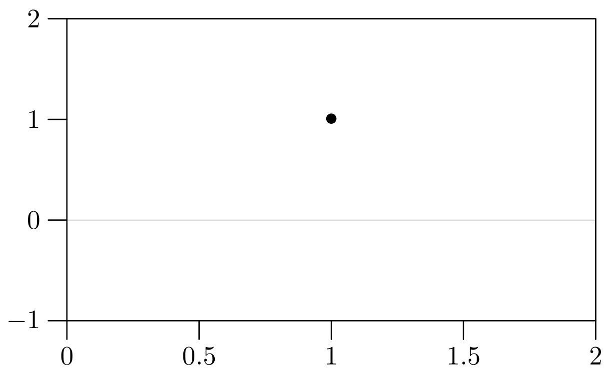
Známe-li distribuční funkci $\Phi(x)$ náhodné veličiny X (resp. pravděpodobnostní funkci $\pi(x)$ v diskrétním případě resp. hustotu pravděpodobnosti $\varphi(x)$ ve spojitém případě), pak řekneme, že známe rozložení pravděpodobností (zkráceně rozložení) náhodné veličiny X. Toto rozložení závisí na nějakém parametru ν, což nejčastěji bývá reálné číslo nebo reálný vektor. Zápis $X \sim L(\nu)$ čteme: náhodná veličina X má rozložení L s parametrem ν.

8.2. Definice

Nejprve se seznámíme s vybranými rozloženími diskrétních náhodných veličin.

a) *Degenerované rozložení*: $X \sim Dg(\mu)$
Tato náhodná veličina nabývá pouze konstantní hodnotu μ.

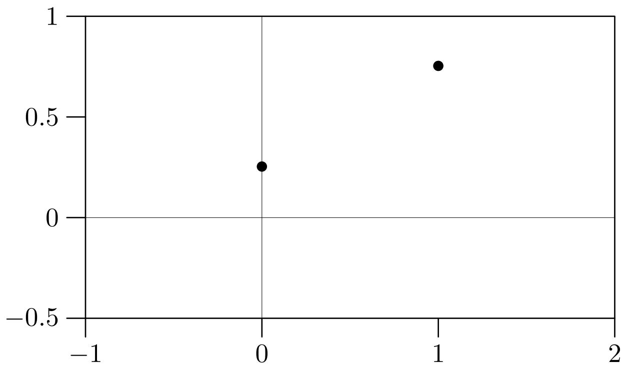
$$\pi(x) = \begin{cases} 1 & \text{pro } x = \mu, \\ 0 & \text{jinak.} \end{cases}$$



Pravděpodobnostní funkce $Dg(1)$.

b) *Alternativní rozložení*: $X \sim A(\nu)$
Náhodná veličina X udává počet úspěchů v jednom pokusu, přičemž pravděpodobnost úspěchu je ν.

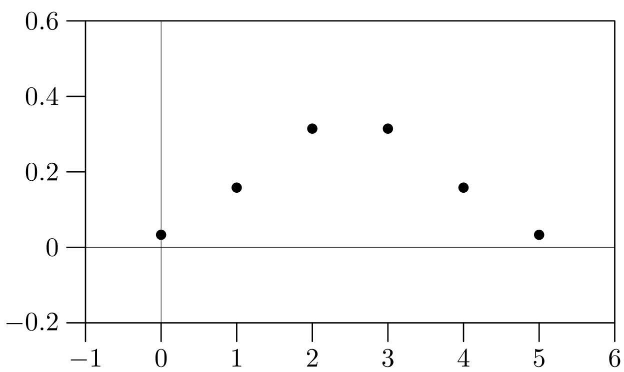
$$\pi(x) = \begin{cases} 1-\nu & \text{pro } x = 0, \\ \nu & \text{pro } x = 1, \\ 0 & \text{jinak.} \end{cases}$$



Pravděpodobnostní funkce $A(0{,}75)$.

c) *Binomické rozložení*: $X \sim Bi(n,\nu)$
Náhodná veličina X udává počet úspěchů v posloupnosti n nezávislých opakovaných pokusů, přičemž pravděpodobnost úspěchu je v každém pokusu ν.

$$\pi(x) = \begin{cases} \binom{n}{x}\nu^x(1-\nu)^{n-x} & \text{pro } x = 0,1,\ldots,n \\ 0 & \text{jinak.} \end{cases}$$



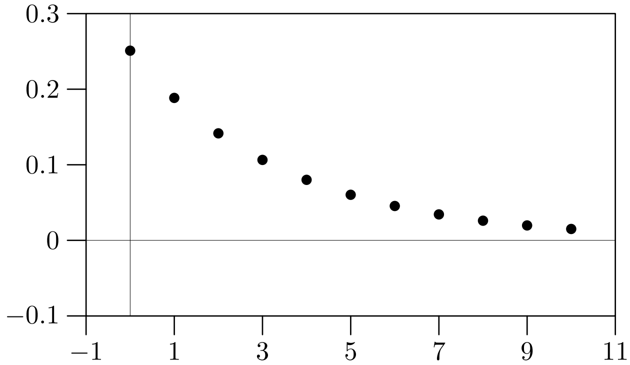
Pravděpodobnostní funkce $Bi(5; 0{,}5)$.

(Odvození – viz př. 6.3 (b).) Alternativní rozložení je speciálním případem binomického rozložení pro $n = 1$. Jsou-li $X_1,\ldots,X_n$ stochasticky nezávislé náhodné veličiny, $X_i \sim A(\nu)$, $i = 1,\ldots,n$, pak

$$X = \sum_{i=1}^{n} X_i \sim Bi(n,\nu).$$

d) *Geometrické rozložení*: $X \sim Ge(\nu)$
Náhodná veličina X udává počet neúspěchů v posloupnosti opakovaných nezávislých pokusů předcházejících prvnímu úspěchu, přičemž pravděpodobnost úspěchu je v každém pokusu ν.

$$\pi(x) = \begin{cases} (1-\nu)^x \nu & \text{pro } x = 0, 1, \dots \\ 0 & \text{jinak.} \end{cases}$$

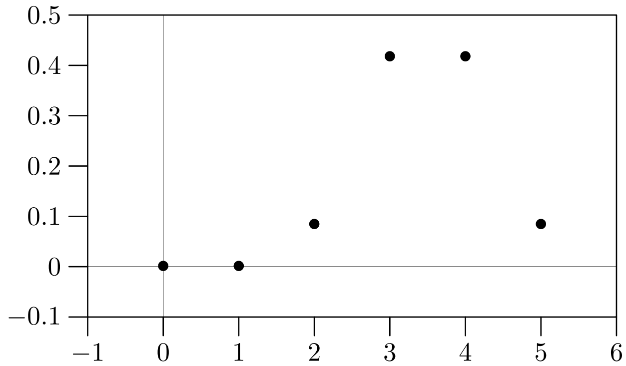


Pravděpodobnostní funkce $Ge(0{,}25)$.

(Odvození – viz př. 6.3 (a).)

e) *Hypergeometrické rozložení*: $X \sim Hg(N, M, n)$
V souboru N prvků je M prvků označeno. Náhodně vybereme n prvků bez vracení. Náhodná veličina X udává počet vybraných označených prvků.

$$\pi(x) = \begin{cases} \dfrac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} & \text{pro } x = \max\{0, M-N+n\}, \dots \min\{M, n\}, \\ 0 & \text{jinak.} \end{cases}$$

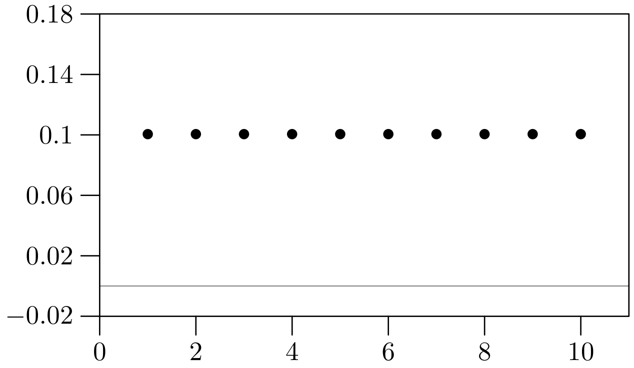


Pravděpodobnostní funkce $Hg(10, 7, 5)$.

f) *Rovnoměrné diskrétní rozložení*: $X \sim Rd(G)$
Nechť G je konečná množina o n prvcích. Náhodná veličina X nabývá se stejnou pravděpodobností každé hodnoty z množiny G.

$$\pi(x) = \begin{cases} \frac{1}{n} & \text{pro } x \in G, \\ 0 & \text{jinak.} \end{cases}$$

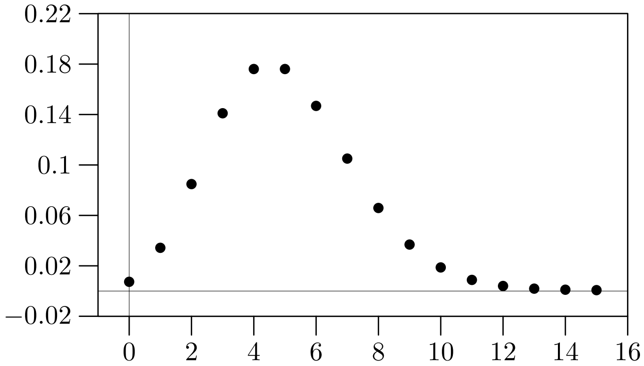
(Typickým příkladem je náhodná veličina udávající počet ok při hodu kostkou.)



Pravděpodobnostní funkce $Rd(\{1, 2, \ldots, 10\})$.

g) *Poissonovo rozložení*: $X \sim Po(\lambda)$
Náhodná veličina X udává počet událostí, které nastanou v jednotkovém časovém intervalu, přičemž události nastávají náhodně, jednotlivě a vzájemně nezávisle. Parametr $\lambda > 0$ je střední počet těchto událostí.

$$\pi(x) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda} & \text{pro } x = 0, 1, \ldots, \\ 0 & \text{jinak.} \end{cases}$$



Pravděpodobnostní funkce $Po(5)$.

8.3. Příklad
V rodině je 10 dětí. Za předpokladu, že chlapci i dívky se rodí s pravděpodobností 0,5 a pohlaví se formuje nezávisle na sobě, určete pravděpodobnost, že v této rodině jsou nejméně 3 a nejvýše 8 chlapců.

Řešení:
X – počet chlapců v této rodině, $X \sim Bi(10; 0{,}5)$,

$$P(3 \leq X \leq 8) = \sum_{x=3}^{8} \binom{10}{x} \left(\frac{1}{2}\right)^x \left(1 - \frac{1}{2}\right)^{10-x} = \frac{957}{1024} = 0{,}935.$$

8.4. Příklad
Jaká je pravděpodobnost, že při hře „Člověče, nezlob se!“ nasadíme nejpozději při třetím hodu?

Řešení:
X – počet neúspěchů před první šestkou, $X \sim Ge(\frac{1}{6})$,

$$P(X \leq 2) = \sum_{x=0}^{2} \left(1 - \frac{1}{6}\right)^x \frac{1}{6} = 0{,}4213.$$

8.5. Příklad
Při provozu balicího automatu vznikají během směny náhodné poruchy, které se řídí rozložením $Po(2)$. Jaká je pravděpodobnost, že během směny dojde aspoň k jedné poruše?

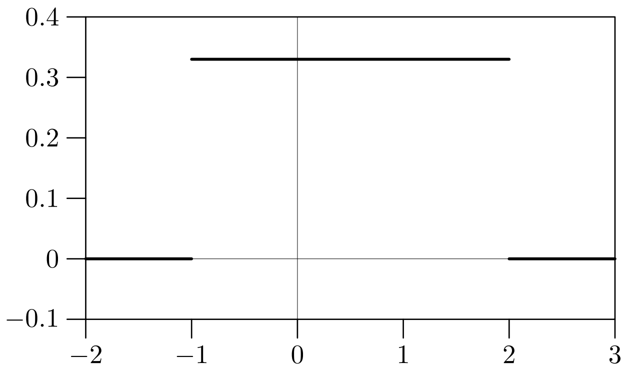
Řešení:
X – počet poruch během směny, $X \sim Po(2)$,

$$P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) = 1 - \frac{2^0}{0!}e^{-2} = 0{,}8647.$$

8.6. Definice
Nyní uvedeme vybrané typy spojitých rozložení.

a) *Rovnoměrné spojité rozložení*: $X \sim Rs(a, b)$
Náhodná veličina X nabývá se stejnou pravděpodobností každé hodnoty z intervalu (a, b).

$$\varphi(x) = \begin{cases} \frac{1}{b-a} & \text{pro } x \in (a, b), \\ 0 & \text{jinak.} \end{cases}$$

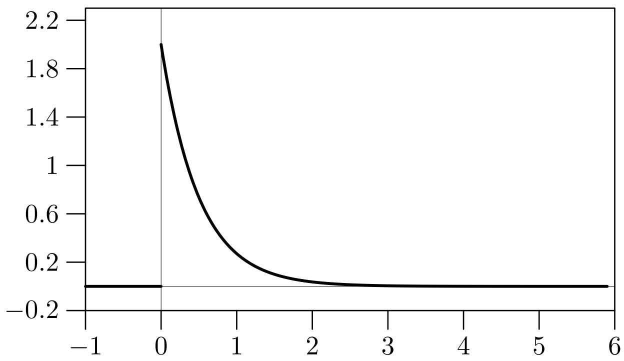


Hustota $Rs(-1, 2)$.

b) *Exponenciální rozložení*: $X \sim Ex(\lambda)$
Náhodná veličina X udává dobu čekání na příchod nějaké události,

která se může dostavit každým okamžikem se stejnou šancí bez ohledu na dosud pročekanou dobu. Přitom $\frac{1}{\lambda}$ vyjadřuje střední dobu čekání.

$$\varphi(x) = \begin{cases} \lambda e^{-\lambda x} & \text{pro } x > 0, \\ 0 & \text{jinak.} \end{cases}$$



Hustota $Ex(2)$.

c) *Normální rozložení*: $X \sim N(\mu, \sigma^2)$
Tato náhodná veličina vzniká např. tak, že ke konstantě μ se přičítá velké množství nezávislých náhodných vlivů mírně kolísajících kolem 0. Proměnlivost těchto vlivů je vyjádřena konstantou $\sigma > 0$.

$$\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

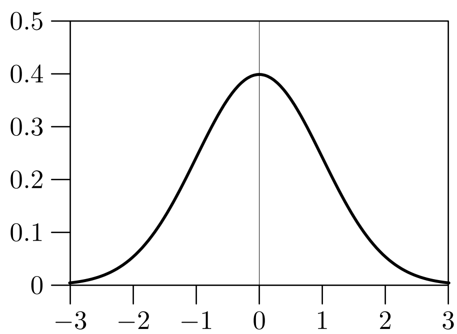
Pro $\mu = 0$, $\sigma^2 = 1$ se jedná o standardizované normální rozložení, píšeme $U \sim N(0,1)$. Hustota pravděpodobnosti má v tomto případě tvar

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}.$$

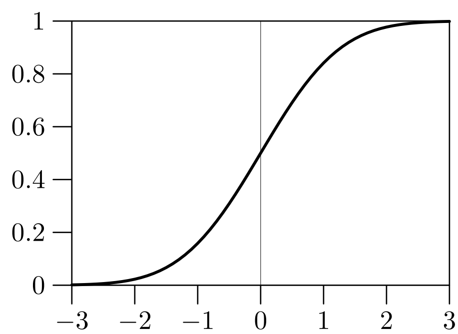
Distribuční funkce standardizovaného normálního rozložení

$$\Phi(u) = \int\limits_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt$$

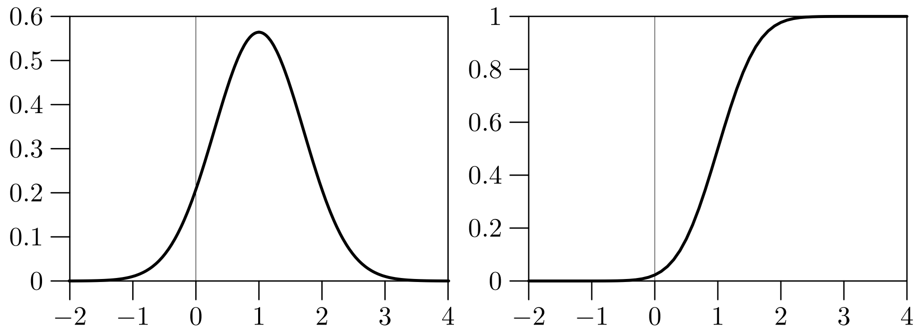
je tabelována pro $u \geq 0$, pro $u < 0$ se používá přepočtový vzorec $\Phi(-u) = 1 - \Phi(u)$. Má-li $X \sim N(\mu, \sigma^2)$, pak $U = \frac{X-\mu}{\sigma} \sim N(0,1)$.



Hustota $N(0,1)$



Distribuční funkce $N(0,1)$



Hustota $N(1; 0{,}5)$ Distribuční funkce $N(1; 0{,}5)$

(Normální rozložení hraje ústřední roli v počtu pravděpodobnosti i matematické statistice. Jeho význam spočívá jednak v tom, že normálním rozložením se řídí pravděpodobnostní chování mnoha náhodných veličin a jednak v tom, že za určitých podmínek konverguje k normálnímu rozložení součet nezávislých náhodných veličin s týmž rozložením.)

d) *Dvourozměrné normální rozložení*:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$$

Náhodný vektor $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ vzniká ve dvourozměrných situacích podobně jako skalární náhodná veličina v bodě (e).

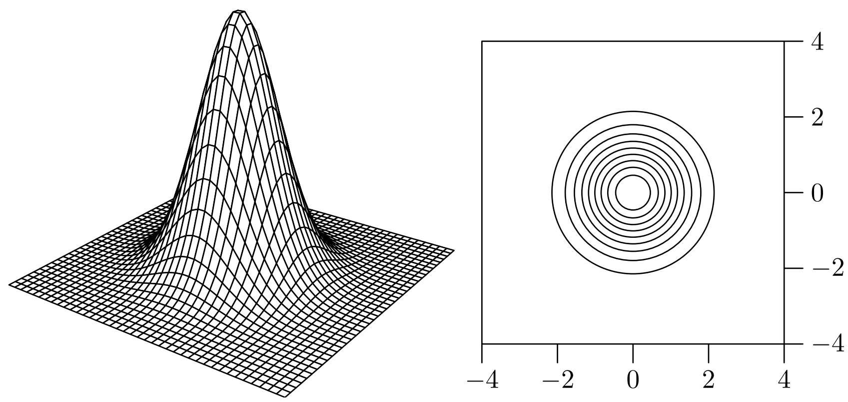
$$\varphi(x_1, x_2) = \frac{1}{\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{q(x_1,x_2)}{2}},$$

kde

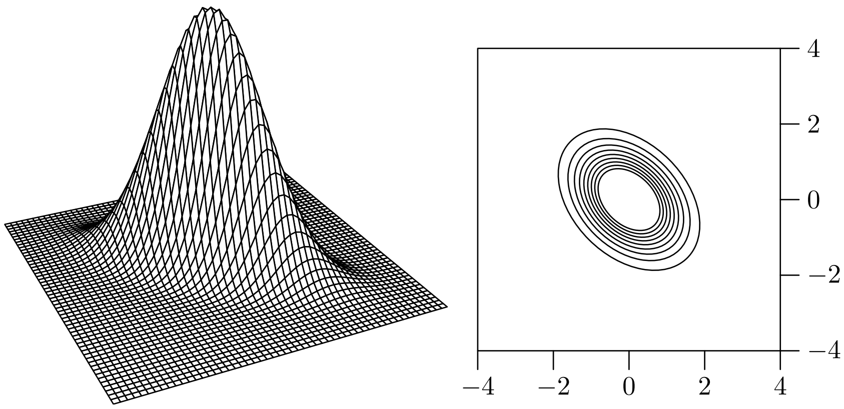
$$q(x_1, x_2) = \frac{1}{1-\rho^2}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{x_1-\mu_1}{\sigma_1}\frac{x_2-\mu_2}{\sigma_2} + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right].$$

Pro $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1^2 = 1$, $\sigma_2^2 = 1$, $\rho = 0$ se jedná o standardizované dvourozměrné normální rozložení.

Vrstevnice a graf hustoty standardizovaného dvourozměrného normálního rozložení:

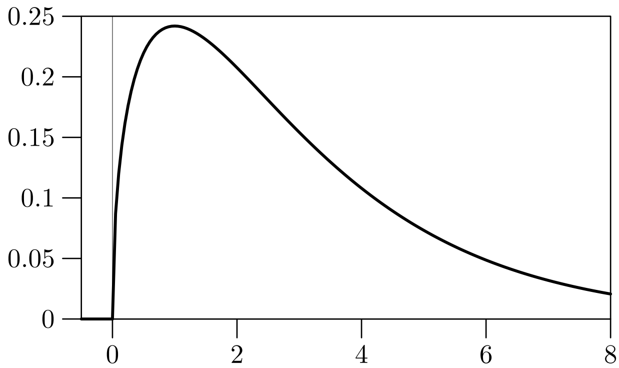


Vrstevnice a graf hustoty dvourozměrného normálního rozložení s parametry $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1^2 = 1$, $\sigma_2^2 = 1$, $\rho = -0{,}75$



Následující tři rozložení – Pearsonovo, Studentovo a Fisherovo-Snedecorovo – jsou odvozena ze standardizovaného normálního rozložení. Mají velký význam především v matematické statistice při konstrukci intervalů spolehlivosti a testování hypotéz. Vyjádření hustot těchto rozložení neuvádíme, je příliš složité – viz např. [3].)

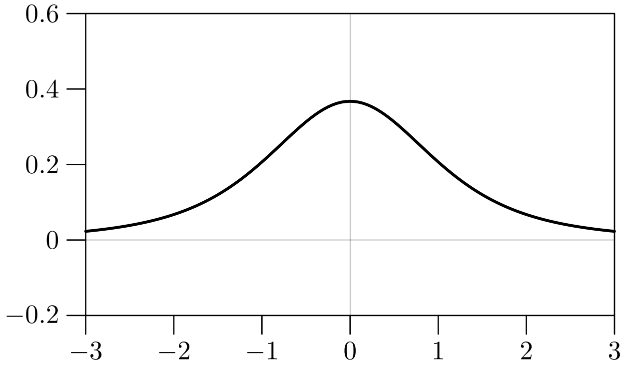
e) *Pearsonovo rozložení chí-kvadrát s n stupni volnosti*: $X \sim \chi^2(n)$
Nechť $X_1, \ldots, X_n$ jsou stochasticky nezávislé náhodné veličiny, $X_i \sim N(0,1)$, $i = 1, \ldots, n$. Pak náhodná veličina $X = X_1^2 + \cdots + X_n^2 \sim \chi^2(n)$.



Hustota $\chi^2(3)$.

f) *Studentovo rozložení s n stupni volnosti*: $X \sim t(n)$
Nechť X_1, X_2 jsou stochasticky nezávislé náhodné veličiny a nechť dále $X_1 \sim N(0,1)$, $X_2 \sim \chi^2(n)$. Pak náhodná veličina

$$X = \frac{X_1}{\sqrt{\frac{X_2}{n}}} \sim t(n).$$



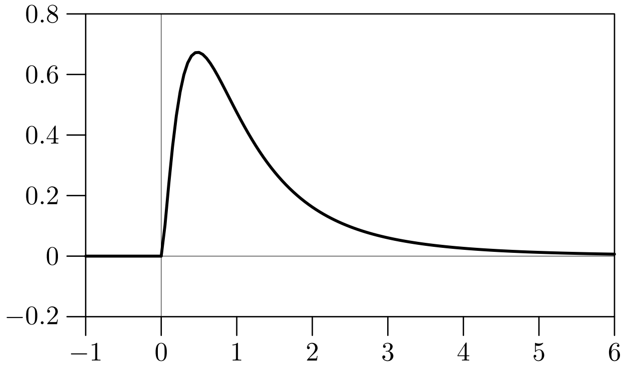
Hustota $t(3)$.

g) *Fisherovo-Snedecorovo rozložení s n_1 a n_2 stupni volnosti*:

$$X \sim F(n_1, n_2)$$

Nechť $X_1, \dots, X_n$ jsou stochasticky nezávislé náhodné veličiny, $X_i \sim \chi^2(n_i)$, $i = 1, 2$. Pak náhodná veličina

$$X = \frac{\frac{X_1}{n_1}}{\frac{X_2}{n_2}} \sim F(n_1, n_2).$$



Hustota $F(5, 8)$.

8.7. Příklad

Na automatické lince se plní láhve mlékem. Působením náhodných vlivů množství mléka kolísá v intervalu (980 ml, 1020 ml). Každé množství mléka v tomto intervalu považujeme za stejně možné. Jaká je pravděpodobnost, že v náhodně vybrané láhvi bude aspoň 1000 ml mléka?

Řešení:

X – množství mléka v náhodně vybrané láhvi, $X \sim Rs(980, 1020)$,

$$\varphi(x) = \begin{cases} \frac{1}{40} & \text{pro } x \in (980, 1020), \\ 0 & \text{jinak.} \end{cases}$$

$$P(X \geq 1000) = \int_{1000}^{1020} \frac{1}{40}\, dx = \frac{1}{40}[x]_{1000}^{1020} = 0{,}5.$$

8.8. Příklad
Doba (v minutách) potřebná k obsloužení zákazníka v prodejně potravin je náhodná veličina, která se řídí rozložením $Ex(\frac{1}{3})$. Jaká je pravděpodobnost, že doba potřebná k obsloužení náhodně vybraného zákazníka v této prodejně bude v rozmezí od 3 do 6 minut?

Řešení:
X – doba potřebná k obsloužení náhodně vybraného zákazníka, $X \sim Ex(\frac{1}{3})$,

$$\varphi(x) = \begin{cases} \frac{1}{3}e^{-\frac{x}{3}} & \text{pro } x > 0, \\ 0 & \text{jinak.} \end{cases}$$

$$P(3 \leq X \leq 6) = \int_3^6 \frac{1}{3}e^{-\frac{x}{3}}\, dx = \frac{1}{3}(-3)\left[e^{-\frac{x}{3}}\right]_3^6 = -e^{-2} + e^{-1} = 0{,}233.$$

8.9. Příklad
Výsledky u přijímacích zkoušek na jistou VŠ jsou normálně rozloženy s parametry $\mu = 550$ bodů, $\sigma = 100$ bodů. S jakou pravděpodobností bude mít náhodně vybraný uchazeč aspoň 600 bodů?

Řešení:
X – výsledek náhodně vybraného uchazeče, $X \sim N(550, 100^2)$,

$$\begin{aligned} P(X \geq 600) &= 1 - P(X \leq 600) + P(X = 600) = 1 - P(X \leq 600) = \\ &= 1 - P\left(\frac{X-\mu}{\sigma} \leq \frac{600-\mu}{\sigma}\right) = 1 - P\left(U \leq \frac{600-550}{100}\right) = \\ &= 1 - \Phi(0{,}5) = 1 - 0{,}69146 = 0{,}31. \end{aligned}$$

8.10. Příklad
Nechť X_1, X_2, X_3, X_4 jsou stochasticky nezávislé náhodné veličiny, $X_i \sim N(0,1)$, $i = 1,2,3,4$. Jaké rozložení má transformovaná náhodná veličina

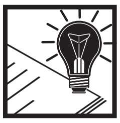
$$X = \frac{X\sqrt{3}}{\sqrt{X_2^2 + X_3^2 + X_4^2}}?$$

Řešení:
$X \sim t(3)$, protože $X_1 \sim N(0,1)$ a $X_2^2 + X_3^2 + X_4^2 \sim \chi^2(3)$.

Shrnutí kapitoly

Degenerované rozložení popisuje pravděpodobnostní chování konstanty, což je nepochybně patologický případ. Zajímavější je **alternativní**, **geometrické** a zvláště **binomické rozložení**. Všechna tato rozložení souvisejí s počty úspěchů či neúspěchů v posloupnosti opakovaných nezávislých pokusů. **Hypergeometrické rozložení** se vyskytuje v situacích, kdy provádíme výběr bez vracení ze souboru, který obsahuje označené prvky. **Rovnoměrné rozložení** na dané množině je charakteristické tím, že náhodná veličina, která se jím řídí, nabývá každé hodnoty z této množiny se stejnou pravděpodobností. Podle **Poissonova rozložení** se chová např. náhodná veličina udávající počet událostí, které nastanou v jednotkovém čase.

Za spojitých rozložení je nejjednodušší **rovnoměrné spojité rozložení**. Jeho hustota je na daném intervalu konstantní a jinde nulová. Náhodná veličina s **exponenciálním rozložením** udává dobu čekání na příchod nějaké události, přičemž toto čekání probíhá „bez paměti“. Vůbec nejdůležitějším rozložením je **normální rozložení**, které vzniká např. tak, že k nějaké konstantě se přičítá velké množství nezávislých náhodných vlivů mírně kolísajících kolem nuly. Tím se z konstanty stane náhodná veličina. Grafem normální hustoty pravděpodobnosti je známá Gaussova křivka. Pomocí standardizovaného rozložení lze zavést další tři typy speciálních rozložení, a to **Pearsonovo**, **Studentovo** a **Fisherovo-Snedecorovo**. Nacházejí uplatnění především v matematické statistice.



Kontrolní otázky a úkoly

1. (S) Pomocí systému STATISTICA nakreslete grafy hustot a distribučních funkcí uvedených spojitých rozložení. Sledujte vliv parametrů na tvar hustot a distribučních funkcí. Návod: viz příloha B.
2. (S) Pojišťovna zjistila, že 12% pojistných událostí je způsobeno vloupáním. Jaká je pravděpodobnost, že mezi 30 náhodně vybranými pojistnými událostmi bude způsobeno vloupáním nejvýše 6?
3. Doba (v hodinách), která uplyne mezi dvěma naléhavými příjmy v jisté nemocnici, se řídí rozložením $Ex(0{,}5)$. Jaká je pravděpodobnost, že uplyne více než 5 hodin bez naléhavého příjmu?
4. Jaká je pravděpodobnost, že náhodná veličina $X \sim N(20, 16)$ nabude hodnotu menši než 12 nebo větší než 28?
5. Nechť $X \sim Rs(a, b)$, přičemž

   $$\Phi(x) = \begin{cases} 0 & \text{pro } x \le a \\ \frac{x+20}{55} & \text{pro } a < x < b \\ 1 & \text{pro } x \ge b \end{cases}$$

   Určete a, b.
6. Nechť X_1, X_2 jsou stochasticky nezávislé náhodné veličiny takové, že $X_i \sim N(0, 1)$, $i = 1, 2$. Jaké rozložení má transformovaná náhodná veličina

   $$X = \frac{X_1^2}{X_2^2}?$$

9. Číselné charakteristiky náhodných veličin

Cíl kapitoly

Po prostudování této kapitoly budete umět:

- spočítat kvantily spojitých náhodných veličin
- hledat kvantily některých spojitých náhodných veličin ve statistických tabulkách
- určit střední hodnotu a rozptyl náhodné veličiny
- spočítat kovarianci a koeficient korelace dvou náhodných veličin
- využívat vlastností číselných charakteristik náhodných veličin při konkrétních výpočtech

Časová zátěž

Na prostudování této kapitoly budete potřebovat asi 10 hodin studia.

9.1. Motivace

V 7. kapitole jsme se seznámili s funkcionálními charakteristikami náhodných veličin (např. distribuční funkce, pravděpodobnostní funkce, hustota pravděpodobnosti), které plně popisují pravděpodobnostní chování náhodné veličiny. Číselné charakteristiky vystihují pouze některé rysy tohoto chování, např. popisují polohu realizací náhodné veličiny na číselné ose či jejich proměnlivost (variabilitu). Jsou jednodušší než číselné charakteristiky, ale nesou jen částečnou informaci.

9.2. Definice

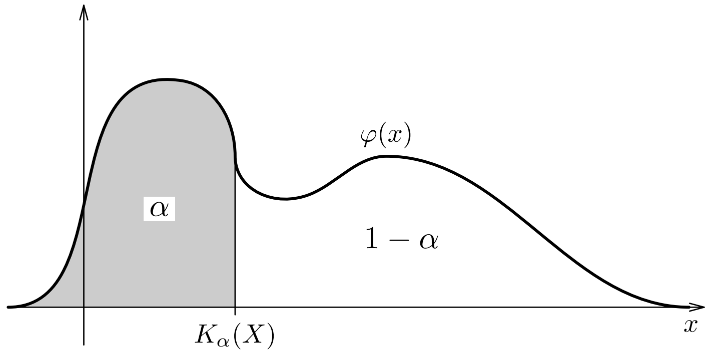
Nechť X je spojitá náhodná veličina aspoň ordinálního charakteru (viz definici 3.2) s distribuční funkcí $\Phi(x)$ a nechť $\alpha \in (0, 1)$. Číslo $K_\alpha(X)$, které splňuje podmínku

$$\alpha = \Phi(K_\alpha(X)) = \int\limits_{-\infty}^{K_\alpha(X)} \varphi(x)\, dx,$$

se nazývá α-kvantil náhodné veličiny X. Kvantil $K_{0,50}(X)$ se nazývá medián, $K_{0,25}(X)$ dolní kvartil, $K_{0,75}(X)$ horní kvartil, $K_{0,10}(X), \ldots, K_{0,90}(X)$ jsou decily, $K_{0,01}(X), \ldots, K_{0,99}(X)$ jsou percentily. Kterýkoliv α-kvantil je charakteristikou polohy číselných realizací náhodné veličiny na číselné ose. Jako charakteristika variability slouží kvartilová odchylka $q = K_{0,75}(X) - K_{0,25}(X)$.

(Lze samozřejmě definovat i kvantily diskrétních náhodných veličin, ale zde se zabýváme jenom kvantily spojitých náhodných veličin, které se v praxi nejčastěji používají.)

Význam α-kvantilu spojité náhodné veličiny ilustruje následující obrázek.



9.3. Označení

$X \sim N(0,1) \Rightarrow K_\alpha(X) = u_\alpha, \quad X \sim \chi^2(n) \Rightarrow K_\alpha(X) = \chi^2_\alpha(n),$
$X \sim t(n) \Rightarrow K_\alpha(X) = t_\alpha(n), \quad X \sim F(n_1, n_2) \Rightarrow K_\alpha(X) = F_\alpha(n_1, n_2).$

Tyto kvantily najdeme ve statistických tabulkách. Používáme vztahy:

$$\begin{aligned} u_\alpha &= -u_{1-\alpha}, \\ t_\alpha(n) &= -t_{1-\alpha}(n), \\ F_\alpha(n_1, n_2) &= \frac{1}{F_{1-\alpha}(n_2, n_1)}. \end{aligned}$$

9.4. Příklad

a) Nechť $U \sim N(0,1)$. Najděte medián a horní a dolní kvartil.
b) Určete $\chi^2_{0,025}(25)$.
c) Určete $t_{0,99}(30)$ a $t_{0,05}(24)$.
d) Určete $F_{0,975}(5,20)$ a $F_{0,05}(2,10)$.

Řešení:

ad a) $u_{0,50} = 0$, $u_{0,25} = -0{,}67449$, $u_{0,75} = 0{,}67449$
ad b) $\chi^2_{0,025}(25) = 13{,}12$
ad c) $t_{0,99}(30) = 2{,}4573$, $t_{0,05}(24) = -1{,}7109$
ad d) $F_{0,975}(5,20) = 3{,}2891$, $F_{0,05}(2,10) = 0{,}05156$

9.5. Věta

Nechť X je spojitá náhodná veličina, $Y = g(X)$ transformovaná náhodná veličina, $\alpha \in (0,1)$.

a) Je-li g všude rostoucí funkce, pak $K_\alpha(Y) = g(K_\alpha(X))$.
b) Je-li g všude klesající funkce, pak $K_\alpha(Y) = g(K_{1-\alpha}(X))$.

9.6. Příklad

Nechť $U \sim N(0,1)$. Najděte devátý decil transformované náhodné veličiny $Y = 3 + 2U$.

Řešení:

Funkce $y = 3 + 2u$ je všude rostoucí funkce, tedy $K_{0,90}(Y) = 3 + 2u_{0,90} = 3 + 2 \cdot 1{,}28155 = 5{,}5631$.

Nyní budeme věnovat pozornost číselným charakteristikám polohy a variability náhodné veličiny intervalového či poměrového charakteru. Jak uvidíme, teoretickým protějškem aritmetického průměru m je střední hodnota $E(X)$ a empirického rozptylu s^2 teoretický rozptyl $D(X)$. Empirický rozptyl s^2 jsme zavedli jako aritmetický průměr kvadrátů centrovaných hodnot. Není tedy překvapivé, že teoretický rozptyl $D(X)$ je střední hodnotou kvadrátů centrovaných hodnot. Naučíme se počítat střední hodnotu a rozptyl transformovaných náhodných veličin a náhodných vektorů. Uvedeme střední hodnoty a rozptyly vybraných typů diskrétních a spojitých rozložení, která jsme poznali v 8. kapitole.

9.7. Definice
Nechť X je náhodná veličina aspoň intervalového charakteru (viz definici 3.2). Její *střední hodnotou* nazýváme číslo $E(X)$, které je v diskrétním případě zavedeno vztahem

$$E(X) = \sum_{x=-\infty}^{\infty} x\pi(x)$$

a ve spojitém případě vztahem

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

za předpokladu, že případná nekonečná suma či integrál vpravo absolutně konverguje. Není-li tato podmínka splněna, pak řekneme, že střední hodnota neexistuje. Transformovaná náhodná veličina $X - E(X)$ se nazývá *centrovaná náhodná veličina*.

(Střední hodnota je číslo, které charakterizuje polohu realizací náhodné veličiny na číselné ose s přihlédnutím k jejich pravděpodobnostem. V diskrétním případě představuje střední hodnota těžiště soustavy hmotných bodů, jejichž hmotnost je popsána pravděpodobnostní funkcí $\pi(x)$ a ve spojitém případě je střední hodnota těžištěm hmotné přímky, na níž je rozprostření hmoty popsáno hustotou pravděpodobnosti $\varphi(x)$. Střední hodnota je teoretickým protějškem váženého aritmetického průměru z definice 3.20.)

9.8. Příklad
Náhodná veličina X udává počet ok při hodu kostkou. Vypočtěte její střední hodnotu.

Řešení:

$$\pi(x) = \begin{cases} \frac{1}{6} & \text{pro } x = 1, 2, \ldots, 6 \\ 0 & \text{jinak,} \end{cases}$$

$$E(X) = \sum_{x=1}^{6} x\pi(x) = \frac{1}{6}(1+2+3+4+5+6) = \frac{7}{2} = 3{,}5.$$

9.9. Věta

a) Skalární případ:

- Nechť X je diskrétní náhodná veličina s pravděpodobnostní funkcí $\pi(x)$ a $Y = g(X)$ je transformovaná náhodná veličina. Pak

$$E(Y) = \sum_{x=-\infty}^{\infty} g(x)\pi(x),$$

pokud suma vpravo absolutně konverguje.

- Nechť X je spojitá náhodná veličina s hustotou pravděpodobnosti $\varphi(x)$ a $Y = g(X)$ je transformovaná náhodná veličina. Pak

$$E(Y) = \int_{-\infty}^{\infty} g(x)\varphi(x)\,dx,$$

pokud integrál vpravo absolutně konverguje.

b) Vektorový případ:

- Nechť (X_1, X_2) je diskrétní náhodný vektor se simultánní pravděpodobnostní funkcí $\pi(x_1, x_2)$ a $Y = g(X_1, X_2)$ je transformovaná náhodná veličina. Pak

$$E(Y) = \sum_{x_1=-\infty}^{\infty}\sum_{x_2=-\infty}^{\infty} g(x_1, x_2)\pi(x_1, x_2),$$

pokud suma vpravo absolutně konverguje.

- Nechť (X_1, X_2) je spojitý náhodný vektor se simultánní hustotou pravděpodobnosti $\varphi(x_1, x_2)$ a $Y = g(X_1, X_2)$ je transformovaná náhodná veličina. Pak

$$E(Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x_1, x_2)\varphi(x_1, x_2)\,dx_1 dx_2,$$

pokud integrál vpravo absolutně konverguje.

9.10. Příklad

Nechť $X \sim Ex(\lambda)$, $Y = e^{-\gamma X}$, kde $\gamma > 0$ je konstanta. Vypočtěte $E(Y)$.

Řešení:

$$\varphi(x) = \begin{cases} \lambda e^{-\lambda x} & \text{pro } x > 0, \\ 0 & \text{jinak}, \end{cases} \qquad E(Y) = \int_0^{\infty} e^{-\gamma x}\lambda e^{-\lambda x}\,dx = \frac{\lambda}{\lambda + \gamma}.$$

9.11. Definice

Rozptylem náhodné veličiny X, která má střední hodnotu $E(X)$, rozumíme číslo $D(X) = E([X - E(X)]^2)$, pokud střední hodnota vpravo existuje. Číslo

$\sqrt{D(X)}$ se nazývá *směrodatná odchylka*. Transformovaná náhodná veličina $\frac{X-E(X)}{\sqrt{D(X)}}$ se nazývá *standardizovaná náhodná veličina.*

Z věty 9.9 (a) plyne, že v diskrétním případě je rozptyl dán vzorcem

$$D(X) = \sum_{x=-\infty}^{\infty} [x - E(X)]^2 \pi(x)$$

a ve spojitém případě vzorcem

$$D(X) = \int_{x=-\infty}^{\infty} [x - E(X)]^2 \varphi(x)\, dx$$

(pokud suma či integrál vpravo absolutně konvergují).

(Rozptyl je číslo, které charakterizuje proměnlivost realizací náhodné veličiny kolem její střední hodnoty s přihlédnutím k jejich pravděpodobnostem. Je teoretickým protějškem váženého rozptylu zavedeného v definici 3.20.)

9.12. Příklad
Náhodná veličina X udává počet ok při hodu kostkou. Vypočtěte její rozptyl.

Řešení:

$$\pi(x) = \begin{cases} \frac{1}{6} & \text{pro } x = 1, 2, \ldots, 6, \\ 0 & \text{jinak,} \end{cases} \qquad E(X) = 3{,}5 \quad \text{(viz př. 9.8),}$$

$$D(X) = \sum_{x=1}^{6} (x - 3{,}5)^2 \frac{1}{6} = \cdots = \frac{35}{12} = 2{,}92.$$

9.13. Věta
Uveďme střední hodnoty a rozptyly vybraných typů diskrétních a spojitých rozložení.

a) $X \sim Dg(\mu) \;\Rightarrow\; E(X) = \mu,\ D(X) = 0,$

b) $X \sim A(\nu) \;\Rightarrow\; E(X) = \nu,\ D(X) = \nu(1-\nu),$

c) $X \sim Bi(n,\nu) \;\Rightarrow\; E(X) = n\nu,\ D(X) = n\nu(1-\nu),$

d) $X \sim Ge(\nu) \;\Rightarrow\; E(X) = \frac{1-\nu}{\nu},\ D(X) = \frac{1-\nu}{\nu^2},$

e) $X \sim Hg(N,M,n) \;\Rightarrow\; E(X) = \frac{M}{N}n,\ D(X) = \frac{MN}{N}(1 - \frac{M}{N})\frac{N-n}{N-1},$

f) $X \sim Rd(G) \;\Rightarrow\; E(X) = \frac{n-1}{2},\ D(X) = \frac{n^2-1}{12},$

g) $X \sim Po(\lambda) \;\Rightarrow\; E(X) = \lambda,\ D(X) = \lambda^2,$

h) $X \sim Rs(a,b) \;\Rightarrow\; E(X) = \frac{a+b}{2},\ D(X) = \frac{(b-a)^2}{12},$

i) $X \sim Ex(\lambda) \;\Rightarrow\; E(X) = \frac{1}{\lambda},\ D(X) = \frac{1}{\lambda^2},$

j) $X \sim N(\mu, \sigma^2) \Rightarrow E(X) = \mu,\ D(X) = \sigma^2$,

k) $X \sim \chi^2(n) \Rightarrow E(X) = n,\ D(X) = 2n$,

l) $X \sim t(n) \Rightarrow E(X) = 0$ pro $n \geq 2$, pro $n = 1$ $E(X)$ neexistuje, $D(X) = \frac{n}{n-2}$ pro $n \geq 3$, pro $n = 1, 2$ $D(X)$ neexistuje,

m) $X \sim F(n_1, n_2) \Rightarrow E(X) = \frac{n_2}{n_2-2}$ pro $n_2 \geq 3$, pro $n_2 = 1, 2$ $E(X)$ neexistuje, $D(X) = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)(n_2-4)}$ pro $n_2 \geq 5$, pro $n_2 = 1, 2, 3, 4$ $D(X)$ neexistuje.

Věnujme se nyní dvěma náhodným veličinám. Budou nás zajímat charakteristiky jejich společné variability a síly těsnosti lineárního vztahu mezi nimi.

Jako motivace pro zavedení těchto charakteristik nám poslouží empirická kovariance s_{12} a empirický koeficient korelace r_{12}. Empirická kovariance s_{12} byla definována jako aritmetický průměr součinů centrovaných hodnot a empirický koeficient korelace r_{12} jako aritmetický průměr součinů standardizovaných hodnot. Lze tedy očekávat, že teoretická kovariance $C(X_1, X_2)$ bude střední hodnotou součinů centrovaných hodnot a teoretický rozptyl $R(X_1, X_2)$ bude střední hodnotou součinů standardizovaných veličin.

Podrobně se seznámíme s řadou vlastností všech výše uvedených číselných charakteristik a využijeme jich při řešení několika příkladů.

Pokud neznáme rozložení pravděpodobnosti náhodné veličiny, ale jenom její střední hodnotu a rozptyl, pak můžeme pomocí tzv. Čebyševovy nerovnosti aspoň odhadnout pravděpodobnost, že tato náhodná veličina se od své střední hodnoty odchýlí o více než t-násobek své směrodatné odchylky.

V závěru kapitoly se soustředíme na vlastnosti střední hodnoty a rozptylu náhodné veličiny s normálním rozložením.

9.14. Definice
Kovariancí náhodných veličin X_1, X_2, které mají střední hodnoty $E(X_1)$, $E(X_2)$, rozumíme číslo

$$C(X_1, X_2) = E([X_1 - E(X_1)][X_2 - E(X_2)])$$

(pokud střední hodnoty vpravo existují). Z věty 9.9 (b) plyne, že v diskrétním případě je kovariance dána vzorcem

$$C(X_1, X_2) = \sum_{x_1=-\infty}^{\infty} \sum_{x_2=-\infty}^{\infty} [x_1 - E(X_1)][x_2 - E(X_2)]\pi(x_1, x_2)$$

a ve spojitém případě vzorcem

$$C(X_1, X_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [x_1 - E(X_1)][x_2 - E(X_2)]\varphi(x_1, x_2)\, dx_1 dx_2$$

(pokud dvojná suma či dvojný integrál vpravo absolutně konvergují).

(Kovariance je číslo, které charakterizuje proměnlivost realizací náhodných veličin X_1, X_2 kolem jejich středních hodnot s přihlédnutím k jejich pravděpodobnostem. Je-li kovariance kladná (záporná), pak to svědčí o existenci jistého stupně přímé (nepřímé) lineární závislosti mezi realizacemi náhodných veličin X_1, X_2. Je-li kovariance nulová, pak říkáme, že náhodné veličiny X_1, X_2 jsou nekorelované a znamená to, že mezi jejich realizacemi není žádný lineární vztah. Pozor – z nekorelovanosti nevyplývá stochastická nezávislost, zatímco ze stochastické nezávislosti plyne nekorelovanost. Kovariance je teoretickým protějškem vážené kovariance z definice 3.20.)

9.15. Příklad

Diskrétní náhodný vektor (X_1, X_2) má simultánní pravděpodobnostní funkci s hodnotami: $\pi(0,-1) = c$, $\pi(0,0) = \pi(0,1) = \pi(1,-1) = \pi(2,-1) = 0$, $\pi(1,0) = \pi(0,1) = \pi(2,1) = 2c$, $\pi(2,0) = 3c$, $\pi(x_1, x_2) = 0$ jinak. Určete konstantu c a vypočtěte $C(X_1, X_2)$.

Řešení:

Hodnoty simultánní pravděpodobnostní funkce a obou marginálních pravděpodobnostních funkcí uspořádáme do kontingenční tabulky.

		x_2			$\pi_1(x_1)$
		-1	0	1	
x_1	0	c	0	0	c
	1	0	$2c$	$2c$	$4c$
	2	0	$3c$	$2c$	$5c$
$\pi_2(x_2)$		c	$5c$	$4c$	1

Z normovanosti pravděpodobnostní funkce diskrétního náhodného vektoru (viz věta 7.8, vektorový případ) dostáváme $10c = 1$, tedy $c = 0{,}1$.

$$E(X_1) = \sum_{x_1=0}^{2} x_1 \pi_1(x_1) = 0 \cdot 0{,}1 + 1 \cdot 0{,}4 + 2 \cdot 0{,}5 = 1{,}4$$

$$E(X_2) = \sum_{x_2=-1}^{1} x_2 \pi_2(x_2) = -1 \cdot 0{,}1 + 0 \cdot 0{,}5 + 1 \cdot 0{,}4 = 0{,}3$$

$$\begin{aligned} C(X_1, X_2) &= \sum_{x_1=0}^{2} \sum_{x_2=-1}^{1} [x_1 - E(X_1)][x_2 - E(X_2)]\pi(x_1, x_2) = \\ &= (0 - 1{,}4) \cdot (-1 - 0{,}3) \cdot 0{,}1 + \cdots + (2 - 1{,}4) \cdot (1 - 0{,}3) \cdot 0{,}2 = 0{,}18. \end{aligned}$$

9.16. Definice

Koeficientem korelace náhodných veličin X_1, X_2 rozumíme číslo

$$R(X_1, X_2) = \begin{cases} E\left(\frac{X_1 - E(X_1)}{\sqrt{D(X_1)}} \cdot \frac{X_2 - E(X_2)}{\sqrt{D(X_2)}} \right) & \text{pro } \sqrt{D(X_1)}\sqrt{D(X_2)} > 0, \\ 0 & \text{jinak.} \end{cases}$$

(Koeficient korelace je číslo, které charakterizuje těsnost lineární závislosti realizací náhodných veličin X_1, X_2. Čím bližší je 1, tím těsnější je přímá lineární závislost, čím bližší je -1, tím těsnější je nepřímá lineární závislost.)

9.17. Věta
Nechť a, a_1, a_2, b, b_1, b_2 jsou reálná čísla, X, $X_1, \dots, X_n$, $Y_1, \dots, Y_m$ jsou náhodné veličiny definované na témž pravděpodobnostním prostoru. V následujících vzorcích vždy z existence číselných charakteristik na pravé straně vyplývá existence výrazu na levé straně.

Vlastnosti střední hodnoty

a) $E(a) = a$,
b) $E(a + bX) = a + bE(X)$,
c) $E(X - E(X)) = 0$,
d) $E\left(\sum\limits_{i=1}^{n} X_i\right) = \sum\limits_{i=1}^{n} E(X_i)$,
e) Jsou-li náhodné veličiny $X_1, \dots, X_n$ stochasticky nezávislé, pak platí $E\left(\prod\limits_{i=1}^{n} X_i\right) = \prod\limits_{i=1}^{n} E(X_i)$.

Vlastnosti kovariance

a) $C(a_1, X_2) = C(X_1, a_2) = C(a_1, a_2) = 0$,
b) $C(a_1 + b_1X_1, a_2 + b_2X_2) = b_1b_2C(X_1, X_2)$,
c) $C(X, X) = D(X)$,
d) $C(X_1, X_2) = C(X_2, X_1)$,
e) $C(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2)$,
f) $C\left(\sum\limits_{i=1}^{n} X_i, \sum\limits_{j=1}^{m} Y_j\right) = \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m} C(X_i, Y_j)$.

Vlastnosti rozptylu

a) $D(a) = 0$,
b) $D(a + bX) = b^2D(X)$,
c) $D(X) = E(X^2) - [E(X)]^2$,
d) $D\left(\sum\limits_{i=1}^{n} X_i\right) = \sum\limits_{i=1}^{n} D(X_i) + 2\sum\limits_{i=1}^{n-1}\sum\limits_{j=i+1}^{n} C(X_i, X_j)$ (Jsou-li náhodné veličiny $X_1, \dots, X_n$ nekorelované, pak $D\left(\sum\limits_{i=1}^{n} X_i\right) = \sum\limits_{i=1}^{n} D(X_i)$.)

Vlastnosti koeficientu korelace

a) $R(a_1, X_2) = R(X_1, a_2) = R(a_1, a_2) = 0$,
b) $R(a_1 + b_1X_1, a_2 + b_2X_2) = \operatorname{sgn}(b_1b_2)R(X_1, X_2)$,
c) $R(X, X) = 1$ pro $D(X) \neq 0$, $R(X, X) = 0$ jinak,
d) $R(X_1, X_2) = R(X_2, X_1)$

e) $R(X1, X2) = \begin{cases} E\left(\frac{C(X_1,X_2)}{\sqrt{D(X_1)}\sqrt{D(X_2)}}\right) & \text{pro } \sqrt{D(X_1)}\sqrt{D(X_2)} > 0, \\ 0 & \text{jinak,} \end{cases}$

f) $|R(X_1, X_2)| \leq 1$ a rovnost nastane tehdy a jen tehdy, když mezi veličinami X_1, X_2 existuje s pravděpodobností 1 úplná lineární závislost, tj. existují konstanty a_1, a_2 tak, že $P(X_2 = a_1 + a_2X_1) = 1$. (Uvedená nerovnost se nazývá Cauchyova-Schwarzova-Buňakovského nerovnost.)

9.18. Příklad
Vypočtěte koeficient korelace náhodných veličin X_1, X_2 z příkladu 9.15.

Řešení:
V příkladu 9.15 byla vypočtena kovariance $C(X_1, X_2) = 0{,}18$. Stačí tedy vypočítat směrodatné odchylky veličin X_1, X_2.

$$D(X_1) = \sum_{x_1=0}^{2} [x_1 - E(X_1)]^2 \pi_1(x_1) =$$
$$= (0 - 1{,}4)^2 \cdot 0{,}1 + (1 - 1{,}4)^2 \cdot 0{,}4 + (2 - 1{,}4)^2 \cdot 0{,}5 = 0{,}44$$
$$D(X_2) = \sum_{x_2=0}^{2} [x_2 - E(X_2)]^2 \pi_1(x_2) =$$
$$= (-1 - 0{,}3)^2 \cdot 0{,}1 + (0 - 0{,}3)^2 \cdot 0{,}5 + (1 - 0{,}3)^2 \cdot 0{,}4 = 0{,}41$$
$$R(X_1, X_2) = \frac{C(X_1, X_2)}{\sqrt{D(X_1)}\sqrt{D(X_2)}} = \frac{0{,}18}{\sqrt{0{,}44}\sqrt{0{,}41}} = 0{,}42.$$

9.19. Příklad
Náhodná veličina X má střední hodnotu μ a rozptyl σ^2. Vypočtěte střední hodnotu a rozptyl centrované náhodné veličiny $Y = X - \mu$ a střední hodnotu a rozptyl standardizované náhodné veličiny $U = \frac{X-\mu}{\sigma}$.

Řešení:

$$E(Y) = E(X - \mu) = E(X) - E(\mu) = \mu - \mu = 0,$$
$$D(Y) = D(X - \mu) = D(X) = \sigma^2,$$
$$E(U) = E\left(\frac{X - \mu}{\sigma}\right) = \frac{1}{\sigma}E(X - \mu) = \frac{1}{\sigma} \cdot 0 = 0,$$
$$D(U) = D\left(\frac{X - \mu}{\sigma}\right) = \frac{1}{\sigma^2}D(X - \mu) = \frac{1}{\sigma^2}\sigma^2 = 1.$$

9.20. Příklad
Náhodné veličiny X, Y jsou náhodné chyby, které vznikají na vstupním zařízení. Mají střední hodnoty $E(X) = -2$, $E(Y) = 4$ a rozptyly $D(X) = 4$,

$D(Y) = 9$. Koeficient korelace těchto chyb je $R(X, Y) = -0{,}5$. Chyba na výstupu zařízení souvisí s chybami na vstupu funkční závislostí $Z = 3X^2 - 2XY + Y^2 - 3$. Najděte střední hodnotu chyby na výstupu.

Řešení:

$$\begin{aligned}
E(Z) &= E(3X^2 - 2XY + Y^2 - 3) = 3E(X^2) - 2E(XY) + E(Y^2) - E(3) = \\
&= 3\left\{D(X) + [E(X)]^2\right\} - 2\left[C(X,Y) + E(X)E(Y)\right] + D(Y) + [E(Y)]^2 - 3 = \\
&= 3[D(X) + [E(X)]^2] - 2[R(X,Y)\sqrt{D(X)}\sqrt{D(Y)} + E(X)E(Y)] + D(Y) + \\
&\quad + [E(Y)]^2 - 3 = 3(4+4) - 2[-0{,}5 \cdot 2 \cdot 3 + (-2) \cdot 4] + 9 + 16 - 3 = \\
&= 24 + 22 + 25 - 3 = 68.
\end{aligned}$$

9.21. Věta

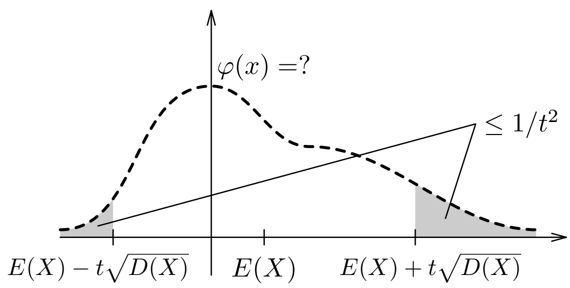
Nechť náhodná veličina X má střední hodnotu μ a rozptyl σ^2. Pak platí Čebyševova nerovnost

$$\forall \varepsilon > 0 : P(|X - \mu| > \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}.$$

Označíme-li $\varepsilon = t\sigma$, pak pro

$$\forall t > 0 : P(|X - \mu| > t\sigma) \leq \frac{1}{t^2}.$$

(Význam Čebyševovy nerovnosti spočívá v tom, že pokud neznáme rozložení náhodné veličiny, ale známe její střední hodnotu a rozptyl, pak můžeme odhadnout pravděpodobnost, s jakou se od své střední hodnoty odchýlí o více než t-násobek své směrodatné odchylky.)



9.22. Příklad

Nechť $E(X) = \mu$, $D(X) = \sigma^2$.

a) Odhadněte $P(|X - \mu > 3\sigma)$.
b) Jestliže $X \sim N(\mu, \sigma^2)$, vypočtěte $P(|X - \mu| > 3\sigma)$.

Řešení:

ad a) $P(|X - \mu| > 3\sigma) \leq \frac{1}{3^2} = \frac{1}{9} = 0{,}\overline{1}$.

(Tento výsledek je znám jako pravidlo 3σ a říká, že nejvýše 11,1% realizací

náhodné veličiny leží vně intervalu $(\mu - 3\sigma, \mu + 3\sigma)$.)
ad b) $P(|X-\mu| > 3\sigma) = 1 - P(-3\sigma \le X - \mu \le 3\sigma) = 1 - P\left(-3 \le \frac{X-\mu}{\sigma} \le 3\right)$ $= 1 - \Phi(3) + \Phi(-3) = 2[1 - \Phi(3)] = 2(1 - 0{,}99865) = 0{,}0027$.
(Má-li náhodná veličina normální rozdělení, pak pouze 0,27% realizací leží vně intervalu $(\mu - 3\sigma, \mu + 3\sigma)$.)

9.23. Věta

a) Jestliže $X \sim N(\mu, \sigma^2)$, pak $E(X) = \mu$, $D(X) = \sigma^2$.
b) Jestliže $X \sim N(\mu, \sigma^2)$ a $Y = a + bX$, pak $Y \sim N(a + b\mu, b^2\sigma^2)$.
c) Jestliže $X_1, \dots, X_n$ jsou stochasticky nezávislé náhodné veličiny a nechť $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \dots, n$, $Y = \sum\limits_{i=1}^{n} X_i$, pak

$$Y \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

.

9.24. Příklad

Nechť X_1, X_2 jsou stochasticky nezávislé náhodné veličiny, $X_i \sim N(0,1)$, $i = 1, 2$. Zjistěte, jaké rozložení má transformovaná náhodná veličina $Y = 3 + X_1 - 2X_2$, určete jeho parametry a najděte dolní kvartil náhodné veličiny Y.

Řešení:
$Y \sim N(E(Y), D(Y))$, přičemž

$$\begin{aligned} E(Y) &= E(3 + X_1 - 2X_2) = 3 + E(X_1) - 2E(X_2) = 3 + 0 - 2 \cdot 0 = 3, \\ D(Y) &= D(3 + X_1 - 2X_2) = D(X_1) + (-2)^2 D(X_2) = 1 + 4 \cdot 1 = 5, \end{aligned}$$

tedy $Y \sim N(3, 5)$. Nyní vypočítáme dolní kvartil. Využijeme toho, že $U = \frac{Y-2}{\sqrt{5}} \sim N(0,1)$, tedy $K_{0,25}(Y) = 3 + \sqrt{5}u_{0,25} = 3 - \sqrt{5} \cdot 0{,}67449 = 1{,}4918$.

Shrnutí kapitoly

Při zavádění číselných charakteristik náhodných veličin nás motivují číselné charakteristiky znaků, jak jsme je poznali ve 3. kapitole.

Jako charakteristika polohy číselných realizací spojité náhodné veličiny aspoň ordinálního typu slouží **α-kvantil** a jeho speciální případy: **medián**, **dolní** a **horní kvartil**. Variabilitu charakterizujeme **kvartilovou odchylkou**. Výpočet kvantilů není příliš jednoduchá záležitost, proto jsou kvantily několika typů rozložení tabelovány nebo je lze získat pomocí speciálního statistického software.

Pro náhodné veličiny intervalového a poměrového typu používáme jako charakteristiku polohy **střední hodnotu** – teoretický protějšek aritmetického průměru. Pomocí střední hodnoty pak definujeme další číselné charakteristiky: **rozptyl** a jeho druhou odmocninu – **směrodatnou odchylku**, **kovarianci** a **koeficient korelace**.

Řešení konkrétních příkladů velmi usnadňují vzorce, které popisují **vlastnosti číselných charakteristik**.

Kontrolní otázky a úkoly

1 Pomocí statistických tabulek vypočtěte následující kvantily: $u_{0,95}$, $u_{0,10}$, $\chi^2_{0,975}(10)$, $\chi^2_{0,025}(9)$, $t_{0,90}(8)$, $t_{0,05}(6)$, $F_{0,975}(5,7)$, $F_{0,055}(8,6)$.

2 Nechť $X \sim N(-1,4)$. Najděte $K_{0,025}(X)$.

3 Nechť X_1, X_2 jsou stochasticky nezávislé náhodné veličiny takové, že $X_1 \sim N(2,4)$, $X_2 \sim N(-1,9)$. Vypočtěte 99% kvantil transformované náhodné veličiny $Y = 2X_1 - 3X_2 + 5$.

4 V zásilce 15 výrobků je 5 nekvalitních. Náhodná veličina X udává počet nekvalitních výrobků mezi čtyřmi náhodně vybranými výrobky. Vypočtěte její střední hodnotu a rozptyl, jestliže výběr byl proveden a) s vracením, b) bez vracení. (Návod: v bodě (a) má X binomické rozložení, v bodě (b) hypergeometrické.)

5 Sledovaná železniční trasa vykazuje velké nerovnosti, takže zatížení jednotlivé vozové nápravy náhodně kolísá, teoreticky spojitým způsobem. Prakticky jsou známy jen částečné informace, takže uvažujeme o diskrétní náhodné veličině X (náhodné zatížení v tunách) s pravděpodobnostní funkcí $\pi(x) = 0{,}15$ pro $x = 6$, $\pi(x) = 0{,}65$ pro $x = 30$, $\pi(x) = 0{,}2$ pro $x = 70$, $\pi(x) = 0$ jinak. Při kalkulaci nákladů se ekonom zajímá o střední opotřebení náprav dané vzorcem $Y = 1{,}15X^2$. Vypočtěte střední hodnotu opotřebení.

6 Počet různých druhů zboží, které zákazník nakoupí při jedné návštěvě obchodu, je náhodná veličina X. Dlouhodobým sledováním bylo zjištěno, že X nabývá hodnot $0, 1, 2, 3, 4$ s pravděpodobnostmi 0,25, 0,55, 0,11, 0,07 a 0,02.

a) Najděte distribuční funkci náhodné veličiny X a nakreslete její graf.
b) Vypočtěte střední hodnotu náhodné veličiny X.
c) Vypočtěte rozptyl náhodné veličiny X.

7 Střelec střílí 3× nezávisle na sobě do terče. Při každém výstřelu se trefí s pravděpodobností $\frac{3}{4}$. Za zásah získá 2 body, jinak ztratí 2 body. Vypočtěte střední hodnotu a rozptyl počtu získaných bodů.

8 Uvažme rodinu se třemi dětmi. Předpokládáme, že pravděpodobnost narození chlapce i dívky je stejná. Náhodná veličina X udává počet dívek v této rodině (má binomické rozložení) , transformovaná náhodná veličina $Y = -100X^2 + 300X + 500$ udává roční náklady (v dolarech) na ošacení dětí. Vypočtěte střední hodnotu náhodné veličiny Y.

9 Náhodná veličina X udává příjem manžela (v tisících dolarů) a náhodná veličina Y udává příjem manželky (v tisících dolarů). Je známa simultánní pravděpodobnostní funkce $\pi(x,y)$ diskrétního náhodného vektoru (X,Y): $\pi(10,10) = 0{,}2$, $\pi(10,20) = 0{,}04$, $\pi(10,30) = 0{,}01$, $\pi(10,40) = 0$, $\pi(20,10) = 0{,}1$, $\pi(20,20) = 0{,}36$, $\pi(20,30) = 0{,}09$, $\pi(20,40) = 0$, $\pi(30,10) = 0$, $\pi(30,20) = 0{,}05$, $\pi(30,30) = 0{,}1$,

$\pi(30,40) = 0$, $\pi(40,10) = 0$, $\pi(40,20) = 0$, $\pi(40,30) = 0$, $\pi(40,40) = 0{,}05$, $\pi(x,y) = 0$ jinak.

a) Vypočtěte korelační koeficient náhodných veličin X, Y.

b) Vypočtěte střední hodnotu a směrodatnou odchylku náhodné veličiny $Z = 0{,}1X + 0{,}2Y$, která vyjadřuje příspěvek obou manželů na důchod. (Náhodná veličina Z vyjadřuje, že příspěvek na důchod činí 10% manželova platu a 20% manželčina platu.)

10 Náhodné veličiny X_1, X_2 mají kovarianci 12. Vypočtěte kovarianci náhodných veličin $Y_1 = -8 + 11X_1$, $Y_2 = 6 - 4X_2$.

11 Náhodná veličina X udává výšku v metrech a náhodná veličina Y udává hmotnost v gramech. Jak se změní kovariance a koeficient korelace, jestliže výšku vyjádříme v cm a hmotnost v kg?

12 Náhodná veličina X má střední hodnotu μ a směrodatnou odchylku σ. Kolik procent realizací této náhodné veličiny se bude nacházet v intervalu $(\mu - 2\sigma, \mu + 2\sigma)$?

13 Použijte Čebyševovu nerovnost k odhadu pravděpodobnosti, že při 600 hodech kostkou padne šestka aspoň 75× a nejvýše 125×.

10. Zákon velkých čísel a centrální limitní věta

Cíl kapitoly

Po prostudování této kapitoly budete umět:

- odhadnout pravděpodobnost, s níž se náhodná veličina realizuje v určité vzdálenosti od své střední hodnoty
- odhadnout pravděpodobnost úspěchu v posloupnosti opakovaných nezávislých pokusů relativní četností tohoto úspěchu
- aproximovat distribuční funkci binomického rozložení distribuční funkcí standardizovaného normálního rozložení

Časová zátěž

Na prostudování této kapitoly budete potřebovat asi 5 hodin studia.

V 5. kapitole, konkrétně v definici 5.6, jsme se seznámili s empirickým zákonem velkých čísel, který tvrdil, že při mnohonásobném nezávislém opakování téhož náhodného pokusu se relativní četnost jevu blíží pravděpodobnosti tohoto jevu. Jak uvidíme, je empirický zákon velkých čísel speciálním případem obecnějšího zákona velkých čísel. Tento důsledek uvedeme jako Bernoulliovu větu.

10.1. Motivace
Zákon velkých čísel vyjadřuje skutečnost, že s rostoucím počtem nezávislých opakování náhodného pokusu se empirické charakteristiky, které popisují výsledky těchto pokusů, blíží teoretickým charakteristikám, např. relativní četnost úspěchu se blíží pravděpodobnosti úspěchu, četnostní funkce se blíží pravděpodobnostní funkci, hustota četnosti se blíží hustotě pravděpodobnosti apod.

Centrální limitní věta tvrdí, že za jistých podmínek má součet nezávislých náhodných veličin s týmž rozložením přibližně normální rozložení. Normální rozložení je tedy rozložením limitním, k němuž se blíží všechna rozložení, proto hraje velmi důležitou roli v počtu pravděpodobnosti a matematické statistice.

10.2. Věta
Nechť $\{X_n\}_{n=1}^{\infty}$ je posloupnost stochasticky nezávislých náhodných veličin, které mají střední hodnoty μ a rozptyly σ^2. Pak pro posloupnost aritmetických průměrů $\{\frac{1}{n}\sum\limits_{i=1}^{n} X_i\}_{i=1}^{\infty}$ platí:

$$\forall \varepsilon > 0 : P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| < \varepsilon\right) \geq 1 - \frac{\sigma^2}{n\varepsilon^2},$$

neboli

$$\forall \varepsilon > 0 : \lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \varepsilon\right) = 0.$$

(Uvedená věta se nazývá zákon velkých čísel nebo též Čebyševova věta. Její tvrzení říká, že posloupnost aritmetických průměrů konverguje podle pravděpodobnosti ke střední hodnotě μ. Tedy při dostatečně velkém počtu pokusů lze střední hodnotu odhadnout průměrem výsledků jednotlivých pokusů.)

10.3. Důsledek

Nechť náhodná veličina Y_n udává počet úspěchů v posloupnosti n opakovaných nezávislých pokusů, přičemž v každém pokusu nastává úspěch s pravděpodobností ν. (Podle definice 8.2 (c) $Y_n \sim Bi(n, \nu)$). Pak pro posloupnost relativních četností $\{\frac{Y_n}{n}\}_{n=1}^{\infty}$ platí:

$$\forall \varepsilon > 0 : P\left(\left|\frac{Y_n}{n} - \vartheta\right| < \varepsilon\right) \geq 1 - \frac{\vartheta(1-\vartheta)}{n\varepsilon^2} > 1 - \frac{1}{4n\varepsilon^2},$$

neboli

$$\forall \varepsilon > 0 : \lim_{n\to\infty} P\left(\left|\frac{Y_n}{n} - \vartheta\right| > \varepsilon\right) = 0.$$

(Tento důsledek Čebyševovy věty se nazývá Bernoulliova věta. Vyjadřuje skutečnost, že posloupnost relativních četností konverguje podle pravděpodobnosti k pravděpodobnosti úspěchu ν. Tedy při dostatečně velkém počtu pokusů lze pravděpodobnost úspěchu odhadnout relativní četností úspěchu.)

10.4. Příklad

Při výstupní kontrole bylo zjištěno, že mezi 3000 kontrolovanými výrobky je 12 zmetků. Jaká je pravděpodobnost, že relativní četnost výskytu zmetku se od pravděpodobnosti výskytu zmetku neliší o více než 0, 01?

Řešení:

Y_{3000} – počet zmetků mezi kontrolovanými výrobky, $Y_{3000} \sim Bi(3000, \nu)$, $\nu \approx \frac{12}{3000}$. Podle Bernoulliovy věty dostáváme:

$$\forall \varepsilon > 0 : P\left(\left|\frac{Y_n}{n} - \vartheta\right| < \varepsilon\right) \geq 1 - \frac{\vartheta(1-\vartheta)}{n\varepsilon^2} > 1 - \frac{1}{4n\varepsilon^2}.$$

V našem případě $\varepsilon = 0{,}01$, $n = 3000$, $\nu \approx \frac{12}{3000}$, tedy

$$P\left(\left|\frac{Y_{3000}}{3000} - \vartheta\right| < 0{,}01\right) \geq 1 - \frac{\frac{12}{3000}\frac{2988}{3000}}{3000 \cdot 0{,}0001} = 0{,}872.$$

Již několikrát jsme se zmínili o tom, že normální rozložení je vůbec nejdůležitější typ rozložení. Centrální limitní věta nám dá odpověď na otázku, proč tomu tak je.

Při praktických výpočtech se často používá důsledek centrální limitní věty, a to Moivreova-Laplaceova věta, která za určitých podmínek umožní nahradit složitý výpočet distribuční funkce binomického rozložení jednoduchým hledáním v tabulkách hodnot distribuční funkce standardizovaného normálního rozložení. Pokud však máme k dispozici statistický software, dáme přednost přesnému výpočtu před aproximativním.

10.5. Věta
Nechť $\{X_n\}_{n=1}^{\infty}$ je posloupnost stochasticky nezávislých náhodných veličin, které mají všechny totéž rozložení se střední hodnotou μ a rozptylem σ^2. Pak pro posloupnost standardizovaných součtů

$$U_n = \frac{\sum\limits_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}, \qquad n = 1, 2, \ldots$$

platí: $\forall x \in \mathbb{R} : \lim\limits_{x\to\infty} P(U_n \leq x) = \Phi(x)$, kde $\Phi(x)$ je distribuční funkce rozložení $N(0,1)$.

(Lindebergova-Lévyova centrální limitní věta říká, že pro dostatečně velká n (praktickz stačí $n \geq 30$) lze rozložení součtu stochasticky nezávislých a stejně rozložených náhodných veličin aproximovat normálním rozložením $N(n\mu, n\sigma^2)$.)

10.6. Důsledek
Nechť $\{Y_n\}_{n=1}^{\infty}$ je posloupnost stochasticky nezávislých náhodných veličin, $Y_n \sim Bi(n,\nu)$, $n = 1, 2, \ldots$ Pak platí:

$$\forall y \in \mathbb{R} : \lim_{n\to\infty} P(Y_n \leq y) = \lim_{n\to\infty} P\left(\frac{Y_n - n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}} \leq \frac{y - n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}}\right) \approx$$
$$\approx \Phi\left(\frac{y - n\vartheta}{\sqrt{n\vartheta(1-\vartheta)}}\right),$$

kde $\Phi(x)$ je distribuční funkce rozložení $N(0,1)$.

(Moivreova-Laplaceova věta tvrdí, že za určitých podmínek lze binomické rozložení aproximovat standardizovaným normálním rozložením. Aproximace se považuje za vyhovující, když jsou splněny podmínky $\frac{1}{n+1} < \nu < \frac{n}{n+1}$ a $n\nu(1-\nu) > 9$.)

10.7. Příklad
V určité skupině zaměstnanců je 10% s příjmem, který překračuje celostátní průměr. Kolik zaměstnanců z této skupiny je třeba vybrat, aby s pravděpodobností aspoň 0,95 bylo mezi nimi 8% až 12% zaměstnanců s nadprůměrným příjmem?

Řešení:
X – počet zaměstnanců s nadprůměrným příjmem, $Y_n \sim Bi(n; 0{,}1)$, $E(X) =$

$0{,}1n,\ D(X) = 0{,}09n,$

$$0{,}95 \le P\left(0{,}08 \le \frac{X}{n} \le 0{,}12\right) = P(0{,}08n \le X \le 0{,}12n) =$$
$$= P\left(\frac{0{,}08 - 0{,}1n}{\sqrt{0{,}09n}} \le \frac{X - 0{,}1n}{\sqrt{0{,}09n}} \le \frac{0{,}12 - 0{,}1n}{\sqrt{0{,}09n}}\right) =$$
$$= P\left(\frac{-\sqrt{n}}{15} \le \frac{X - 0{,}1n}{\sqrt{0{,}09n}} \le \frac{\sqrt{n}}{15}\right) \approx \Phi\left(\frac{\sqrt{n}}{15}\right) - \Phi\left(-\frac{\sqrt{n}}{15}\right) =$$
$$= 2\Phi\left(\frac{\sqrt{n}}{15}\right) - 1 \Rightarrow \Phi\left(\frac{\sqrt{n}}{15}\right) \ge 0{,}975$$

tedy $\frac{\sqrt{n}}{15} \ge u_{0{,}975} = 1{,}96 \Rightarrow \sqrt{n} \ge 29{,}4 \Rightarrow n \ge 865$. Pro splnění podmínek je zapotřebí vybrat aspoň 865 zaměstnanců.

Shrnutí kapitoly

V této kapitole jsme ukázali, že již dříve vyslovený empirický zákon velkých čísel je speciálním případem obecnějšího **zákona velkých čísel**, který popisuje pravděpodobnostní chování posloupností aritmetických průměrů stochasticky nezávislých náhodných veličin s touž střední hodnotou a rozptylem. Důsledek tohoto zákona (zvaného též **Čebyševova věta**) jsme uvedli jako **Bernoulliovu větu**.

Seznámili jsme se též s **Lindebergovou-Lévyovou centrální větou**, která tvrdí, že za určitých podmínek lze rozložení součtu náhodných věličin s jakýmkoliv rozložením aproximovat normálním rozložením. Toto tvrzení tedy vysvětluje důležitost normálního rozložení. Historicky starší než tato věta je její důsledek uváděný jako **Moivreova-Laplaceova věta**, která umožňuje aproximovat binomické rozložení normálním rozložením.

Kontrolní otázky a úkoly

1 Pravděpodobnost, že výrobek má 1. jakost, je $\nu = 0{,}9$. Kolik výrobků je třeba zkontrolovat, aby s pravděpodobností aspoň 0,99 bylo zaručeno, že rozdíl relativní četnosti počtu výrobků 1. jakosti a pravděpodobnosti $\nu = 0{,}9$ byl v absolutní hodnotě menší než 0,03? K výpočtu použijte jak Bernoulliovu větu, tak Moivreovu-Laplaceovu větu a výsledky porovnejte.

2 Pravděpodobnost narození chlapce je 0,515. Jaká je pravděpodobnost, že mezi 10 000 novorozenci bude
- a) více děvčat než chlapců,
- b) chlapců od 5 000 do 5 300,
- c) relativní četnost chlapců v mezích od 0,515 do 0,517?

3 Pravděpodobnost zásahu terče jedním výstřelem je 0,4. Kolikrát je třeba vystřelit, aby absolutní hodnota odchylky relativní četnosti zásahů od uvedené pravděpodobnosti byla menší než 0,02 s pravděpodobností aspoň 0,95?

11. Základní pojmy matematické statistiky

Cíl kapitoly

Po prostudování této kapitoly budete umět:

- definovat náhodné výběry z jednorozměrného i vícerozměrného rozložení pravděpodobností
- stanovit důležité statistiky pro náhodný výběr z jednorozměrného a dvourozměrného rozložení pravděpodobností
- popsat vlastnosti těchto statistik
- využít vlastností statistik odvozených z náhodného výběru z normálního rozložení při výpočtu konkrétních pravděpodobností

Časová zátěž

Pro zvládnutí této kapitoly budete potřebovat asi 7 hodin studia.

Nejprve zavedeme pojem náhodného výběru a vysvětlíme jeho souvislost s datovým souborem. Musíme si však uvědomit následující skutečnost: datový soubor obsahuje konstantní hodnoty znaků, zatímco složkami náhodného výběru jsou náhodné veličiny spojené s nějakým náhodným pokusem.

11.1. Definice

a) Nechť $X_1, \dots, X_n$ jsou stochasticky nezávislé náhodné veličiny, které mají všechny stejné rozložení $L(\nu)$. Řekneme, že $X_1, \dots, X_n$ je *náhodný výběr rozsahu* n z rozložení $L(\nu)$. (Číselné realizace $x_1, \dots, x_n$ náhodného výběru $X_1, \dots, X_n$ uspořádané do sloupcového vektoru představují datový soubor zavedený v popisné statistice v definici 1.9)

b) Nechť $(X_1, Y_1), \dots, (X_n, Y_n)$ jsou stochasticky nezávislé dvourozměrné náhodné vektory, které mají všechny stejné dvourozměrné rozložení $L_2(\nu)$. Řekneme, že $(X_1, Y_1), \dots, (X_n, Y_n)$ je *dvourozměrný náhodný výběr rozsahu* n z dvourozměrného rozložení $L_2(\nu)$. (Číselné realizace $(x_1, y_1), \dots, (x_n, y_n)$ náhodného výběru $(X_1, Y_1), \dots, (X_n, Y_n)$ uspořádané do matice typu $2 \times n$ představují dvourozměrný datový soubor zavedený v popisné statistice.)

(Analogicky lze definovat p-rozměrný náhodný výběr rozsahu n z p-rozměrného rozložení $L_p(\nu)$.)

V matematické statistice velmi často pracujeme s transformacemi náhodného výběru. Těmto transformovaným náhodným veličinám říkáme statistiky. Zavedeme několik důležitých statistik a upozorníme na jejich souvislost s číselnými charakteristikami znaků, které jsme poznali ve 3. kapitole v popisné statistice.

Protože statistiky jsou náhodnými veličinami, lze počítat jejich střední hodnotu a rozptyl. Ukážeme, jak se chovají tyto číselné charakteristiky některých statistik.

11.2. Definice
Libovolná funkce $T = T(X_1, \dots, X_n)$ náhodného výběru $X_1, \dots, X_n$ (resp. $T = T(X_1, Y_1, \dots, X_n, Y_n)$ náhodného výběru $(X_1, Y_1), \dots, (X_n, Y_n)$) se nazývá *(výběrová) statistika*.

Statistika
$$M = \frac{1}{n} \sum_{i=1}^{n} X_i$$
se nazývá *výběrový průměr*,
$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - M)^2$$
výběrový rozptyl,
$$S = \sqrt{S^2}$$
výběrová směrodatná odchylka,
$$S_{12} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - M_1)(Y_i - M_2)$$
výběrová kovariance (přitom $M_1 = \frac{1}{n} \sum\limits_{i=1}^{n} X_i$, $M_2 = \frac{1}{n} \sum\limits_{i=1}^{n} Y_i$) a
$$R_{12} = \begin{cases} \frac{1}{n-1} \sum\limits_{i=1}^{n} \frac{X_i - M_1}{S_1} \frac{Y_i - M_2}{S_2} & \text{pro } S_1, S_2 \neq 0, \\ 0 & \text{jinak,} \end{cases}$$
se nazývá *výběrový koeficient korelace*.

(Číselné realizace m, s^2, s, s_{12}, r_{12} statistik M, S^2, S, S_{12}, R_{12} odpovídají číselným charakteristikám znaků v popisné statistice zavedeným definicích 3.6, 3.10 a 3.12, ale u rozptylu, směrodatné odchylky, kovariance a koeficientu korelace je multiplikativní konstanta $\frac{1}{n-1}$, nikoli $\frac{1}{n}$, jak tomu bylo v popisné statistice.)

11.3. Věta

a) Nechť $X_1, \dots, X_n$ je náhodný výběr z rozložení se střední hodnotou μ a rozptylem σ^2. Pak $E(M) = \mu$, $D(M) = \frac{\sigma^2}{n}$, $E(S^2) = \sigma^2$, ať jsou hodnoty parametrů μ, σ^2 jakékoli.
b) Nechť $(X_1, Y_1), \dots, (X_n, Y_n)$ je náhodný výběr z dvourozměrného rozložení s kovariancí σ_{12} a koeficientem korelace ρ. Pak $E(S_{12}) = \sigma_{12}$, ať je hodnota parametru σ_{12} jakákoli, avšak $E(R_{12})$ je rovno ρ pouze přibližně (shoda je vyhovující pro $n \geq 30$), ať je hodnota parametru ρ jakákoli.

Nyní se budeme zabývat náhodným váběrem z normálního rozložení. Zavedeme několik statistik vzniklých transformací výběrového průměru a výběrového rozptylu (jsou to tzv. pivotové statistiky) a ukážeme, jakým způsobem se tyto statistiky řídí. V příští kapitole využijeme těchto pivotových statistik při konstrukci intervalů spolehlivosti pro parametry normálních rozložení. V této kapitole nám uvedené vlastnosti poslouží při výpočtu různých pravděpodobností.

11.4. Věta

Nechť $X_1, \dots, X_n$ je náhodný výběr z rozložení $N(\mu, \sigma^2)$. Pak platí

a) Výběrový průměr M a výběrový rozptyl S^2 jsou stochasticky nezávislé.

b) $M \sim N(\mu, \frac{\sigma^2}{n})$, tedy $U = \frac{M-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$. (Statistika U slouží ke konstrukci intervalu spolehlivosti pro μ, když σ^2 známe.)

c) $K = (n-1)S^2\sigma^2 \sim \chi^2(n-1)$. (Statistika K slouží ke konstrukci intervalu spolehlivosti pro σ^2, když μ neznáme.)

d) $\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \sim \chi^2(n)$. (Tato statistika, která nemá speciální označení, slouží ke konstrukci intervalu spolehlivosti pro σ^2, když μ známe.)

e) $T = \frac{M-\mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$. (Statistika T slouží ke konstrukci intervalu spolehlivosti pro μ, když σ^2 neznáme.)

11.5. Příklad

Hmotnost jedné porce kávy považujeme za náhodnou veličinu s normálním rozložením $X \sim N(7\,\mathrm{g}, 0{,}25\,\mathrm{g}^2)$. Jaká je pravděpodobnost, že k přípravě 28 porcí kávy postačí dva 100 g balíčky?

Řešení:

$X_1, \dots, X_{28}$ je náhodný výběr z $N(7, 0{,}25)$. Počítáme

$$P\left(\sum_{i=1}^{28} X_i \leq 200\right) = P\left(\frac{1}{28}\sum_{i=1}^{28} X_i \leq \frac{200}{28}\right) = P\left(M \leq \frac{200}{28}\right) =$$

$$= P\left(\frac{M-7}{\frac{0{,}5}{\sqrt{28}}} \leq \frac{\frac{200}{28}-1}{\frac{0{,}5}{\sqrt{28}}}\right) = P(U \leq 1{,}51) = \Phi(1{,}51) = 0{,}9345.$$

S pravděpodobností 93,45% můžeme předpokládat, že k přípravě 28 porcí kávy postačí dva 100 g balíčky.

11.6. Příklad

Odběratel provede kontrolu stejnorodosti dodávky výrobků tak, že změří sledovaný rozměr u 25 náhodně vybraných výrobků. Dodávku přijme, jestliže výběrová směrodatná odchylka se bude realizovat hodnotou menší nebo rovnou 0,2 mm. Je známo, že sledovaný rozměr výrobku má normální rozložení $N(50\,\mathrm{mm}, 0{,}2632\,\mathrm{mm}^2)$. Jaká je pravděpodobnost přijetí dodávky?

Řešení:
$X_1, \ldots, X_{25}$ je náhodný výběr z $N(50, 0{,}2632)$. Počítáme

$$P(S \leq 0{,}2) = P(S^2 \leq 0{,}04) = P \leq \left(\frac{(n-1)S^2}{\sigma^2} \leq \frac{(n-1)0{,}04}{\sigma^2} \right) =$$
$$= P\left(K \leq \frac{24 \cdot 0{,}04}{0{,}263^2} \right) = P(K \leq 13{,}879),$$

tedy číslo 13,879 je α-kvantil Pearsonova rozložení $\chi^2(24)$. V tabulkách kvantilů Pearsonova rozložení najdeme, že $\alpha = 0{,}05$. S pravděpodobností pouhých 5% lze očekávat, že odběratel přijme dodávku.

Přejdeme nyní ke dvěma nezávislým náhodným výběrům z normálního rozložení. I v této situaci nás zajímá rozložení pivotových statistik vzniklých transformací výběrových průměrů a výběrových rozptylů.

11.7. Věta
Nechť $X_{11}, \ldots, X_{n_1 1}$ je náhodný výběr z rozložení $N(\mu_1, \sigma_1^2)$ a $X_{12}, \ldots, X_{n_2 2}$ je na něm nezávislý náhodný výběr rozložení $N(\mu_2, \sigma_2^2)$, přičemž $n_1 \geq 2$ a $n_2 \geq 2$. Označme M_1, M_2 výběrové průměry a S_1^2, S_2^2 výběrové rozptyly. Pak platí:

a) Statistiky $M_1 - M_2$ (rozdíl výběrových průměrů) a

$$S_*^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

(vážený průměr výběrových rozptylů) jsou stochasticky nezávislé.

b) $M_1 - M_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$, tedy $U = \frac{(M_1-M_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$.

(Statistika U slouží ke konstrukci intervalu spolehlivosti pro rozdíl středních hodnot $\mu_1 - \mu_2$, když rozptyly σ_1^2, σ_2^2 známe.)

c) Jestliže $\sigma_1^2 = \sigma_2^2 = \sigma^2$, pak $K = \frac{(n_1+n_2-2)S_*^2}{\sigma^2} \sim \chi^2(n_1+n_2-2)$. (Statistika K slouží ke konstrukci intervalu spolehlivosti pro společný rozptyl σ^2, když střední hodnoty $\mu_1 - \mu_2$ neznáme.)

d) Jestliže $\sigma_1^2 = \sigma_2^2 = \sigma^2$, pak $T = \frac{(M_1-M_2)-(\mu_1-\mu_2)}{S_*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2)$.

e) $F = \frac{\frac{S_1^2}{S_2^2}}{\frac{\sigma_1^2}{\sigma_2^2}} \sim F(n_1-1, n_2-1)$. (Statistika F slouží ke konstrukci intervalu spolehlivosti pro podíl rozptylů $\frac{\sigma_1^2}{\sigma_2^2}$, když střední hodnoty μ_1, μ_2 neznáme.)

11.8. Příklad
Nechť jsou dány dva nezávislé náhodné výběry, první pochází z rozložení $N(2; 1{,}5)$ a má rozsah 10, druhý pochází z rozložení $N(3, 4)$ a má rozsah 5. Jaká je pravděpodobnost, že výběrový průměr 1. výběru bude menší než výběrový průměr 2. výběru?

Řešení:

$$P(M_1 < M_2) = P(M_1 - M_2 < 0) =$$

$$= P\left(\frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < \frac{0 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right) =$$

$$= P\left(U < \frac{-2+3}{\sqrt{\frac{1{,}5}{10} + \frac{4}{5}}}\right) = P(U < 1{,}05) = \Phi(1{,}05) = 0{,}85314.$$

S pravděpodobností 85,3% je výběrový průměr 1. výběru menší než výběrový průměr 2. výběru.

Shrnutí kapitoly

Ústředním pojmem matematické statistiky je pojem **náhodného výběru**, a to jednorozměrného i vícerozměrného. Transformací jednoho nebo více náhodných výběrů vzniká náhodná veličina zvaná **(výběrová) statistika**. K nejdůležitějším statistikám patří **výběrový průměr**, **výběrový rozptyl**, **výběrová směrodatná odchylka**, **výběrová kovariance**, **výběrový koeficient korelace**.

Jelikož statistika je náhodná veličina, má smysl počítat její střední hodnotu a rozptyl. Ukázali jsme si **vlastnosti střední hodnoty** a **rozptylu výběrového průměru** a **střední hodnoty výběrového rozptylu**, **výběrové kovariance** a **výběrového koeficientu korelace**.

Zabývali jsme se rovněž **rozložením výběrových statistik pro náhodné výběry z normálních rozložení**, tzv. pivotových statistik. Jak uvidíme v dalších kapitolách, lze pomocí těchto pivotových statistik konstruovat intervaly spolehlivosti pro parametry normálních rozložení a testovat hypotézy o těchto rozloženích.

Kontrolní otázky a úkoly

1 Kdy lze posloupnost náhodných veličin $X_1, \ldots, X_n$ považovat za náhodný výběr?

2 Uveďte nejdůležitější statistiky odvozené z náhodného výběru, který pochází a) z jednorozměrného rozložení, b) z dvourozměrného rozložení.

3 Jaký je vztah mezi výběrovým rozptylem a rozptylem v popisné statistice?

4 Nechť $X_1, \ldots, X_{10}$ je náhodný výběr z $N(100, 100)$. Jaké rozložení má výběrový průměr?

5 Předpokládáme. že velký ročník na vysoké škole má výsledky ze statistiky normálně rozloženy kolem střední hodnoty 72 bodů se směrodatnou odchylkou 9 bodů. Vypočtěte pravděpodobnost, že
   a) náhodně vybraný student bude mít výsledek nad 80 bodů
   b) průměr výsledků náhodně vybraných 10 studentů bude nad 80 bodů.

6 Nechť $X_1, \ldots, X_{20}$ je náhodný výběr z $N(\mu, \sigma^2)$. Najděte čísla k_1, k_2 tak, aby platilo $P(\frac{S^2}{\sigma^2} < k_1) = 0{,}05$ a $P(\frac{S^2}{\sigma^2} > k_2) = 0{,}05$.

12. Bodové a intervalové odhady parametrů a parametrických funkcí

Cíl kapitoly

Po prostudování této kapitoly budete umět:

- posoudit nestrannost a asymptotickou nestrannost bodových odhadů parametrické funkce a pomocí rozptylu ohodnotit jejich kvalitu
- sestrojit intervaly spolehlivosti pro parametry jednoho a dvou normálních rozložení
- stanovit rozsah náhodného výběru tak, aby šířka intervalu spolehlivosti nepřesáhla dané číslo

Časová zátěž

Pro zvládnutí této kapitoly budete potřebovat asi 8 hodin studia.

Jak jsme poznali v předešlé kapitole, náhodný výběr je posloupnost stochasticky nezávislých náhodných veličin se stejným rozložením. Každé rozložení závisí na nějakém parametru nebo i více parametrech. Např. alternativní rozložení závisí na parametru ν, exponenciální rozložení na parametru λ, normální rozložení na parametrech μ a σ^2 apod. Tyto parametry neznáme, známe jenom náhodný výběr. Ukážeme si, jak lze na základě znalosti náhodného výběru odhadnout neznámý parametr či jeho funkci, tzv. parametrickou funkci.

Je-li odhadem statistika, hovoříme o bodovém odhadu parametrické funkce. Existují různé typy bodových odhadů, nás budou zajímat odhady nestranné, asymptoticky nestranné a konzistentní.

Je-li odhadem interval, jehož meze jsou statistiky a který s dostatečně velkou pravděpodobností pokrývá neznámou hodnotu parametrické funkce, jedná se o interval spolehlivosti.

12.1. Motivace

Vycházíme z náhodného výběru $X_1, \ldots, X_n$ z rozložení $L(\nu)$, které závisí na parametru ν. Množinu všech přípustných hodnot tohoto parametru označíme Ξ. Parametr ν neznáme a chceme ho odhadnout pomocí daného náhodného výběru (případně chceme odhadnout nějakou parametrickou funkci $h(\nu)$).

Bodovým odhadem parametrické funkce $h(\nu)$ budeme rozumět statistiku $T_n = T(X_1, \ldots, X_n)$, která nabývá hodnot blízkých $h(\nu)$, ať je hodnota parametru ν jakákoliv. Existují různé metody, jak konstruovat bodové odhady (např. metoda momentů či metoda maximální věrohodnosti, ale těmi se zde zabývat nebudeme) a také různé typy bodových odhadů. Omezíme se na odhady nestranné a asymptoticky nestranné.

Intervalovým odhadem parametrické funkce $h(\nu)$ rozumíme interval (D, H), jehož meze jsou statistiky $D = D(X_1, \ldots, X_n)$, $H = H(X_1, \ldots, X_n)$ a který s dostatečně velkou pravděpodobností pokrývá $h(\nu)$, ať je hodnota parametru ν jakákoliv. Zaměříme se na intervalové odhady parametrů a parametrických funkcí normálního rozložení.

Bodový odhad parametrické funkce by měl mít určité vhodné vlastnosti. Takovou vlastností může být pro jeden odhad nestrannost a pro posloupnost odhadů asymptotická nestrannost či konzistence. Kvalitu nestranného bodového odhadu lze posoudit pomocí rozptylu tohoto odhadu: čím menší rozptyl, tím kvalitnější odhad.

12.2. Definice
Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozložení $L(\nu)$, $h(\nu)$ je parametrická funkce, T, T_1, T_2, ... jsou statistiky.

a) Řekneme, že statistika T je nestranným odhadem parametrické funkce $h(\nu)$, jestliže $\forall \vartheta \in \Xi : E(T) = h(\nu)$.
(Význam nestrannosti spočívá v tom, že odhad T nesmí parametrickou funkci $h(\nu)$ systematicky nadhodnocovat ani podhodnocovat. Není-li tato podmínka splněna, jde o vychýlený odhad.)

b) Jsou-li T_1, T_2 nestranné odhady téže parametrické funkce $h(\nu)$, pak řekneme, že T_1 je lepší odhad než T_2, jestliže $\forall \vartheta \in \Xi : D(T_1) < D(T_2)$.

c) Posloupnost se nazývá posloupnost asymptoticky nestranných odhadů parametrické funkce $h(\nu)$, jestliže $\forall \vartheta \in \Xi : \lim\limits_{n\to\infty} E(T_n) = h(\nu)$.
(Význam asymptotické nestrannosti spočívá v tom, že s rostoucím rozsahem výběru klesá vychýlení odhadu. Je zřejmé, že z nestrannosti okamžitě vyplývá asymptotická nestrannost.)

c) Posloupnost se nazývá posloupnost konzistentních odhadů parametrické funkce $h(\nu)$, jestliže $\forall \vartheta \in \Xi, \forall \varepsilon > 0 : \lim\limits_{n\to\infty} P(|T_n - h(\nu)| > \varepsilon) = 0$.
(Význam konzistence spočívá v tom, že s rostoucím rozsahem výběru klesá pravděpodobnost, že se odhad bude realizovat „daleko“ od skutečné hodnoty parametrické funkce. Lze ukázat, že z asymptotické nestrannosti vyplývá konzistence, pokud posloupnost rozptylů konverguje k 0.)

12.3. Příklad
Nezávisle opakovaná měření určité konstanty μ jsou charakterizována náhodným výběrem $X_1, \ldots, X_n$ z rozložení se střední hodnotou $E(X_i) = \mu$ a rozptylem $D(X_i) = \sigma^2$, $i = 1, \ldots, n$. Uvažme statistiky $M = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$ a $L = \frac{X_1+X_n}{2}$.

a) Dokažte, že M a L jsou nestranné odhady střední hodnoty μ.
b) Zjistěte, který z těchto dvou odhadů je lepší.

Řešení:
ad a)

$$E(M) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n} n\mu = \mu$$

$$E(L) = \left(\frac{X_1 + X_n}{2}\right) = \frac{1}{2}E(X_1 + X_n) = \frac{1}{2}[E(X_1) + E(X_n)] =$$
$$= \frac{1}{2}(\mu + \mu) = \mu$$

ad b)

$$D(M) = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{\sigma^2}n\sigma^2 = \frac{\sigma^2}{n}$$

$$D(L) = D\left(\frac{X_1 + X_n}{2}\right) = \frac{1}{4}D(X_1 + X_n) = \frac{1}{4}[D(X_1) + D(X_n)] =$$
$$= \frac{\sigma^2 + \sigma^2}{4} = \frac{\sigma^2}{2}$$

Vidíme tedy, že M je lepší odhad než L pro $n \geq 3$.

12.4. Poznámka
Ve větě 11.3, tvrzení (a), bylo uvedeno, že $E(S^2) = \sigma^2$, tedy výběrový rozptyl S^2 je nestranným odhadem rozptylu σ^2. (Odtud je také vidět, že ve vzorci pro výběrový rozptyl musí být konstanta $\frac{1}{n-1}$, nikoli $\frac{1}{n}$, aby platilo $E(S^2) = \sigma^2$.) Výběrová směrodatná odchylka S však není nestranným odhadem směrodatné odchylky σ. Pak by totiž platilo $E(S) = \sigma$, ovšem $E(S^2) = \sigma^2$, tedy $D(S) = E(S^2) - [E(S)]^2 = \sigma^2 - \sigma^2 = 0$, což je možné jen tak, že S by byla konstanta.

Nyní budeme definovat interval spolehlivosti pro parametrickou funkci, a to jak oboustranný, tak levostranný či pravostranný. Uvedeme doporučený postup při konstrukci intervalu spolehlivosti a ukážeme si, jaký vliv na šířku intervalu spolehlivosti má riziko a rozsah výběru.

12.5. Definice
Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozložení $L(\nu)$, $h(\nu)$ je parametrická funkce, $\alpha \in (0, 1)$, $D = D(X_1, \ldots, X_n)$, $H = H(X_1, \ldots, X_n)$ jsou statistiky.

a) Interval (D, H) se nazývá $100(1-\alpha)\%$ (oboustranný) interval spolehlivosti pro parametrickou funkci $h(\nu)$, jestliže:

$$\forall\vartheta \in \Xi: \; P(D < h(\nu) < H) \geq 1 - \alpha.$$

b) Interval (D, ∞) se nazývá $100(1-\alpha)\%$ levostranný interval spolehlivosti pro parametrickou funkci $h(\nu)$, jestliže:

$$\forall\vartheta \in \Xi: \; P(D < h(\nu)) \geq 1 - \alpha.$$

c) Interval $(-\infty, H)$ se nazývá $100(1-\alpha)\%$ pravostranný interval spolehlivosti pro parametrickou funkci $h(\nu)$, jestliže:

$$\forall\vartheta \in \Xi: \; P(h(\nu) < H) \geq 1 - \alpha.$$

d) Číslo α se nazývá riziko (zpravidla $\alpha = 0{,}05$, méně často 0,1 či 0,01), číslo $1 - \alpha$ se nazývá spolehlivost.

12.6. Poznámka

Doporučený postup při konstrukci intervalu spolehlivosti:

a) Vyjdeme ze statistiky V, která je nestranným bodovým odhadem parametrické funkce $h(\nu)$.

b) Najdeme tzv. pivotovou statistiku W, která vznikne transformací statistiky V, je monotónní funkcí $h(\nu)$ a přitom její rozložení je známé a na $h(\nu)$ nezávisí. (Při konstrukci intervalů spolehlivosti pro parametry jednoho a dvou normálních rozložení používáme jako pivotové statistiky statistiky M, K, T, F z vět 11.4 a 11.7.)

c) Pomocí známého rozložení pivotové statistiky W najdeme kvantily $w_{\alpha/2}$, $w_{1-\alpha/2}$, takže platí:

$$\forall \vartheta \in \Xi : \; P(w_{\alpha/2} < W < w_{1-\alpha/2}) \geq 1 - \alpha.$$

d) Nerovnost $w_{\alpha/2} < W < w_{1-\alpha/2}$ převedeme ekvivalentními úpravami na nerovnost $D < h(\nu) < H$.

e) Statistiky D, H nahradíme jejich číselnými realizacemi d, h a získáme tak $100(1-\alpha)\%$ empirický interval spolehlivosti, o němž prohlásíme, že pokrývá $h(\nu)$ s pravděpodobností aspoň $1-\alpha$. (Tvrzení, že (d,h) pokrývá $h(\nu)$ s pravděpodobností aspoň $1-\alpha$ je třeba chápat takto: jestliže mnohonásobně nezávisle získáme realizace $x_1, \ldots, x_n$ náhodného výběru $X_1, \ldots, X_n$ z rozložení $L(\nu)$ a pomocí každé této realizace sestrojíme $100(1-\alpha)\%$ empirický interval spolehlivosti pro $h(\nu)$, pak podíl počtu těch intervalů, které pokrývají $h(\nu)$ k počtu všech sestrojených intervalů bude přibližně $1-\alpha$.)

12.7. Věta

Nechť (d,h) je $100(1-\alpha)\%$ empirický interval spolehlivosti pro $h(\nu)$ zkonstruovaný pomocí číselných realizací $x_1, \ldots, x_n$ náhodného výběru $X_1, \ldots, X_n$ z rozložení $L(\nu)$.

a) Při konstantním riziku klesá šířka $h-d$ s rostoucím rozsahem náhodného výběru.

b) Při konstantním rozsahu náhodného výběru klesá šířka $h-d$ s rostoucím rizikem.

Nadále se budeme zabývat konstrukcí intervalů spolehlivosti pro parametry normálních rozložení. Vždy pro jednu konkrétní situaci podrobně odvodíme meze intervalu spolehlivosti a pro ostatní situace jen uvedeme přehled vzorců. Těm z vás, kteří mají hlubší zájem o statistiku, lze doporučit, abyste se pokusili uvedené vzorce odvodit a s využitím vlastností příslušných pivotových statistik, jak byly uvedeny ve větách 11.4 a 11.7.

12.8. Příklad

Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozložení $N(\mu, \sigma^2)$, přičemž $n \geq 2$ a parametry μ, σ^2 neznáme. Sestrojte $100(1-\alpha)\%$ interval spolehlivosti pro střední hodnotu μ a to

a) oboustranný,

b) levostranný,
c) pravostranný.

Řešení:
$h(\nu) = \mu$, $V = M = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$, $W = T = \frac{M-\mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$ (viz věta 11.4, tvrzení (e)), $w_{\alpha/2} = t_{\alpha/2}(n-1) = -t_{1-\alpha/2}(n-1)$, $w_{1-\alpha/2} = t_{1-\alpha/2}(n-1)$

ad a)

$$\forall\vartheta \in \Xi:\ 1-\alpha \le P(-t_{1-\alpha/2}(n-1) < T < t_{1-\alpha/2}(n-1)) =$$
$$= P\left(-t_{1-\alpha/2}(n-1) < \frac{M-\mu}{\frac{S}{\sqrt{n}}} < t_{1-\alpha/2}(n-1)\right) =$$
$$= P\left(M - \frac{S}{\sqrt{n}}t_{1-\alpha/2}(n-1) < \mu < M + \frac{S}{\sqrt{n}}t_{1-\alpha/2}(n-1)\right)$$

ad b)

$$\forall\vartheta \in \Xi:\ 1-\alpha \le P(T < t_{1-\alpha}(n-1)) =$$
$$= P\left(\frac{M-\mu}{\frac{S}{\sqrt{n}}} < t_{1-\alpha}(n-1)\right) = P\left(M - \frac{S}{\sqrt{n}}t_{1-\alpha}(n-1) < \mu\right)$$

ad c)

$$\forall\vartheta \in \Xi:\ 1-\alpha \le P(t_\alpha(n-1) < T) = P\left(t_\alpha(n-1) < \frac{M-\mu}{\frac{S}{\sqrt{n}}}\right) =$$
$$= P\left(\mu < M - \frac{S}{\sqrt{n}}t_\alpha(n-1)\right) = P\left(\mu < M + \frac{S}{\sqrt{n}}t_{1-\alpha}(n-1)\right)$$

Konkrétní aplikace: 10 krát nezávisle na sobě byla změřena jistá konstanta μ. Výsledky měření byly: 2; 1,8; 2,1; 2,4; 1,9; 2,1; 2; 1,8; 2,3; 2,2. Tyto výsledky považujeme za číselné realizace náhodného výběru $X_1, \dots, X_{10}$ z rozložení $N(\mu, \sigma^2)$, kde parametry μ, σ^2 neznáme. Najděte 95% empirický interval spolehlivosti pro μ, a to

a) oboustranný,
b) levostranný,
c) pravostranný.

Řešení:
$m = 2{,}06$, $s^2 = 0{,}0404$, $s = 0{,}2011$, $\alpha = 0{,}05$, $t_{0,975}(9) = 2{,}2622$, $t_{0,95}(9) = 1{,}8331$.

ad a) $d = m - \frac{s}{\sqrt{n}}t_{1-\alpha/2}(n-1) = 2{,}06 - \frac{0{,}2011}{\sqrt{10}}2{,}2622 = 1{,}92$
$h = m + \frac{s}{\sqrt{n}}t_{1-\alpha/2}(n-1) = 2{,}06 + \frac{0{,}2011}{\sqrt{10}}2{,}2622 = 2{,}20$
$1{,}92 < \mu < 2{,}20$ s pravděpodobností aspoň 0,95.

ad b) $d = m - \frac{s}{\sqrt{n}}t_{1-\alpha}(n-1) = 2{,}06 - \frac{0{,}2011}{\sqrt{10}}1{,}8331 = 1{,}94$
$1{,}94 < \mu$ s pravděpodobností aspoň 0,95.

ad c) $h = m + \frac{s}{\sqrt{n}}t_{1-\alpha}(n-1) = 2{,}06 + \frac{0{,}2011}{\sqrt{10}}1{,}8331 = 2{,}18$
$\mu < 2{,}18$ s pravděpodobností aspoň 0,95.

12.9. Věta

Přehled vzorců pro meze $100(1-\alpha)\%$ empirických intervalů spolehlivosti pro parametry jednoho normálního rozložení. Nechť $X_1,\dots,X_n$ je náhodný výběr z rozložení $N(\mu,\sigma^2)$, přičemž $n \geq 2$.

a) *Interval spolehlivosti pro μ, když σ^2 známe*

Oboustranný: $(d,h) = \left(m - \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2}, m + \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2}\right)$

Levostranný: $(d,\infty) = \left(m - \frac{\sigma}{\sqrt{n}}u_{1-\alpha}, \infty\right)$

Pravostranný: $(-\infty,h) = \left(-\infty, m + \frac{\sigma}{\sqrt{n}}u_{1-\alpha}\right)$

b) *Interval spolehlivosti pro μ, když σ^2 neznáme*

Oboustranný: $(d,h) = \left(m - \frac{s}{\sqrt{n}}t_{1-\alpha/2}(n-1), m + \frac{s}{\sqrt{n}}t_{1-\alpha/2}(n-1)\right)$

Levostranný: $(d,\infty) = \left(m - \frac{s}{\sqrt{n}}t_{1-\alpha}(n-1), \infty\right)$

Pravostranný: $(-\infty,h) = \left(-\infty, m + \frac{s}{\sqrt{n}}t_{1-\alpha}(n-1)\right)$

c) *Interval spolehlivosti pro σ^2, když μ neznáme*

Oboustranný: $(d,h) = \left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}, \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}\right)$

Levostranný: $(d,\infty) = \left(\frac{(n-1)s^2}{\chi^2_{1-\alpha}(n-1)}, \infty\right)$

Pravostranný: $(-\infty,h) = \left(-\infty, \frac{(n-1)s^2}{\chi^2_{\alpha}(n-1)}\right)$

d) *Interval spolehlivosti pro σ^2, když μ známe*

Oboustranný: $(d,h) = \left(\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{1-\alpha/2}(n)}, \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{\alpha/2}(n)}\right)$

Levostranný: $(d,\infty) = \left(\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{1-\alpha}(n)}, \infty\right)$

Pravostranný: $(-\infty,h) = \left(-\infty, \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{\alpha}(n)}\right)$

12.10. Příklad

Nechť $X_1,\dots,X_n$ je náhodný výběr z rozložení $N(\mu, 0{,}04)$. Jaký musí být minimální rozsah výběru, aby šířka 95% intervalu spolehlivosti pro μ nepřesáhla číslo 0,16?

Řešení:

Podle 12.9 (a) dostáváme:

$$0{,}16 \geq h - d = m + \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2} - m + \frac{\sigma}{\sqrt{n}}u_{1-\alpha/2} = 2\frac{\sigma}{\sqrt{n}}u_{1-\alpha/2} \Rightarrow$$

$$\Rightarrow n \geq \frac{4\sigma^2 u^2_{1-\alpha/2}}{0{,}16^2} = \frac{4\cdot 0{,}04\cdot 1{,}96^2}{0{,}16^2} = 24{,}01 \Rightarrow n \geq 25.$$

12.11. Příklad
Jsou dány dva nezávislé náhodné výběry o rozsazích $n_1 \geq 2$, $n_2 \geq 2$, první pochází z rozložení $N(\mu_1, \sigma^2)$, druhý z rozložení $N(\mu_2, \sigma^2)$, kde parametry μ_1, μ_2, σ^2 neznáme. Sestrojte $100(1-\alpha)\%$ interval spolehlivosti pro rozdíl středních hodnot $\mu_1 - \mu_2$.

Řešení:

$$h(\nu) = \mu,\ V = M_1 - M_2,\ W = T = \frac{(M_1 - M_2) - (\mu_2 - \mu_2)}{S_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

(viz věta 11.7, tvrzení (d)), $w_{\alpha/2} = t_{\alpha/2}(n_1 + n_2 - 2) = -t_{1-\alpha/2}(n_1 + n_2 - 2)$, $w_{1-\alpha/2} = t_{1-\alpha/2}(n_1 + n_2 - 2)$.

$$\begin{aligned}
&\forall \vartheta \in \Xi: \ 1 - \alpha \leq P(-t_{1-\alpha/2}(n_1 + n_2 - 2) < T < t_{1-\alpha/2}(n_1 + n_2 - 2)) = \\
&= P\left(-t_{1-\alpha/2}(n_1 + n_2 - 2) < \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{S_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < t_{1-\alpha/2}(n_1 + n_2 - 2)\right) = \\
&= P\left(M_1 - M_2 - S_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{1-\alpha/2}(n_1 + n_2 - 2) < \mu_1 - \mu_2 < \right. \\
&\qquad\qquad \left. < M_1 - M_2 + S_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{1-\alpha/2}(n_1 + n_2 - 2)\right)
\end{aligned}$$

Konkrétní aplikace: Ve dvou nádržích se zkoumal obsah chlóru (v g/l). Z první nádrže bylo odebráno 25 vzorků, z druhé nádrže 10 vzorků. Byly vypočteny realizace výběrových průměrů a rozptylů: $m_1 = 34{,}48$, $m_2 = 35{,}59$, $s_1^2 = 1{,}7482$, $s_2^2 = 1{,}7121$. Hodnoty zjištěné z odebraných vzorků považujeme za realizace dvou nezávislých náhodných výběrů z rozložení $N(\mu_1, \sigma^2)$ a $N(\mu_2, \sigma^2)$. Sestrojte 95% empirický interval spolehlivosti pro rozdíl středních hodnot $\mu_1 - \mu_2$.

Řešení:
$s_*^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{24 \cdot 1{,}7482 + 9 \cdot 1{,}7121}{33} = 1{,}7384$, $\ t_{0{,}975}(33) = 2{,}035$

$$\begin{aligned}
d &= m_1 - m_2 - s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{1-\alpha/2}(n_1 + n_2 - 2) = \\
&= 34{,}46 - 35{,}59 - \sqrt{1{,}7384} \sqrt{\frac{1}{25} + \frac{1}{10}} \cdot 2{,}035 = -2{,}114
\end{aligned}$$

$$\begin{aligned}
h &= m_1 - m_2 + s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{1-\alpha/2}(n_1 + n_2 - 2) = \\
&= 34{,}46 - 35{,}59 + \sqrt{1{,}7384} \sqrt{\frac{1}{25} + \frac{1}{10}} \cdot 2{,}035 = -0{,}106
\end{aligned}$$

$-2{,}114\,\mathrm{g/l} < \mu_1 - \mu_2 < -0{,}106\,\mathrm{g/l}$ s pravděpodobností aspoň 0,95.

12.12. Příklad
Jsou dány dva nezávislé náhodné výběry o rozsazích $n_1 \geq 2$, $n_2 \geq 2$, první pochází z rozložení $N(\mu_1, \sigma_1^2)$, druhý z rozložení $N(\mu_2, \sigma_2^2)$, kde parametry μ_1, μ_2, σ_1^2, σ_2^2 neznáme. Sestrojte $100(1-\alpha)\%$ interval spolehlivosti pro podíl rozptylů $\frac{\sigma_1^2}{\sigma_2^2}$.

Řešení:
$h(\nu) = \frac{\sigma_1^2}{\sigma_2^2}$, $V = \frac{S_1^2}{S_2^2}$, $W = F = \frac{\frac{S_1^2}{S_2^2}}{\frac{\sigma_1^2}{\sigma_2^2}} \sim F(n_1-1, n_2-1)$ (viz věta 11.7, tvrzení (e)), $w_{\alpha/2} = F_{\alpha/2}(n_1-1, n_2-1)$, $w_{1-\alpha/2} = F_{1-\alpha/2}(n_1-1, n_2-1)$.

$$\forall \vartheta \in \Xi: \ 1-\alpha \leq P(F_{\alpha/2}(n_1-1, n_2-1) < F < F_{1-\alpha/2}(n_1-1, n_2-1)) =$$
$$= P\left(F_{\alpha/2}(n_1-1, n_2-1) < \frac{\frac{S_1^2}{S_2^2}}{\frac{\sigma_1^2}{\sigma_2^2}} < F_{1-\alpha/2}(n_1-1, n_2-1)\right) =$$
$$= P\left(\frac{\frac{S_1^2}{S_2^2}}{F_{1-\alpha/2}(n_1-1, n_2-1)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{\frac{S_1^2}{S_2^2}}{F_{\alpha/2}(n_1-1, n_2-1)}\right)$$

Konkrétní aplikace: V předešlém příkladě nyní předpokládáme, že dané dva náhodné výběry pocházejí z rozložení $N(\mu_1, \sigma_1^2)$ a $N(\mu_2, \sigma_2^2)$. Sestrojte 95% empirický interval spolehlivosti pro podíl rozptylů.

Řešení:
$$d = \frac{\frac{S_1^2}{S_2^2}}{F_{1-\alpha/2}(n_1-1, n_2-1)} = \frac{\frac{1{,}7482}{1{,}7121}}{F_{0{,}975}(24,9)} = \frac{\frac{1{,}7482}{1{,}7121}}{3{,}6142} = 0{,}28$$
$$h = \frac{\frac{S_1^2}{S_2^2}}{F_{\alpha/2}(n_1-1, n_2-1)} = \frac{\frac{1{,}7482}{1{,}7121}}{F_{0{,}025}(24,9)} = \frac{\frac{1{,}7482}{1{,}7121}}{\frac{1}{F_{0{,}975}(9,24)}} =$$
$$= \frac{\frac{1{,}7482}{1{,}7121}}{\frac{1}{2{,}7027}} = 2{,}76$$

$0{,}28 < \frac{\sigma_1^2}{\sigma_2^2} < 2{,}76$ s pravděpodobností aspoň 0,95.

12.13. Věta
Přehled vzorců pro meze $100(1-\alpha)\%$ empirických intervalů spolehlivosti pro parametry dvou normálních rozložení. Nechť $X_{11}, \ldots, X_{n_1 1}$ je náhodný výběr z rozložení $N(\mu_1, \sigma_1^2)$ a $X_{12}, \ldots, X_{n_2 2}$ je na něm nezávislý náhodný výběr rozložení $N(\mu_2, \sigma_2^2)$, přičemž $n_1 \geq 2$ a $n_2 \geq 2$.

a) *Interval spolehlivosti pro $\mu_1 - \mu_2$, když σ_1^2, σ_2^2 známe*
Oboustranný:
$$(d,h) = \left(m_1 - m_2 - \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} u_{1-\alpha/2}, m_1 - m_2 - \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} u_{1-\alpha/2}\right)$$
Levostranný: $(d, \infty) = \left(m_1 - m_2 - \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} u_{1-\alpha}, \infty\right)$

Pravostranný: $(-\infty, h) = \left(-\infty, m_1 - m_2 - \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}u_{1-\alpha}\right)$

b) *Interval spolehlivosti pro* $\mu_1 - \mu_2$, *když* σ_1^2, σ_2^2 *neznáme, ale víme, že jsou shodné*

Oboustranný: $$\left(m_1 - m_2 - s_*\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}t_{1-\alpha/2}(n_1 + n_2 - 2),\right.$$
$$\left. m_1 - m_2 + s_*\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}t_{1-\alpha/2}(n_1 + n_2 - 2)\right)$$

Levostranný: $(d, \infty) = \left(m_1 - m_2 - s_*\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}t_{1-\alpha/2}(n_1 + n_2 - 2), \infty\right)$

Pravostranný:
$$(-\infty, h) = \left(-\infty, m_1 - m_2 + s_*\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}t_{1-\alpha/2}(n_1 + n_2 - 2)\right)$$

c) *Interval spolehlivosti pro společný neznámý rozptyl* σ^2

Oboustranný: $(d, h) = \left(\frac{(n_1+n_2-2)s_*^2}{\chi^2_{1-\alpha/2}(n_1+n_2-2)}, \frac{(n_1+n_2-2)s_*^2}{\chi^2_{\alpha/2}(n_1+n_2-2)}\right)$

Levostranný: $(d, \infty) = \left(\frac{(n_1+n_2-2)s_*^2}{\chi^2_{1-\alpha}(n_1+n_2-2)}, \infty\right)$

Pravostranný: $(-\infty, h) = \left(-\infty, \frac{(n_1+n_2-2)s_*^2}{\chi^2_{\alpha}(n_1+n_2-2)}\right)$

d) *Interval spolehlivosti pro podíl rozptylů* $\frac{\sigma_1^2}{\sigma_2^2}$

Oboustranný: $(d, h) = \left(\frac{\frac{s_1^2}{s_2^2}}{F_{1-\alpha/2}(n_1-1,n_2-1)}, \frac{\frac{s_1^2}{s_2^2}}{F_{\alpha/2}(n_1-1,n_2-1)}\right)$

Levostranný: $(d, \infty) = \left(\frac{\frac{s_1^2}{s_2^2}}{F_{1-\alpha}(n_1-1,n_2-1)}, \infty\right)$

Pravostranný: $(-\infty, h) = \left(-\infty, \frac{\frac{s_1^2}{s_2^2}}{F_{\alpha}(n_1-1,n_2-1)}\right)$

12.14. Poznámka

Není-li v bodě (b) věty 12.13 splněn předpoklad o shodě rozptylů, lze sestrojit aspoň přibližný $100(1-\alpha)\%$ interval spolehlivosti pro $\mu_1 - \mu_2$. V tomto případě má statistika T přibližně rozložení $t(\nu)$, kde počet stupňů volnosti

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}.$$

Není-li ν celé číslo, použijeme v tabulkách kvantilů Studentova rozložení lineární interpolaci.

Předpoklad o shodě rozptylů lze ověřit tak, že sestrojíme $100(1-\alpha)\%$ interval spolehlivosti pro $\frac{\sigma_1^2}{\sigma_2^2}$. Pokud tento interval bude obsahovat 1, lze s pravděpodobností $1-\alpha$ považovat rozptyly za shodné.

12.15. Věta
Nechť $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \dots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ je náhodný výběr z rozložení

$$N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right),$$

přičemž $n \geq 2$. Označíme $\mu = \mu_1 - \mu_2$ a zavedeme rozdílový náhodný výběr $Z_1 = X_1 - Y_1, \dots, Z_n = X_n - Y_n$. Nechť

$$M = \frac{1}{n}\sum_{i=1}^{n} Z_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - M)^2.$$

Pak statistika $T = \frac{M-\mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$, tudíž meze $100(1-\alpha)\%$ intervalu spolehlivosti pro μ jsou $M \pm \frac{S}{\sqrt{n}} t_{1-\alpha/2}(n-1)$.

12.16. Příklad
Bylo vybráno šest nových automobilů téže značky a po určité době bylo zjištěno, o kolik mm se sjely jejich pravé a levé přední pneumatiky.

číslo automobilu	1	2	3	4	5	6
pravá pneumatika se sjela o:	1,8	1,0	2,2	0,9	1,5	1,6
levá pneumatika se sjela o:	1,5	1,1	2,0	1,1	1,4	1,4

Za předpokladu, že naměřené dvojice hodnot představují číselné realizace náhodného výběru rozsahu 6 z dvourozměrného normálního rozložení

$$N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right),$$

sestrojte 95% empirický interval spolehlivosti pro rozdíl středních hodnot $\mu_1 - \mu_2$.

Řešení:
$z_1 = 0{,}3$, $z_2 = -0{,}1$, $z_3 = 0{,}2$, $z_4 = -0{,}2$, $z_5 = 0{,}1$, $z_6 = 0{,}2$, $m = 0{,}0833$, $s = 0{,}1941$, $\alpha = 0{,}05$.

$$\begin{aligned} d = m - \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1) = 0{,}0833 - \frac{0{,}1941}{\sqrt{6}} t_{0{,}975}(5) &= \\ = 0{,}0833 - \frac{0{,}1941}{\sqrt{6}} 2{,}5706 = -0{,}12 \end{aligned}$$

$$\begin{aligned} h = m + \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1) = 0{,}0833 + \frac{0{,}1941}{\sqrt{6}} t_{0{,}975}(5) &= \\ = 0{,}0833 + \frac{0{,}1941}{\sqrt{6}} 2{,}5706 = 0{,}29. \end{aligned}$$

$-0{,}12\,\text{mm} < \mu_1 - \mu_2 < 0{,}29\,\text{mm}$ s pravděpodobností aspoň 0,95.

Shrnutí kapitoly

Na základě znalosti náhodného výběru aproximujeme neznámou hodnotu parametru či parametrické funkce bodovým odhadem parametrické funkce. Zpravidla požadujeme, aby tento odhad měl jisté žádoucí vlastnosti. K těm patří nestrannost, resp. asymptotická nestrannost či konzistence, pokud pracujeme s posloupností bodových odhadů téže parametrické funkce.

Bodové odhady však mají jednu značnou nevýhodu – nevíme, s jakou pravděpodobností odhadují hodnotu neznámé parametrické funkce. Tuto nevýhodu odtraňují intervalové odhady parametrické funkce: jsou to intervaly, jejichž meze jsou statistiky a které s předem danou dostatečně velkou pravděpodobností pokrývají hodnotu neznámé parametrické funkce. Pokud do vzorců pro meze $100(1-\alpha)\%$ intervalu spolehlivosti pro danou parametrickou funkci dosadíme číselné realizace náhodného výběru, dostaneme $100(1-\alpha)\%$ empirický interval spolehlivosti.

V praxi se nejčastěji používají intervaly spolehlivosti pro parametry normálních rozložení. Proto jsme si uvedly předhled vzorců pro meze $100(1-\alpha)\%$ empirických intervalů spolehlivosti pro parametry jednoho a dvou normálních rozložení.

Kontrolní otázky a úkoly

1 Definujte nestranný odhad a asymptoticky nestranný odhad parametrické funkce. V čem spočívá význam nestrannosti a asymptotické nestrannosti?

2 (S) Přírůstky cen akcií na burze v New Yorku u 10 náhodně vybraných společností dosáhly těchto hodnot: 10, 16, 5, 10, 12, 8, 4, 6, 5, 4. Najděte nestranné bodové odhady střední hodnoty a rozptylu přírůstků cen akcií.

3 Nechť $X_1, \dots, X_n$ je náhodný výběr z rozložení $Rs(0,b)$, kde $b > 0$ je neznámý parametr. Jsou definovány statistiky $T_1 = X_1 + \frac{1}{2}X_2 + \frac{1}{3}X_3 + \frac{1}{6}X_4$ a $T_2 = \frac{1}{2}(X_1 + X_2 + X_3 + X_4)$. Ukažte, že T_1, T_2 jsou nestranné odhady parametru b a určete, který odhad je lepší.

4 Definujte $100(1-\alpha)\%$ interval spolehlivosti pro parametrickou funkci, a to jak oboustranný, tak jednostranné intervaly spolehlivosti.

5 Jaký vliv na šířku intervalu spolehlivosti má zvýšení rizika při konstantním rozsahu výběru?

6 Jaký vliv na šířku intervalu spolehlivosti má zvětšení rozsahu výběru při konstantním riziku?

7 Hloubka moře se měří přístrojem, jehož systematická chyba je nulová a náhodné chyby měření mají normální rozložení se směrodatnou odchylkou $\sigma = 1\,\mathrm{m}$. Kolik měření je nutno provést, aby se hloubka moře stanovila s chybou nejvýše $\pm 0{,}25\,\mathrm{m}$ při riziku 0,05?

8 U jistého měřicího zařízení má být posouzena jeho přesnost. Proto na něm byla nezávisle změřena délka téhož výrobku. Výsledky měření v cm byly: 15,15; 15,20; 15,04; 15,14; 15,22. Předpokládáme, že tyto výsledky jsou číselné realizace náhodného výběru rozsahu 5 z rozložení $N(\mu, \sigma^2)$. Sestrojte 95% empirický interval spolehlivosti pro rozptyl σ^2.

9 Sponzor televizních pořadů pro děti chce vědět, kolik času stráví děti sledováním televize, protože na těchto informacích závisí typy a počty programů. Náhodným výběrem 100 dětí se zjistilo, že sledování televize věnují týdně průměrně 27,5 h se směrodatnou odchylkou 8 h. Za předpokladu, že počet hodin strávený za týden sledováním televize se řídí normálním rozložením, sestrojte 95% empirický interval spolehlivosti pro střední hodnotu počtu hodin strávených týdně sledováním televize.

10 (S) Na jisté velké americké univerzitě bylo v r. 1969 náhodně vybráno 5 profesorů a nezávisle na tom 5 profesorek a byl zjištěn jejich roční příjem (v tisících dolarů). Muži: 16, 19, 12, 11, 22, ženy: 9, 12, 8, 10, 16. Předpokládáme, že uvedené údaje tvoří realizace dvou nezávislých náhodných výběrů z rozložení $N(\mu_1, \sigma_1^2)$ a $N(\mu_2, \sigma_2^2)$.

a) Sestrojte 95% empirický interval spolehlivosti pro podíl rozptylů příjmů mužů a žen.

b) Pokud bude uvedený interval spolehlivosti obsahovat 1, sestrojte 95% empirický interval spolehlivosti pro rozdíl středních hodnot příjmů mužů a žen. V opačném případě sestrojte aspoň přibližný interval spolehlivosti.

11 (S) Pět mužů se rozhodlo, že budou hubnout. Zjistili svou hmotnost před zahájením diety a po ukončení diety.

Číslo osoby	1	2	3	4	5
Hmotnost před dietou	84	77,5	91,5	84,5	97,5
Hmotnost po dietě	78,5	73,5	88,5	80	97

Za předpokladu, že uvedené údaje jsou číselné realizace náhodného výběru rozsahu 5 z dvourozměrného normálního rozložení

$$N_2\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{pmatrix}\right)$$

sestrojte 95% empirický interval spolehlivosti pro středních hodnotu úbytku hmotnosti.

13. Úvod do testování hypotéz a testy o parametrech normálního rozložení

Cíl kapitola

Po prostudování této kapitoly budete umět:

- formulovat nulovou a alternativní hypotézu
- stanovit testové kritérium a kritický obor pro test nulové hypotézy proti oboustranné alternativě i proti jednostranným alternativám
- posoudit sílu testu pomocí grafu silofunkce
- provádět testy hypotéz o parametrech normálního rozložení třemi různými způsoby

Časová zátěž

Pro zvládnutí této kapitoly budete potřebovat asi 8 hodin studia.

V této kapitole se budeme zabývat problémem, jak pomocí statistiky vzniklé transformací daného náhodného výběru rozhodnout, zda naše domněnka o parametru rozložení, z něhož náhodný výběr pochází, je správná. Například známe průměrnou hmotnost automaticky balených potravinářských výrobků určitého druhu zjištěnou před a po seřízení balícího automatu. S pravděpodobností 95% máme prokázat, že střední hodnota hmotnosti balíčků se seřízením automatu změnila. Statistické postupy, které řeší podobné problémy, se nazývají testy hypotéz.

Nejprve objasníme pojmy nulová hypotéza a alternativní hypotéza a vysvětlíme, kdy dojde k chybě 1. druhu či 2. druhu.

13.1. Motivace

Testování hypotéz patří k nejdůležitějším metodám matematické statistiky. Na základě znalosti náhodného výběru umožní s předem danou pravděpodobností ověřovat domněnky o parametrech rozložení, z něhož daný náhodný výběr pochází.

13.2. Definice

Nechť $X_1, \ldots, X_n$ je náhodný výběr z rozložení $L(\nu)$, kde parametr $\vartheta \in \Xi$ neznáme. Nechť $h(\nu)$ je parametrická funkce a c daná reálná konstanta. Tvrzení $H_0 : h(\nu) = c$ se nazývá *nulová hypotéza*, tvrzení $H_1 : h(\nu) \neq c$ se nazývá *oboustranná alternativní hypotéza*, tvrzení $H_1 : h(\nu) < c$ se nazývá *levostranná alternativní hypotéza*, tvrzení $H_1 : h(\nu) > c$ se nazývá *pravostranná alternativní hypotéza*. Testováním H_0 proti H_1 rozumíme rozhodovací postup založený na náhodném výběru $X_1, \ldots, X_n$, s jehož pomocí zamítneme či nezamítneme platnost nulové hypotézy.

13.3. Poznámka

Volba alternativní hypotézy není libovolná, ale vyplývá z konkrétní situace. Např. při současné technologii je pravděpodobnost vyrobení zmetku $\nu = 0{,}01$.

a) Po rekonstrukci výrobní linky byla obnovena výroba, přičemž technologie zůstala stejná. Chceme ověřit, zda se změnila kvalita výrobků. Testujeme $H_0 : \nu = 0{,}01$ proti $H_1 : \nu \neq 0{,}01$.

b) Byly provedeny změny v technologii výroby s cílem zvýšit kvalitu. V tomto případě tedy testujeme $H_0 : \nu = 0{,}01$ proti $H_1 : \nu < 0{,}01$.
c) Byly provedeny změny v technologii výroby s cílem snížit náklady. V této situaci testujeme $H_0 : \nu = 0{,}01$ proti $H_1 : \nu > 0{,}01$.

13.4. Definice
Při testování H_0 proti H_1 se můžeme dopustit jedné ze dvou chyb: chyba 1. druhu spočívá v tom, že H_0 zamítneme, ač ve skutečnosti platí a chyba 2. druhu spočívá v tom, že H_0 nezamítneme, ač ve skutečnosti neplatí. Situaci přehledně znázorňuje tabulka:

skutečnost	rozhodnutí	
	H_0 nezamítáme	H_0 zamítáme
H_0 platí	správné rozhodnutí	chyba 1. druhu
H_0 neplatí	chyba 2. druhu	správné rozhodnutí

Pravděpodobnost chyby 1. druhu se značí α a nazývá se hladina významnosti (většinou bývá $\alpha = 0{,}05$, méně často 0,1 či 0,01). Pravděpodobnost chyby 2. druhu se značí β. Číslo $1 - \beta$ se nazývá síla testu a vyjadřuje pravděpodobnost, s jakou test vypoví, že H_0 neplatí. Při daném rozsahu výběru vede snižování α ke růstu β a obráceně.

Nyní si ukážeme tři způsoby, jimiž lze provést test nulové hypotézy proti alternativní hypotéze. Klasický způsob spočívá v nalezení kritického oboru. Testování pomocí intervalu spolehlivosti navazuje na poznatky získané ve 12. kapitole. Moderní způsob založenýý na p-hodnotě je vhodný především tehdy, máme-li k dispozici statistický software. Všechny tři způsoby použijeme při řešení konkrétnho příkladu.

13.5. Poznámka
Testování H_0 proti H_1 na hladině významnosti α je možno provádět třemi různými způsoby:

a) pomocí kritického oboru
b) pomocí intervalu spolehlivosti
c) pomocí p-hodnoty.

ad a) Najdeme statistiku $T_0 = T_0(X_1, \ldots, X_n)$, kterou nazveme testovým kritériem. Množina hodnot, jichž může testové kritérium nabýt, se rozpadá na dva neslučitelné obory: obor nezamítnutí nulové hypotézy (značí se V) a obor zamítnutí nulové hypotézy (značí se W a nazývá se též *kritický obor*). Tyto dva obory jsou odděleny krytickými hodnotami (pro danou hladinu významnosti α je lze najít ve statistických tabulkách).

Jestliže číselná realizace t_0 testového krytéria T_0 padne do kritického oboru W, pak nulovou hypotézu zamítáme na hladině významnosti α a znamená to skutečné vyvrácení testované hypotézy. Jestliže t_0 padne do oboru nezamítnutí V, pak jde o pouhé mlčení, které platnost nulové hypotézy jenom připouští.

Pravděpodobnosti chyb 1. a 2. druhu nyní zapíšeme takto:

$$P(t_0 \in W | H_0 \text{ platí}) = \alpha, \qquad P(t_0 \in V | H_1 \text{ platí}) = \beta.$$

Stanovení kritického oboru pro danou hladinu významnosti α:
Označme $t_{\min}$ (resp. $t_{\max}$) nejmenší (resp. největší) hodnotu testového kritéria. Kritický obor v případě oboustranné alternativy má tvar

$$W = (t_{\min}, K_{\alpha/2}(T)\rangle \cup \langle K_{1-\alpha/2}(T), t_{\max}),$$

kde $K_{\alpha/2}(T)$ a $K_{1-\alpha/2}(T)$ jsou kvantily rozložení, jímž se řídí testové kritérium T_0, je-li testová hypotéza pravdivá. Kritický obor v případě levostranné alternativy má tvar:

$$W = (t_{\min}, K_{\alpha/2}(T)\rangle,$$

v případě pravdostranné alternativy má kritický obor tvar

$$W = \langle K_{1-\alpha/2}(T), t_{\max}).$$

ad b) Sestrojíme $100(1-\alpha)\%$ empirický interval spolehlivosti pro parametrickou funkci $h(\nu)$. Pokryje-li tento interval hodnotu c, pak H_0 nezamítáme na hladině významnosti α, v opačném případě H_0 zamítáme na hladině významnosti α.

Pro test H_0 proti oboustranné alternativě sestrojíme oboustranný interval spolehlivosti. Pro test H_0 proti levostranné alternativě sestrojíme pravostranný interval spolehlivosti. Pro test H_0 proti pravostranné alternativě sestrojíme levostranný interval spolehlivosti.

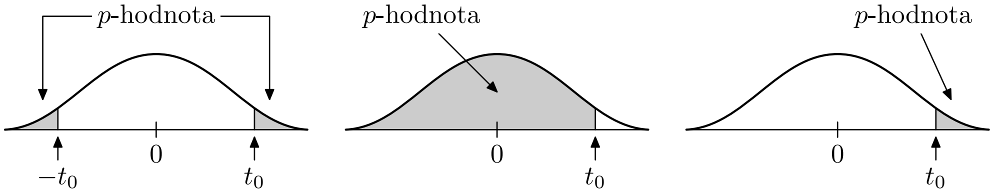
ad c) p-hodnota udává nejnižší možnou hladinu významnosti pro zamítnutí nulové hypotézy. Je-li p-hodnota $\leq \alpha$, pak H_0 zamítáme na hladině významnosti α, je-li p-hodnota $> \alpha$, pak H_0 nezamítáme na hladině významnosti α.

Způsob výpočtu p-hodnoty:
Pro oboustrannou alternativu: $p = 2\min\{P(T_0 \leq t_0), P(T_0 \geq t_0)\}$. Pro levostrannou alternativu: $p = P(T_0 \leq t_0)$, pro pravostrannou alternativu: $p = P(T_0 \geq t_0)$.

p-hodnota vyjadřuje pravděpodobnost, s jakou číselné realizace $x_1, \ldots, x_n$ náhodného výběru $X_1, \ldots, X_n$ podporují H_0, je-li pravdivá. Statistické programové systémy poskytují ve svých výstupech p-hodnotu. Její výpočet vyžaduje znalost distribuční funkce rozložení, kterým se řídí testové kritérium T_0, je-li H_0 pravdivá.

Vzhledem k tomu, že v běžných statistických tabulkách jsou uvedeny pouze hodnoty distribuční funkce standardizovaného normálního rozložení, bez použití speciálního software jsme schopni vypočítat p-hodnotu pouze pro test hypotézy o střední hodnotě normálního rozložení při známém rozptylu.

Ilustrace významu p-hodnoty pro test nulové hypotéza proti oboustranné, levostranné a pravostranné alternativě:



(Zvonovitá křivka reprezentuje hustotu rozložení, kterým se řídí testové kritérium, je-li nulová hypotéza pravdivá.)

13.6. Poznámka
Provádíme-li test nulové hypotézy proti alternativní hypotéze pomocí kritického oboru, doporučuje se dodržet následující postup:

1. Stanovíme nulovou hypotézu a alternativní hypotézu. Přitom je vhodné zvolit jako alternativní hypotézu ten předpoklad, jehož přijetí znamená závažné opatření a mělo by k němu dojít jen s malým rizikem omylu.
2. Zvolíme hladinu významnosti α. Zpravidla volíme $\alpha = 0{,}05$, méně často 0,1 nebo 0,01.
3. Najdeme vhodné testové kritérium a na základě zjištěných dat vypočítáme jeho realizaci.
4. Stanovíme kritický obor.
5. Jestliže realizace testového kritéria padla do kritického oboru, nulovou hypotézu zamítáme na hladině významnosti α. V opačném případě nulovou hypotézu nezamítáme na hladině významnosti α.

13.7. Příklad
10× nezávisle na sobě byla změřena jistá konstanta μ. Výsledky měření byly: 2; 1,8; 2,1; 2,4; 1,9; 2,1; 2; 1,8; 2,3; 2,2. Tyto výsledky považujeme za číselné realizace náhodného výběru $X_1, \dots, X_{10}$ z rozložení $N(\mu,\, 0{,}04)$. Nějaká teorie tvrdí, že $\mu = 1{,}95$. Proti nulové hypotéze $H_0 : \mu = 1{,}95$ postavíme oboustrannou alternativu $H_1 : \mu \neq 1{,}95$. Na hladině významnosti 0,05 testujte H_0 proti H_1.

Řešení:
$m = \frac{1}{10}(2 + \cdots + 2{,}2) = 2{,}06$, $\sigma^2 = 0{,}04$, $n = 10$, $\alpha = 0{,}05$, $c = 1{,}95$

a) Test provedeme pomocí kritického oboru.
Pro úlohy o střední hodnotě normálního rozložení při známém rozptylu používáme pivotovou statistiku $U = \frac{M-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$ (viz věta 11.4 (a)). Testové kritérium tedy bude $T_0 = \frac{M-c}{\frac{\sigma}{\sqrt{n}}}$ a bude mít rozložení $N(0,1)$, pokud je H_0 pravdivá. Vypočítáme realizaci testového kritéria: $t_0 = \frac{2{,}06-1{,}95}{\frac{0{,}2}{\sqrt{10}}} = 1{,}74$. Stanovíme kritický obor:

$$\begin{aligned} W &= (t_{\min}, K_{\alpha/2}(T)\rangle \cup \langle K_{1-\alpha/2}(T), t_{\max}) = (-\infty, u_{\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty) = \\ &= (-\infty, -u_{1-\alpha/2}\rangle \cup \langle u_{1-\alpha/2}, \infty) = (-\infty, -u_{0{,}975}\rangle \cup \langle u_{0{,}975}, \infty) = \\ &= (-\infty, -1{,}96\rangle \cup \langle 1{,}96, \infty) \end{aligned}$$

Protože $1{,}74 \notin W$, H_0 nezamítáme na hladině významnosti 0,05.

b) Test provedeme pomocí intervalu spolehlivosti.
Meze $100(1-\alpha)\%$ intervalu spolehlivosti pro střední hodnotu μ při známém rozptylu σ^2 jsou (viz věta 12.9 (a)): $(d,h) = \left(m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}, m + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}\right)$.
V našem případě $d = 2{,}06 - \frac{0{,}2}{\sqrt{10}} u_{0{,}975} = 2{,}06 - \frac{0{,}2}{\sqrt{10}} 1{,}96 = 1{,}936$, $h = 2{,}184$.
Protože $1{,}95 \in (1{,}936; 2{,}184)$, H_0 nezamítáme na hladině významnosti 0,05.

c) Test provedeme pomocí p-hodnoty.
Protože proti nulové hypotéze stavíme oboustrannou alternativu, použijeme vzorec

$$p = 2\min\{P(T_0 \le t_0), P(T_0 \ge t_0)\} = 2\min\{P(T_0 \le 1{,}74), P(T_0 \ge 1{,}74)\} = \\ = 2\min\{\Phi(1{,}74), 1-\Phi(1{,}74)\} = 2\min\{0{,}95907, 1-0{,}95907\} = 0{,}08186$$

Jelikož $0{,}08186 > 0{,}05$, H_0 nezamítáme na hladině významnosti 0,05.

Nadále se budeme zabývat tastováním hypotéz o parametrech normálního rozložení. Ukážeme si různé typy testů a naučíme se je provádět pomocí kritického oboru.

13.8. Definice

a) Nechť $X_1, \dots, X_n$ je náhodný výběr $N(\mu, \sigma^2)$, kde σ^2 známe. Nechť $n \ge 2$ a c je konstanta. Test $H_0: \mu = c$ proti $H_1: \mu \ne c$ se nazývá *z-test*.

b) Nechť $X_1, \dots, X_n$ je náhodný výběr $N(\mu, \sigma^2)$, kde σ^2 neznáme. Nechť $n \ge 2$ a c je konstanta. Test $H_0: \mu = c$ proti $H_1: \mu \ne c$ se nazývá *jednovýběrový t-test*.

c) Nechť $X_{11}, \dots, X_{n_1 1}$ je náhodný výběr z rozložení $N(\mu_1, \sigma^2)$ a $X_{12}, \dots, X_{n_2 2}$ je na něm nezávislý náhodný výběr rozložení $N(\mu_2, \sigma^2)$, přičemž $n_1 \ge 2$ a $n_2 \ge 2$ a σ^2 neznáme. Nechť c je konstanta. Test $H_0: \mu_1 - \mu_2 = c$ proti $H_1: \mu_1 - \mu_2 \ne c$ se nazývá *dvouvýběrový t-test*.

d) Nechť $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \dots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ je náhodný výběr z rozložení

$$N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right),$$

přičemž $n \ge 2$ a žádný parametr neznáme. Nechť c je konstanta. Test $H_0: \mu_1 - \mu_2 = c$ proti $H_1: \mu_1 - \mu_2 \ne c$ se nazývá *párový t-test*.

e) Nechť $X_{11}, \dots, X_{n_1 1}$ je náhodný výběr z rozložení $N(\mu_1, \sigma_1^2)$ a $X_{12}, \dots, X_{n_2 2}$ je na něm nezávislý náhodný výběr rozložení $N(\mu_2, \sigma_2^2)$, přičemž $n_1 \ge 2$ a $n_2 \ge 2$. Test $H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$ proti $H_1: \frac{\sigma_1^2}{\sigma_2^2} \ne 1$ se nazývá *F-test*.

f) Nechť $X_1, \dots, X_n$ je náhodný výběr $N(\mu, \sigma^2)$, kde μ neznáme. Nechť $n \ge 2$ a c je konstanta. Test $H_0: \sigma^2 = c$ proti $H_1: \sigma^2 \ne c$ se nazývá *test o rozptylu*.

13.9. Věta
Návody na provedení výše popsaných šesti typů testů pomocí kritického oboru.

a) Provedení z-testu
Hypotézu $H_0: \mu = c$ proti $H_1: \mu \neq c$ (resp. $H_1: \mu < c$ resp. $H_1: \mu > c$) zamítáme na hladině významnosti α, jestliže $\left|\frac{m-c}{\frac{\sigma}{\sqrt{n}}}\right| \geq u_{1-\alpha/2}$ (resp. $\frac{m-c}{\frac{\sigma}{\sqrt{n}}} \leq u_{1-\alpha}$ resp. $\frac{m-c}{\frac{\sigma}{\sqrt{n}}} \geq u_{1-\alpha}$).

b) Provedení jednovýběrového t-testu
Hypotézu $H_0: \mu = c$ proti $H_1: \mu \neq c$ (resp. $H_1: \mu < c$ resp. $H_1: \mu > c$) zamítáme na hladině významnosti α, jestliže $\left|\frac{m-c}{\frac{s}{\sqrt{n}}}\right| \geq t_{1-\alpha/2}(n-1)$ (resp. $\frac{m-c}{\frac{s}{\sqrt{n}}} \leq t_{1-\alpha}(n-1)$ resp. $\frac{m-c}{\frac{s}{\sqrt{n}}} \geq t_{1-\alpha}(n-1)$).

c) Provedení dvouvýběrového t-testu
Hypotézu $H_0: \mu_1 - \mu_2 = c$ proti $H_1: \mu_1 - \mu_2 \neq c$ (resp. $H_1: \mu_1 - \mu_2 < c$ resp. $H_1: \mu_1 - \mu_2 > c$) zamítáme na hladině významnosti α, jestliže

$$\left|\frac{m_1 - m_2 - c}{s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right| \geq t_{1-\alpha/2}(n_1 + n_2 - 2)$$

(resp. $\frac{m_1-m_2-c}{s_*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \leq t_{1-\alpha}(n_1+n_2-2)$ resp. $\frac{m_1-m_2-c}{s_*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \geq t_{1-\alpha}(n_1+n_2-2)$).

d) Provedení párového t-testu
Od náhodného výběru $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$ z dvourozměrného normálního rozložení přejdeme k rozdílovému náhodnému výběru $Z_1 = X_1 - Y_1, \ldots, Z_n = X_n - Y_n$. Označíme $\mu = \mu_1 - \mu_2$. Pak jde o test hypotézy $H_0: \mu = c$ proti $H_1: \mu \neq c$ a úloha je převedna na jednováběrový t-test.

e) Provedení F-testu
Hypotézu $H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$ proti $H_1: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$ (resp. $H_1: \frac{\sigma_1^2}{\sigma_2^2} < 1$ resp. $H_1: \frac{\sigma_1^2}{\sigma_2^2} > 1$) zamítáme na hladině významnosti α, jestliže

$$\frac{s_1^2}{s_2^2} \leq F_{\alpha/2}(n_1 + n_2 - 2) \qquad \text{nebo} \qquad \frac{s_1^2}{s_2^2} \geq F_{1-\alpha/2}(n_1 + n_2 - 2)$$

(resp. $\frac{s_1^2}{s_2^2} \leq F_{\alpha}(n_1 + n_2 - 2)$ resp. $\frac{s_1^2}{s_2^2} \geq F_{1-\alpha}(n_1 + n_2 - 2)$).

f) Provedení testu o rozptylu
Hypotézu $H_0: \sigma^2 = c$ proti $H_1: \sigma^2 \neq c$ (resp. $H_1: \sigma^2 < c$ resp. $H_1: \sigma^2 > c$) zamítáme na hladině významnosti α, jestliže

$$\frac{(n-1)s^2}{c} \leq \chi^2_{\alpha/2}(n-1) \qquad \text{nebo} \qquad \frac{(n-1)s^2}{c} \geq \chi^2_{1-\alpha/2}(n-1)$$

(resp. $\frac{(n-1)s^2}{c} \leq \chi^2_{\alpha}(n-1)$ resp. $\frac{(n-1)s^2}{c} \geq \chi^2_{1-\alpha}(n-1)$).

13.10. Příklad
Je-li u automatického obráběcího stroje rozptyl délky obráběných součístek větší než $380\,\mu\mathrm{m}^2$, je třeba stroj znova nastavit. Náhodně jsme vybrali 15 součástek a změřili jejich délku. Výběrový rozptyl zjištěných 15-ti délek činil $680\,\mu\mathrm{m}^2$. Za předpokladu, že délky se řídí normálním rozložením testujte na hladině významnosti 0,05 hypotézu, že stroj je třeba znova nastavit.

Řešení:
$X_1, \dots, X_{15}$ je náhodný výběr z rozložení $N(\mu, \sigma^2)$, přičemž $s^2 = 680\,\mu m^2$. Testujeme $H_0 : \sigma^2 = 380\,\mu\mathrm{m}^2$ proti pravostranné alternativě, která má tvar $H_1 : \sigma^2 > 380\,\mu\mathrm{m}^2$, na hladině významnosti 0,05.

Podle bodu (f) věty 13.9 dostáváme: realizace testového kritéria

$$\frac{(n-1)s^2}{c} = \frac{14 \cdot 680}{380} = 25{,}05.$$

Přitom $\chi^2_{1-\alpha}(n-1) = \chi^2_{0{,}95}(14) = 23{,}685$. Protože $25{,}05 \geq 23{,}685$, H_0 zamítáme na hladině významnosti 0,05. Zjištěná data nás tedy opravňují k tomu, abycho stroj znovu seřídili (s rizikem 5%, že budeme provádět zbytečnou práci).

Shrnutí kapitoly

Tvrzení o parametrech rozložení, z něhož pochází daný náhodný výběr, nazýváme **nulovou hypotézou**. Proti nulové hypotéze stavíme **alternativní hypotézu**, která říká, co platí, když neplatí nulová hypotéza. Při testování nulové hypotézy proti alternativní hypotéze se můžeme dopustit buď **chyby 1. druhu** (nulovou hypotézu zamítneme, ač ve skutečnosti platí) nebo **chyby 2. druhu** (nulovou hypotézu nezamítneme, ač ve skutečnosti neplatí). Pravděpodobnost chyby 1. druhu se značí α a nazývá se hladina **významnosti testu**.

Klasický přístup k testování hypotéz spočívá v nalezení vhodného **testového kritéria**. Množina hodnot, jichž může testové kritérium nabýt, se rozpadá na **obor nezamítnutí nulové hypotézy** a na **kritický obor**. Tyto dva neslučitelné obory jsou odděleny **kritickými hodnotami**. Pokud se testové kritérium realizuje v kritickém oboru, nulovou hypotézu zamítáme na hladině významnosti α a přijímáme alternativní hypotézu. V opačném případě nulovou hypotézu nezamítáme na hladině významnosti α. Tím jsme ovšem neprokázali její pravdivost, můžeme pouze říci, že naše data nejsou natolik průkazná, abychom mohli nulovou hypotézu zamítnout.

Test nulové hypotézy proti alternativní hypotéze lze též provést pomocí intervalu spolehlivosti a s využitím metod popsaných ve 12. kapitole.

Máme-li k dispozici statistický software, můžeme vypočítat ***p*-hodnotu** jako nejmenší možnou hladinu významnosti pro zamítnutí nulové hypotézy.

V praxi se nejčastěji setkáváme s **testy hypotéz o parametrech normálního rozložení**. K těmto testům patří například z-test, jednovýběrový, párový či dvouvýběrový t-test apod.

Kontrolní otázky a úkoly

1 Vysvětlete pojem „nulová hypotéza“ a „alternativní hypotéza“.

2 V čem spočívá testování nulové hypotézy proti alternativní hypotéze?

3 Kdy se dopustíme chyby 1. druhu (2. druhu)?

4 Co rozumíme testovým kritériem a kritickým oborem?

5 Popište tři způsoby testování hypotéz.

6 Jaké znáte testy o parametrech normálního rozložení?

7 Podle údajů na obalu čokolády by její čistá hmotnost měla být 125 g. Výrobce dostal několik stížností od kupujících, ve kterých tvrdili, že hmotnost čokolád je nižší než deklarovaných 125 g. Z tohoto důvodu oddělení kontroly náhodně vybralo 50 čokolád a zjistilo, že jejich průměrná hmotnost je 122 g a směrodatná odchylka 8,6 g. Za předpokladu, že hmotnost čokolád se řídí normálním rozložením, můžeme na hladině významnosti 0,01 považovat stížnosti kupujících za oprávněné?

8 (S) V restauraci „U bílého koníčka“ měřili ve 20 případech čas obsluhy zákazníka. Výsledky v minutách: 6, 8, 11, 4, 7, 6, 10, 6, 9, 8, 5, 12, 13, 10, 9, 8, 7, 11, 10, 5. V restauraci „Zlatý lev“ bylo dané pozorování uskutečněno v 15 případech s těmito výsledky: 9, 11, 10, 7, 6, 4, 8, 13, 5, 15, 8, 5, 6, 8, 7. Na hladině významnosti 0,05 testujte hypotézu, že střední hodnoty doby obsluhy jsou v obou restauracích stejné.

9 (S) Na 10 automobilech stejného typu se testovaly dva druhy benzínu lišící se oktanovým číslem. U každého automobilu se při průměrné rychlosti 90 km/h měřil dojezd (tj. dráha, kterou ujede na dané množství benzínu) při použití každého z obou druhů benzínu. Výsledky:

č.auta	1	2	3	4	5	6	7	8	9	10
benzín A	17,5	20,0	18,9	17,9	16,4	18,9	17,2	17,5	18,5	18,2
benzín B	17,8	20,8	19,5	18,3	16,6	19,5	17,5	17,9	19,1	18,6

Za předpokladu, že dojezd se řídí normálním rozložením, testujte na hladině významnosti 0,05 hypotézu, že rozdíl středních hodnot dojezdu při dvou druzích benzínu se neliší.

10 Pevnost vlákna bavlněné příze lze pokládat za náhodnou veličinu s rozložením $N(\mu, \sigma^2)$. Je-li $\sigma^2 > 0{,}36\,\mathrm{kg}^2$, vznikají potíže při tkaní. Při zkoušce 11 náhodně vybraných vláken byly zjištěny hodnoty jejich pevnosti a vypočten empirický rozptyl $s^2 = 0{,}92\,\mathrm{kg}^2$. Na hladině významnosti 0,05 je třeba zjistit, zda je příze vyhovující.

11 Normálně rozložená náhodné veličiny představují výsledek měření téže konstanty dvěma různými metodami a jejich neznámé směrodatné odchylky σ_1, σ_2 charakterizují nespolehlivost těchto metod způsobenou náhodnými chybami. Při realizaci dvou nezávislých náhodných výběrů rozsahu $n_1 = 25$, $n_2 = 31$ jsme získali empirické směrodatné odchylky $s_1 = 0{,}523$, $s_2 = 0{,}363$. Je možno na hladině významnosti 0,05 považovat obě metody za stejně spolehlivé?

Příloha A – Statistické tabulky

Distribuční funkce standardizovaného normálního rozložení

u	$\Phi(u)$	u	$\Phi(u)$	u	$\Phi(u)$	u	$\Phi(u)$
0,00	0,50000	0,50	0,69146	1,00	0,84134	1,50	0,93319
0,01	0,50399	0,51	0,69497	1,01	0,84375	1,51	0,93448
0,02	0,50798	0,52	0,69847	1,02	0,84614	1,52	0,93574
0,03	0,51197	0,53	0,70194	1,03	0,84850	1,53	0,93699
0,04	0,51595	0,54	0,70540	1,04	0,85083	1,54	0,93822
0,05	0,51994	0,55	0,70884	1,05	0,85314	1,55	0,93943
0,06	0,52392	0,56	0,71226	1,06	0,85543	1,56	0,94062
0,07	0,52790	0,57	0,71566	1,07	0,85769	1,57	0,94179
0,08	0,53188	0,58	0,71904	1,08	0,85993	1,58	0,94295
0,09	0,53586	0,59	0,72240	1,09	0,86214	1,59	0,94408
0,10	0,53983	0,60	0,72575	1,10	0,86433	1,60	0,94520
0,11	0,54380	0,61	0,72907	1,11	0,86650	1,61	0,94630
0,12	0,54776	0,62	0,73237	1,12	0,86864	1,62	0,94738
0,13	0,55172	0,63	0,73565	1,13	0,87076	1,63	0,94845
0,14	0,55567	0,64	0,73891	1,14	0,87286	1,64	0,94950
0,15	0,55962	0,65	0,74215	1,15	0,87493	1,65	0,95053
0,16	0,56356	0,66	0,74537	1,16	0,87698	1,66	0,95154
0,17	0,56749	0,67	0,74857	1,17	0,87900	1,67	0,95254
0,18	0,57142	0,68	0,75175	1,18	0,88100	1,68	0,95352
0,19	0,57535	0,69	0,75490	1,19	0,88298	1,69	0,95449
0,20	0,57926	0,70	0,75804	1,20	0,88493	1,70	0,95543
0,21	0,58317	0,71	0,76115	1,21	0,88686	1,71	0,95637
0,22	0,58706	0,72	0,76424	1,22	0,88877	1,72	0,95728
0,23	0,59095	0,73	0,76730	1,23	0,89065	1,73	0,95818
0,24	0,59483	0,74	0,77035	1,24	0,89251	1,74	0,95907
0,25	0,59871	0,75	0,77337	1,25	0,89435	1,75	0,95994
0,26	0,60257	0,76	0,77637	1,26	0,89617	1,76	0,96080
0,27	0,60642	0,77	0,77935	1,27	0,89796	1,77	0,96164
0,28	0,61026	0,78	0,78230	1,28	0,89973	1,78	0,96246
0,29	0,61409	0,79	0,78524	1,29	0,90147	1,79	0,96327
0,30	0,61791	0,80	0,78814	1,30	0,90320	1,80	0,96407
0,31	0,62172	0,81	0,79103	1,31	0,90490	1,81	0,96485
0,32	0,62552	0,82	0,79389	1,32	0,90658	1,82	0,96562
0,33	0,62930	0,83	0,79673	1,33	0,90824	1,83	0,96638
0,34	0,63307	0,84	0,79955	1,34	0,90988	1,84	0,96712
0,35	0,63683	0,85	0,80234	1,35	0,91149	1,85	0,96784
0,36	0,64058	0,86	0,80511	1,36	0,91309	1,86	0,96856
0,37	0,64431	0,87	0,80785	1,37	0,91466	1,87	0,96926
0,38	0,64803	0,88	0,81057	1,38	0,91621	1,88	0,96995
0,39	0,65173	0,89	0,81327	1,39	0,91774	1,89	0,97062
0,40	0,65542	0,90	0,81594	1,40	0,91924	1,90	0,97128
0,41	0,65910	0,91	0,81859	1,41	0,92073	1,91	0,97193
0,42	0,66276	0,92	0,82121	1,42	0,92220	1,92	0,97257
0,43	0,66640	0,93	0,82381	1,43	0,92364	1,93	0,97320
0,44	0,67003	0,94	0,82639	1,44	0,92507	1,94	0,97381
0,45	0,67364	0,95	0,82894	1,45	0,92647	1,95	0,97441
0,46	0,67724	0,96	0,83147	1,46	0,92785	1,96	0,97500
0,47	0,68082	0,97	0,83398	1,47	0,92922	1,97	0,97558
0,48	0,68439	0,98	0,83646	1,48	0,93056	1,98	0,97615
0,49	0,68793	0,99	0,83891	1,49	0,93189	1,99	0,97670

$$\Phi(-u) = 1 - \Phi(u)$$

Distribuční funkce standardizovaného normálního rozložení

u	$\Phi(u)$	u	$\Phi(u)$	u	$\Phi(u)$	u	$\Phi(u)$
2,00	0,97725	2,50	0,99379	3,00	0,99865	3,50	0,99977
2,01	0,97778	2,51	0,99396	3,01	0,99869	3,51	0,99978
2,02	0,97831	2,52	0,99413	3,02	0,99874	3,52	0,99978
2,03	0,97882	2,53	0,99430	3,03	0,99878	3,53	0,99979
2,04	0,97932	2,54	0,99446	3,04	0,99882	3,54	0,99980
2,05	0,97982	2,55	0,99461	3,05	0,99886	3,55	0,99981
2,06	0,98030	2,56	0,99477	3,06	0,99889	3,56	0,99981
2,07	0,98077	2,57	0,99492	3,07	0,99893	3,57	0,99982
2,08	0,98124	2,58	0,99506	3,08	0,99897	3,58	0,99983
2,09	0,98169	2,59	0,99520	3,09	0,99900	3,59	0,99983
2,10	0,98214	2,60	0,99534	3,10	0,99903	3,60	0,99984
2,11	0,98257	2,61	0,99547	3,11	0,99906	3,61	0,99985
2,12	0,98300	2,62	0,99560	3,12	0,99910	3,62	0,99985
2,13	0,98341	2,63	0,99573	3,13	0,99913	3,63	0,99986
2,14	0,98382	2,64	0,99585	3,14	0,99916	3,64	0,99986
2,15	0,98422	2,65	0,99598	3,15	0,99918	3,65	0,99987
2,16	0,98461	2,66	0,99609	3,16	0,99921	3,66	0,99987
2,17	0,98500	2,67	0,99621	3,17	0,99924	3,67	0,99988
2,18	0,98537	2,68	0,99632	3,18	0,99926	3,68	0,99988
2,19	0,98574	2,69	0,99643	3,19	0,99929	3,69	0,99989
2,20	0,98610	2,70	0,99653	3,20	0,99931	3,70	0,99989
2,21	0,98645	2,71	0,99664	3,21	0,99934	3,71	0,99990
2,22	0,98679	2,72	0,99674	3,22	0,99936	3,72	0,99990
2,23	0,98713	2,73	0,99683	3,23	0,99938	3,73	0,99990
2,24	0,98745	2,74	0,99693	3,24	0,99940	3,74	0,99991
2,25	0,98778	2,75	0,99702	3,25	0,99942	3,75	0,99991
2,26	0,98809	2,76	0,99711	3,26	0,99944	3,76	0,99992
2,27	0,98840	2,77	0,99720	3,27	0,99946	3,77	0,99992
2,28	0,98870	2,78	0,99728	3,28	0,99948	3,78	0,99992
2,29	0,98899	2,79	0,99736	3,29	0,99950	3,79	0,99992
2,30	0,98928	2,80	0,99744	3,30	0,99952	3,80	0,99993
2,31	0,98956	2,81	0,99752	3,31	0,99953	3,81	0,99993
2,32	0,98983	2,82	0,99760	3,32	0,99955	3,82	0,99993
2,33	0,99010	2,83	0,99767	3,33	0,99957	3,83	0,99994
2,34	0,99036	2,84	0,99774	3,34	0,99958	3,84	0,99994
2,35	0,99061	2,85	0,99781	3,35	0,99960	3,85	0,99994
2,36	0,99086	2,86	0,99788	3,36	0,99961	3,86	0,99994
2,37	0,99111	2,87	0,99795	3,37	0,99962	3,87	0,99995
2,38	0,99134	2,88	0,99801	3,38	0,99964	3,88	0,99995
2,39	0,99158	2,89	0,99807	3,39	0,99965	3,89	0,99995
2,40	0,99180	2,90	0,99813	3,40	0,99966	3,90	0,99995
2,41	0,99202	2,91	0,99819	3,41	0,99968	3,91	0,99995
2,42	0,99224	2,92	0,99825	3,42	0,99969	3,92	0,99996
2,43	0,99245	2,93	0,99831	3,43	0,99970	3,93	0,99996
2,44	0,99266	2,94	0,99836	3,44	0,99971	3,94	0,99996
2,45	0,99286	2,95	0,99841	3,45	0,99972	3,95	0,99996
2,46	0,99305	2,96	0,99846	3,46	0,99973	3,96	0,99996
2,47	0,99324	2,97	0,99851	3,47	0,99974	3,97	0,99996
2,48	0,99343	2,98	0,99856	3,48	0,99975	3,98	0,99997
2,49	0,99361	2,99	0,99861	3,49	0,99976	3,99	0,99997

$$\Phi(-u) = 1 - \Phi(u)$$

Kvantily standardizovaného normálního rozložení

α	u_α	α	u_α	α	u_α	α	u_α
0,500	0,00000	0,850	1,03643	0,930	1,47579	0,965	1,81191
0,510	0,02507	0,860	1,08032	0,931	1,48328	0,966	1,82501
0,520	0,05015	0,870	1,12639	0,932	1,49085	0,967	1,83842
0,530	0,07527	0,880	1,17499	0,933	1,49851	0,968	1,85218
0,540	0,10043	0,890	1,22653	0,934	1,50626	0,969	1,86630
0,550	0,12566	0,900	1,28155	0,935	1,51410	0,970	1,88079
0,560	0,15097	0,901	1,28727	0,936	1,52204	0,971	1,89570
0,570	0,17637	0,902	1,29303	0,937	1,53007	0,972	1,91104
0,580	0,20189	0,903	1,29884	0,938	1,53820	0,973	1,92684
0,590	0,22754	0,904	1,30469	0,939	1,54643	0,974	1,94313
0,600	0,25335	0,905	1,31058	0,940	1,55477	0,975	1,95996
0,610	0,27932	0,906	1,31652	0,941	1,56322	0,976	1,97737
0,620	0,30548	0,907	1,32251	0,942	1,57179	0,977	1,99539
0,630	0,33185	0,908	1,32854	0,943	1,58047	0,978	2,01409
0,640	0,35846	0,909	1,33462	0,944	1,58927	0,979	2,03352
0,650	0,38532	0,910	1,34076	0,945	1,59819	0,980	2,05375
0,660	0,41246	0,911	1,34694	0,946	1,60725	0,981	2,07485
0,670	0,43991	0,912	1,35317	0,947	1,61644	0,982	2,09693
0,680	0,46770	0,913	1,35946	0,948	1,62576	0,983	2,12007
0,690	0,49585	0,914	1,36581	0,949	1,63523	0,984	2,14441
0,700	0,52440	0,915	1,37220	0,950	1,64485	0,985	2,17009
0,710	0,55338	0,916	1,37866	0,951	1,65463	0,986	2,19729
0,720	0,58284	0,917	1,38517	0,952	1,66456	0,987	2,22621
0,730	0,61281	0,918	1,39174	0,953	1,67466	0,988	2,25713
0,740	0,64335	0,919	1,39838	0,954	1,68494	0,989	2,29037
0,750	0,67449	0,920	1,40507	0,955	1,69540	0,990	2,32635
0,760	0,70630	0,921	1,41183	0,956	1,70604	0,991	2,36562
0,770	0,73885	0,922	1,41865	0,957	1,71689	0,992	2,40892
0,780	0,77219	0,923	1,42554	0,958	1,72793	0,993	2,45726
0,790	0,80642	0,924	1,43250	0,959	1,73920	0,994	2,51214
0,800	0,84162	0,925	1,43953	0,960	1,75069	0,995	2,57583
0,810	0,87790	0,926	1,44663	0,961	1,76241	0,996	2,65207
0,820	0,91537	0,927	1,45381	0,962	1,77438	0,997	2,74778
0,830	0,95417	0,928	1,46106	0,963	1,78661	0,998	2,87816
0,840	0,99446	0,929	1,46838	0,964	1,79912	0,999	3,09023

Kvantily Pearsonova rozložení

			α		
n	0,001	0,005	0,010	0,025	0,050
	0,001	0,005	0,010	0,025	0,050
1	0,000	0,000	0,000	0,001	0,004
2	0,002	0,010	0,020	0,051	0,103
3	0,024	0,072	0,115	0,216	0,352
4	0,091	0,207	0,297	0,484	0,711
5	0,210	0,412	0,554	0,831	1,145
6	0,381	0,676	0,872	1,237	1,635
7	0,598	0,989	1,239	1,690	2,167
8	0,857	1,344	1,646	2,180	2,733
9	1,152	1,735	2,088	2,700	3,325
10	1,479	2,156	2,558	3,247	3,940
11	1,834	2,603	3,053	3,816	4,575
12	2,214	3,074	3,571	4,404	5,226
13	2,617	3,565	4,107	5,009	5,892
14	3,041	4,075	4,660	5,629	6,571
15	3,483	4,601	5,229	6,262	7,261
16	3,942	5,142	5,812	6,908	7,962
17	4,416	5,697	6,408	7,564	8,672
18	4,905	6,265	7,015	8,231	9,390
19	5,407	6,844	7,633	8,907	10,117
20	5,921	7,434	8,260	9,591	10,851
21	6,447	8,034	8,897	10,283	11,591
22	6,983	8,643	9,542	10,982	12,338
23	7,529	9,260	10,196	11,689	13,091
24	8,085	9,886	10,856	12,401	13,848
25	8,649	10,520	11,524	13,120	14,611
26	9,222	11,160	12,198	13,844	15,379
27	9,803	11,808	12,879	14,573	16,151
28	10,391	12,461	13,565	15,308	16,928
29	10,986	13,121	14,256	16,047	17,708
30	11,588	13,787	14,953	16,791	18,493
35	14,688	17,192	18,509	20,569	22,465
40	17,916	20,707	22,164	24,433	26,509
45	21,251	24,311	25,901	28,366	30,612
50	24,674	27,991	29,707	32,357	34,764
55	28,173	31,735	33,570	36,398	38,958
60	31,738	35,534	37,485	40,482	43,188
65	35,362	39,383	41,444	44,603	47,450
70	39,036	43,275	45,442	48,758	51,739
75	42,757	47,206	49,475	52,942	56,054
80	46,520	51,172	53,540	57,153	60,391
85	50,320	55,170	57,634	61,389	64,749
90	54,155	59,196	61,754	65,647	69,126
95	58,022	63,250	65,898	69,925	73,520
100	61,918	67,328	70,065	74,222	77,929

Kvantily Pearsonova rozložení

	α				
n	0,950	0,975	0,990	0,995	0,999
1	3,841	5,024	6,635	7,879	10,828
2	5,991	7,378	9,210	10,597	13,816
3	7,815	9,348	11,345	12,838	16,266
4	9,488	11,143	13,277	14,860	18,467
5	11,070	12,833	15,086	16,750	20,515
6	12,592	14,449	16,812	18,548	22,458
7	14,067	16,013	18,475	20,278	24,322
8	15,507	17,535	20,090	21,955	26,124
9	16,919	19,023	21,666	23,589	27,877
10	18,307	20,483	23,209	25,188	29,588
11	19,675	21,920	24,725	26,757	31,264
12	21,026	23,337	26,217	28,300	32,909
13	22,362	24,736	27,688	29,819	34,528
14	23,685	26,119	29,141	31,319	36,123
15	24,996	27,488	30,578	32,801	37,697
16	26,296	28,845	32,000	34,267	39,252
17	27,587	30,191	33,409	35,718	40,790
18	28,869	31,526	34,805	37,156	42,312
19	30,144	32,852	36,191	38,582	43,820
20	31,410	34,170	37,566	39,997	45,315
21	32,671	35,479	38,932	41,401	46,797
22	33,924	36,781	40,289	42,796	48,268
23	35,172	38,076	41,638	44,181	49,728
24	36,415	39,364	42,980	45,559	51,179
25	37,652	40,646	44,314	46,928	52,620
26	38,885	41,923	45,642	48,290	54,052
27	40,113	43,195	46,963	49,645	55,476
28	41,337	44,461	48,278	50,993	56,892
29	42,557	45,722	49,588	52,336	58,301
30	43,773	46,979	50,892	53,672	59,703
35	49,802	53,203	57,342	60,275	66,619
40	55,758	59,342	63,691	66,766	73,402
45	61,656	65,410	69,957	73,166	80,077
50	67,505	71,420	76,154	79,490	86,661
55	73,311	77,380	82,292	85,749	93,168
60	79,082	83,298	88,379	91,952	99,607
65	84,821	89,177	94,422	98,105	105,988
70	90,531	95,023	100,425	104,215	112,317
75	96,217	100,839	106,393	110,286	118,599
80	101,879	106,629	112,329	116,321	124,839
85	107,522	112,393	118,236	122,325	131,041
90	113,145	118,136	124,116	128,299	137,208
95	118,752	123,858	129,973	134,247	143,344
100	124,342	129,561	135,807	140,169	149,449

Kvantily Studentova rozložení

n	α 0,900	0,950	0,975	0,990	0,995	0,999
1	3,0777	6,3138	12,7062	31,8205	63,6567	318,3088
2	1,8856	2,9200	4,3027	6,9646	9,9248	22,3271
3	1,6377	2,3534	3,1824	4,5407	5,8409	10,2145
4	1,5332	2,1318	2,7764	3,7469	4,6041	7,1732
5	1,4759	2,0150	2,5706	3,3649	4,0321	5,8934
6	1,4398	1,9432	2,4469	3,1427	3,7074	5,2076
7	1,4149	1,8946	2,3646	2,9980	3,4995	4,7853
8	1,3968	1,8595	2,3060	2,8965	3,3554	4,5008
9	1,3830	1,8331	2,2622	2,8214	3,2498	4,2968
10	1,3722	1,8125	2,2281	2,7638	3,1693	4,1437
11	1,3634	1,7959	2,2010	2,7181	3,1058	4,0247
12	1,3562	1,7823	2,1788	2,6810	3,0545	3,9296
13	1,3502	1,7709	2,1604	2,6503	3,0123	3,8520
14	1,3450	1,7613	2,1448	2,6245	2,9768	3,7874
15	1,3406	1,7531	2,1314	2,6025	2,9467	3,7328
16	1,3368	1,7459	2,1199	2,5835	2,9208	3,6862
17	1,3334	1,7396	2,1098	2,5669	2,8982	3,6458
18	1,3304	1,7341	2,1009	2,5524	2,8784	3,6105
19	1,3277	1,7291	2,0930	2,5395	2,8609	3,5794
20	1,3253	1,7247	2,0860	2,5280	2,8453	3,5518
21	1,3232	1,7207	2,0796	2,5176	2,8314	3,5272
22	1,3212	1,7171	2,0739	2,5083	2,8188	3,5050
23	1,3195	1,7139	2,0687	2,4999	2,8073	3,4850
24	1,3178	1,7109	2,0639	2,4922	2,7969	3,4668
25	1,3163	1,7081	2,0595	2,4851	2,7874	3,4502
26	1,3150	1,7056	2,0555	2,4786	2,7787	3,4350
27	1,3137	1,7033	2,0518	2,4727	2,7707	3,4210
28	1,3125	1,7011	2,0484	2,4671	2,7633	3,4082
29	1,3114	1,6991	2,0452	2,4620	2,7564	3,3962
30	1,3104	1,6973	2,0423	2,4573	2,7500	3,3852
∞	1,2816	1,6449	1,9600	2,3263	2,5758	3,0000

Kvantily Fischerova-Snedecorova rozložení pro $\alpha = 0{,}95$

n_2	n_1						
	1	2	3	4	5	6	7
1	161,4500	199,5000	215,7074	224,5832	230,1619	233,9860	236,7684
2	18,5128	19,0000	19,1643	19,2468	19,2964	19,3295	19,3532
3	10,1280	9,5521	9,2766	9,1172	9,0135	8,9406	8,8867
4	7,7086	6,9443	6,5914	6,3882	6,2561	6,1631	6,0942
5	6,6079	5,7861	5,4095	5,1922	5,0503	4,9503	4,8759
6	5,9874	5,1433	4,7571	4,5337	4,3874	4,2839	4,2067
7	5,5914	4,7374	4,3468	4,1203	3,9715	3,8660	3,7870
8	5,3177	4,4590	4,0662	3,8379	3,6875	3,5806	3,5005
9	5,1174	4,2565	3,8625	3,6331	3,4817	3,3738	3,2927
10	4,9646	4,1028	3,7083	3,4780	3,3258	3,2172	3,1355
11	4,8443	3,9823	3,5874	3,3567	3,2039	3,0946	3,0123
12	4,7472	3,8853	3,4903	3,2592	3,1059	2,9961	2,9134
13	4,6672	3,8056	3,4105	3,1791	3,0254	2,9153	2,8321
14	4,6001	3,7389	3,3439	3,1122	2,9582	2,8477	2,7642
15	4,5431	3,6823	3,2874	3,0556	2,9013	2,7905	2,7066
16	4,4940	3,6337	3,2389	3,0069	2,8524	2,7413	2,6572
17	4,4513	3,5915	3,1968	2,9647	2,8100	2,6987	2,6143
18	4,4139	3,5546	3,1599	2,9277	2,7729	2,6613	2,5767
19	4,3807	3,5219	3,1274	2,8951	2,7401	2,6283	2,5435
20	4,3512	3,4928	3,0984	2,8661	2,7109	2,5990	2,5140
21	4,3248	3,4668	3,0725	2,8401	2,6848	2,5727	2,4876
22	4,3009	3,4434	3,0491	2,8167	2,6613	2,5491	2,4638
23	4,2793	3,4221	3,0280	2,7955	2,6400	2,5277	2,4422
24	4,2597	3,4028	3,0088	2,7763	2,6207	2,5082	2,4226
25	4,2417	3,3852	2,9912	2,7587	2,6030	2,4904	2,4047
26	4,2252	3,3690	2,9752	2,7426	2,5868	2,4741	2,3883
27	4,2100	3,3541	2,9604	2,7278	2,5719	2,4591	2,3732
28	4,1960	3,3404	2,9467	2,7141	2,5581	2,4453	2,3593
29	4,1830	3,3277	2,9340	2,7014	2,5454	2,4324	2,3463
30	4,1709	3,3158	2,9223	2,6896	2,5336	2,4205	2,3343
40	4,0847	3,2317	2,8387	2,6060	2,4495	2,3359	2,2490
60	4,0012	3,1504	2,7581	2,5252	2,3683	2,2541	2,1665
80	3,9604	3,1108	2,7188	2,4859	2,3287	2,2142	2,1263
120	3,9201	3,0718	2,6802	2,4472	2,2899	2,1750	2,0868
∞	3,8415	2,9957	2,6049	2,3719	2,2141	2,0986	2,0096

Kvantily Fischerova-Snedecorova rozložení pro $\alpha = 0{,}95$

	n_1						
n_2	8	9	10	11	12	13	14
1	238,8827	240,5433	241,8818	242,9835	243,9060	244,6899	245,3640
2	19,3710	19,3848	19,3959	19,4050	19,4125	19,4189	19,4244
3	8,8452	8,8123	8,7855	8,7633	8,7446	8,7287	8,7149
4	6,0410	5,9988	5,9644	5,9358	5,9117	5,8911	5,8733
5	4,8183	4,7725	4,7351	4,7040	4,6777	4,6552	4,6358
6	4,1468	4,0990	4,0600	4,0274	3,9999	3,9764	3,9559
7	3,7257	3,6767	3,6365	3,6030	3,5747	3,5503	3,5292
8	3,4381	3,3881	3,3472	3,3130	3,2839	3,2590	3,2374
9	3,2296	3,1789	3,1373	3,1025	3,0729	3,0475	3,0255
10	3,0717	3,0204	2,9782	2,9430	2,9130	2,8872	2,8647
11	2,9480	2,8962	2,8536	2,8179	2,7876	2,7614	2,7386
12	2,8486	2,7964	2,7534	2,7173	2,6866	2,6602	2,6371
13	2,7669	2,7144	2,6710	2,6347	2,6037	2,5769	2,5536
14	2,6987	2,6458	2,6022	2,5655	2,5342	2,5073	2,4837
15	2,6408	2,5876	2,5437	2,5068	2,4753	2,4481	2,4244
16	2,5911	2,5377	2,4935	2,4564	2,4247	2,3973	2,3733
17	2,5480	2,4943	2,4499	2,4126	2,3807	2,3531	2,3290
18	2,5102	2,4563	2,4117	2,3742	2,3421	2,3143	2,2900
19	2,4768	2,4227	2,3779	2,3402	2,3080	2,2800	2,2556
20	2,4471	2,3928	2,3479	2,3100	2,2776	2,2495	2,2250
21	2,4205	2,3660	2,3210	2,2829	2,2504	2,2222	2,1975
22	2,3965	2,3419	2,2967	2,2585	2,2258	2,1975	2,1727
23	2,3748	2,3201	2,2747	2,2364	2,2036	2,1752	2,1502
24	2,3551	2,3002	2,2547	2,2163	2,1834	2,1548	2,1298
25	2,3371	2,2821	2,2365	2,1979	2,1649	2,1362	2,1111
26	2,3205	2,2655	2,2197	2,1811	2,1479	2,1192	2,0939
27	2,3053	2,2501	2,2043	2,1655	2,1323	2,1035	2,0781
28	2,2913	2,2360	2,1900	2,1512	2,1179	2,0889	2,0635
29	2,2783	2,2229	2,1768	2,1379	2,1045	2,0755	2,0500
30	2,2662	2,2107	2,1646	2,1256	2,0921	2,0630	2,0374
40	2,1802	2,1240	2,0772	2,0376	2,0035	1,9738	1,9476
60	2,0970	2,0401	1,9926	1,9522	1,9174	1,8870	1,8602
80	2,0564	1,9991	1,9512	1,9105	1,8753	1,8445	1,8174
120	2,0164	1,9588	1,9105	1,8693	1,8337	1,8026	1,7750
∞	1,9384	1,8799	1,8307	1,7886	1,7522	1,7202	1,6918

Kvantily Fischerova-Snedecorova rozložení pro $\alpha = 0{,}95$

n_2	n_1 15	16	17	18	19	20	25
1	245,9499	246,4639	246,9184	247,3232	247,6861	248,0131	249,2601
2	19,4291	19,4333	19,4370	19,4402	19,4431	19,4458	19,4558
3	8,7029	8,6923	8,6829	8,6745	8,6670	8,6602	8,6341
4	5,8578	5,8441	5,8320	5,8211	5,8114	5,8025	5,7687
5	4,6188	4,6038	4,5904	4,5785	4,5678	4,5581	4,5209
6	3,9381	3,9223	3,9083	3,8957	3,8844	3,8742	3,8348
7	3,5107	3,4944	3,4799	3,4669	3,4551	3,4445	3,4036
8	3,2184	3,2016	3,1867	3,1733	3,1613	3,1503	3,1081
9	3,0061	2,9890	2,9737	2,9600	2,9477	2,9365	2,8932
10	2,8450	2,8276	2,8120	2,7980	2,7854	2,7740	2,7298
11	2,7186	2,7009	2,6851	2,6709	2,6581	2,6464	2,6014
12	2,6169	2,5989	2,5828	2,5684	2,5554	2,5436	2,4977
13	2,5331	2,5149	2,4987	2,4841	2,4709	2,4589	2,4123
14	2,4630	2,4446	2,4282	2,4134	2,4000	2,3879	2,3407
15	2,4034	2,3849	2,3683	2,3533	2,3398	2,3275	2,2797
16	2,3522	2,3335	2,3167	2,3016	2,2880	2,2756	2,2272
17	2,3077	2,2888	2,2719	2,2567	2,2429	2,2304	2,1815
18	2,2686	2,2496	2,2325	2,2172	2,2033	2,1906	2,1413
19	2,2341	2,2149	2,1977	2,1823	2,1683	2,1555	2,1057
20	2,2033	2,1840	2,1667	2,1511	2,1370	2,1242	2,0739
21	2,1757	2,1563	2,1389	2,1232	2,1090	2,0960	2,0454
22	2,1508	2,1313	2,1138	2,0980	2,0837	2,0707	2,0196
23	2,1282	2,1086	2,0910	2,0751	2,0608	2,0476	1,9963
24	2,1077	2,0880	2,0703	2,0543	2,0399	2,0267	1,9750
25	2,0889	2,0691	2,0513	2,0353	2,0207	2,0075	1,9554
26	2,0716	2,0518	2,0339	2,0178	2,0032	1,9898	1,9375
27	2,0558	2,0358	2,0179	2,0017	1,9870	1,9736	1,9210
28	2,0411	2,0210	2,0030	1,9868	1,9720	1,9586	1,9057
29	2,0275	2,0073	1,9893	1,9730	1,9581	1,9446	1,8915
30	2,0148	1,9946	1,9765	1,9601	1,9452	1,9317	1,8782
40	1,9245	1,9037	1,8851	1,8682	1,8529	1,8389	1,7835
60	1,8364	1,8151	1,7959	1,7784	1,7625	1,7480	1,6902
80	1,7932	1,7716	1,7520	1,7342	1,7180	1,7032	1,6440
120	1,7505	1,7285	1,7085	1,6904	1,6739	1,6587	1,5980
∞	1,6640	1,6435	1,6228	1,6038	1,5865	1,5705	1,5061

Kvantily Fischerova-Snedecorova rozložení pro $\alpha = 0{,}95$

	n_1					
n_2	30	40	60	80	120	∞
1	250,0952	251,1432	252,1957	252,7237	253,2529	254,3100
2	19,4624	19,4707	19,4791	19,4832	19,4874	19,4960
3	8,6166	8,5944	8,5720	8,5607	8,5494	8,5264
4	5,7459	5,7170	5,6877	5,6730	5,6581	5,6281
5	4,4957	4,4638	4,4314	4,4150	4,3985	4,3650
6	3,8082	3,7743	3,7398	3,7223	3,7047	3,6689
7	3,3758	3,3404	3,3043	3,2860	3,2674	3,2298
8	3,0794	3,0428	3,0053	2,9862	2,9669	2,9276
9	2,8637	2,8259	2,7872	2,7675	2,7475	2,7067
10	2,6996	2,6609	2,6211	2,6008	2,5801	2,5379
11	2,5705	2,5309	2,4901	2,4692	2,4480	2,4045
12	2,4663	2,4259	2,3842	2,3628	2,3410	2,2962
13	2,3803	2,3392	2,2966	2,2747	2,2524	2,2064
14	2,3082	2,2664	2,2229	2,2006	2,1778	2,1307
15	2,2468	2,2043	2,1601	2,1373	2,1141	2,0658
16	2,1938	2,1507	2,1058	2,0826	2,0589	2,0096
17	2,1477	2,1040	2,0584	2,0348	2,0107	1,9604
18	2,1071	2,0629	2,0166	1,9927	1,9681	1,9168
19	2,0712	2,0264	1,9795	1,9552	1,9302	1,8780
20	2,0391	1,9938	1,9464	1,9217	1,8963	1,8432
21	2,0102	1,9645	1,9165	1,8915	1,8657	1,8117
22	1,9842	1,9380	1,8894	1,8641	1,8380	1,7831
23	1,9605	1,9139	1,8648	1,8392	1,8128	1,7570
24	1,9390	1,8920	1,8424	1,8164	1,7896	1,7330
25	1,9192	1,8718	1,8217	1,7955	1,7684	1,7110
26	1,9010	1,8533	1,8027	1,7762	1,7488	1,6906
27	1,8842	1,8361	1,7851	1,7584	1,7306	1,6717
28	1,8687	1,8203	1,7689	1,7418	1,7138	1,6541
29	1,8543	1,8055	1,7537	1,7264	1,6981	1,6376
30	1,8409	1,7918	1,7396	1,7121	1,6835	1,6223
40	1,7444	1,6928	1,6373	1,6077	1,5766	1,5089
60	1,6491	1,5943	1,5343	1,5019	1,4673	1,3893
80	1,6017	1,5449	1,4821	1,4477	1,4107	1,3247
120	1,5543	1,4952	1,4290	1,3922	1,3519	1,2539
∞	1,4591	1,3940	1,3180	1,2735	1,2214	1,0000

Kvantily Fischerova-Snedecorova rozložení pro $\alpha = 0{,}975$

n_2	n_1 = 1	2	3	4	5	6	7
1	647,7890	799,5000	864,1630	899,5833	921,8479	937,1111	948,2169
2	38,5063	39,0000	39,1655	39,2484	39,2982	39,3315	39,3552
3	17,4434	16,0441	15,4392	15,1010	14,8848	14,7347	14,6244
4	12,2179	10,6491	9,9792	9,6045	9,3645	9,1973	9,0741
5	10,0070	8,4336	7,7636	7,3879	7,1464	6,9777	6,8531
6	8,8131	7,2599	6,5988	6,2272	5,9876	5,8198	5,6955
7	8,0727	6,5415	5,8898	5,5226	5,2852	5,1186	4,9949
8	7,5709	6,0595	5,4160	5,0526	4,8173	4,6517	4,5286
9	7,2093	5,7147	5,0781	4,7181	4,4844	4,3197	4,1970
10	6,9367	5,4564	4,8256	4,4683	4,2361	4,0721	3,9498
11	6,7241	5,2559	4,6300	4,2751	4,0440	3,8807	3,7586
12	6,5538	5,0959	4,4742	4,1212	3,8911	3,7283	3,6065
13	6,4143	4,9653	4,3472	3,9959	3,7667	3,6043	3,4827
14	6,2979	4,8567	4,2417	3,8919	3,6634	3,5014	3,3799
15	6,1995	4,7650	4,1528	3,8043	3,5764	3,4147	3,2934
16	6,1151	4,6867	4,0768	3,7294	3,5021	3,3406	3,2194
17	6,0420	4,6189	4,0112	3,6648	3,4379	3,2767	3,1556
18	5,9781	4,5597	3,9539	3,6083	3,3820	3,2209	3,0999
19	5,9216	4,5075	3,9034	3,5587	3,3327	3,1718	3,0509
20	5,8715	4,4613	3,8587	3,5147	3,2891	3,1283	3,0074
21	5,8266	4,4199	3,8188	3,4754	3,2501	3,0895	2,9686
22	5,7863	4,3828	3,7829	3,4401	3,2151	3,0546	2,9338
23	5,7498	4,3492	3,7505	3,4083	3,1835	3,0232	2,9023
24	5,7166	4,3187	3,7211	3,3794	3,1548	2,9946	2,8738
25	5,6864	4,2909	3,6943	3,3530	3,1287	2,9685	2,8478
26	5,6586	4,2655	3,6697	3,3289	3,1048	2,9447	2,8240
27	5,6331	4,2421	3,6472	3,3067	3,0828	2,9228	2,8021
28	5,6096	4,2205	3,6264	3,2863	3,0626	2,9027	2,7820
29	5,5878	4,2006	3,6072	3,2674	3,0438	2,8840	2,7633
30	5,5675	4,1821	3,5894	3,2499	3,0265	2,8667	2,7460
40	5,4239	4,0510	3,4633	3,1261	2,9037	2,7444	2,6238
60	5,2856	3,9253	3,3425	3,0077	2,7863	2,6274	2,5068
80	5,2184	3,8643	3,2841	2,9504	2,7295	2,5708	2,4502
120	5,1523	3,8046	3,2269	2,8943	2,6740	2,5154	2,3948
∞	5,0239	3,6889	3,1161	2,7858	2,5665	2,4082	2,2875

Kvantily Fischerova-Snedecorova rozložení pro $\alpha = 0{,}975$

	n_1						
n_2	8	9	10	11	12	13	14
1	956,6562	963,2846	968,6274	973,0252	976,7080	979,8368	982,5278
2	39,3730	39,3869	39,3980	39,4071	39,4146	39,4210	39,4265
3	14,5399	14,4731	14,4189	14,3742	14,3366	14,3045	14,2768
4	8,9796	8,9047	8,8439	8,7935	8,7512	8,7150	8,6838
5	6,7572	6,6811	6,6192	6,5678	6,5245	6,4876	6,4556
6	5,5996	5,5234	5,4613	5,4098	5,3662	5,3290	5,2968
7	4,8993	4,8232	4,7611	4,7095	4,6658	4,6285	4,5961
8	4,4333	4,3572	4,2951	4,2434	4,1997	4,1622	4,1297
9	4,1020	4,0260	3,9639	3,9121	3,8682	3,8306	3,7980
10	3,8549	3,7790	3,7168	3,6649	3,6209	3,5832	3,5504
11	3,6638	3,5879	3,5257	3,4737	3,4296	3,3917	3,3588
12	3,5118	3,4358	3,3736	3,3215	3,2773	3,2393	3,2062
13	3,3880	3,3120	3,2497	3,1975	3,1532	3,1150	3,0819
14	3,2853	3,2093	3,1469	3,0946	3,0502	3,0119	2,9786
15	3,1987	3,1227	3,0602	3,0078	2,9633	2,9249	2,8915
16	3,1248	3,0488	2,9862	2,9337	2,8890	2,8506	2,8170
17	3,0610	2,9849	2,9222	2,8696	2,8249	2,7863	2,7526
18	3,0053	2,9291	2,8664	2,8137	2,7689	2,7302	2,6964
19	2,9563	2,8801	2,8172	2,7645	2,7196	2,6808	2,6469
20	2,9128	2,8365	2,7737	2,7209	2,6758	2,6369	2,6030
21	2,8740	2,7977	2,7348	2,6819	2,6368	2,5978	2,5638
22	2,8392	2,7628	2,6998	2,6469	2,6017	2,5626	2,5285
23	2,8077	2,7313	2,6682	2,6152	2,5699	2,5308	2,4966
24	2,7791	2,7027	2,6396	2,5865	2,5411	2,5019	2,4677
25	2,7531	2,6766	2,6135	2,5603	2,5149	2,4756	2,4413
26	2,7293	2,6528	2,5896	2,5363	2,4908	2,4515	2,4171
27	2,7074	2,6309	2,5676	2,5143	2,4688	2,4293	2,3949
28	2,6872	2,6106	2,5473	2,4940	2,4484	2,4089	2,3743
29	2,6686	2,5919	2,5286	2,4752	2,4295	2,3900	2,3554
30	2,6513	2,5746	2,5112	2,4577	2,4120	2,3724	2,3378
40	2,5289	2,4519	2,3882	2,3343	2,2882	2,2481	2,2130
60	2,4117	2,3344	2,2702	2,2159	2,1692	2,1286	2,0929
80	2,3549	2,2775	2,2130	2,1584	2,1115	2,0706	2,0346
120	2,2994	2,2217	2,1570	2,1021	2,0548	2,0136	1,9773
∞	2,1918	2,1136	2,0483	1,9927	1,9447	1,9027	1,8656

Kvantily Fischerova-Snedecorova rozložení pro $\alpha = 0{,}975$

n_2	n_1 15	16	17	18	19	20	25
1	984,8668	986,9187	988,7331	990,3490	991,7973	993,1028	998,0808
2	39,4313	39,4354	39,4391	39,4424	39,4453	39,4479	39,4579
3	14,2527	14,2315	14,2127	14,1960	14,1810	14,1674	14,1155
4	8,6565	8,6326	8,6113	8,5924	8,5753	8,5599	8,5010
5	6,4277	6,4032	6,3814	6,3619	6,3444	6,3286	6,2679
6	5,2687	5,2439	5,2218	5,2021	5,1844	5,1684	5,1069
7	4,5678	4,5428	4,5206	4,5008	4,4829	4,4667	4,4045
8	4,1012	4,0761	4,0538	4,0338	4,0158	3,9995	3,9367
9	3,7694	3,7441	3,7216	3,7015	3,6833	3,6669	3,6035
10	3,5217	3,4963	3,4737	3,4534	3,4351	3,4185	3,3546
11	3,3299	3,3044	3,2816	3,2612	3,2428	3,2261	3,1616
12	3,1772	3,1515	3,1286	3,1081	3,0896	3,0728	3,0077
13	3,0527	3,0269	3,0039	2,9832	2,9646	2,9477	2,8821
14	2,9493	2,9234	2,9003	2,8795	2,8607	2,8437	2,7777
15	2,8621	2,8360	2,8128	2,7919	2,7730	2,7559	2,6894
16	2,7875	2,7614	2,7380	2,7170	2,6980	2,6808	2,6138
17	2,7230	2,6968	2,6733	2,6522	2,6331	2,6158	2,5484
18	2,6667	2,6404	2,6168	2,5956	2,5764	2,5590	2,4912
19	2,6171	2,5907	2,5670	2,5457	2,5265	2,5089	2,4408
20	2,5731	2,5465	2,5228	2,5014	2,4821	2,4645	2,3959
21	2,5338	2,5071	2,4833	2,4618	2,4424	2,4247	2,3558
22	2,4984	2,4717	2,4478	2,4262	2,4067	2,3890	2,3198
23	2,4665	2,4396	2,4157	2,3940	2,3745	2,3567	2,2871
24	2,4374	2,4105	2,3865	2,3648	2,3452	2,3273	2,2574
25	2,4110	2,3840	2,3599	2,3381	2,3184	2,3005	2,2303
26	2,3867	2,3597	2,3355	2,3137	2,2939	2,2759	2,2054
27	2,3644	2,3373	2,3131	2,2912	2,2713	2,2533	2,1826
28	2,3438	2,3167	2,2924	2,2704	2,2505	2,2324	2,1615
29	2,3248	2,2976	2,2732	2,2512	2,2313	2,2131	2,1419
30	2,3072	2,2799	2,2554	2,2334	2,2134	2,1952	2,1237
40	2,1819	2,1542	2,1293	2,1068	2,0864	2,0677	1,9943
60	2,0613	2,0330	2,0076	1,9846	1,9636	1,9445	1,8687
80	2,0026	1,9741	1,9483	1,9250	1,9037	1,8843	1,8071
120	1,9450	1,9161	1,8900	1,8663	1,8447	1,8249	1,7462
∞	1,8326	1,8028	1,7759	1,7515	1,7291	1,7085	1,6259

Kvantily Fischerova-Snedecorova rozložení pro $\alpha = 0{,}975$

	n_1					
n_2	30	40	60	80	120	∞
1	1001,4140	1005,5980	1009,8000	1011,9080	1014,0200	1018,3000
2	39,4646	39,4729	39,4812	39,4854	39,4896	39,4980
3	14,0805	14,0365	13,9921	13,9697	13,9473	13,9020
4	8,4613	8,4111	8,3604	8,3349	8,3092	8,2573
5	6,2269	6,1750	6,1225	6,0960	6,0693	6,0153
6	5,0652	5,0125	4,9589	4,9318	4,9044	4,8491
7	4,3624	4,3089	4,2544	4,2268	4,1989	4,1423
8	3,8940	3,8398	3,7844	3,7563	3,7279	3,6702
9	3,5604	3,5055	3,4493	3,4207	3,3918	3,3329
10	3,3110	3,2554	3,1984	3,1694	3,1399	3,0798
11	3,1176	3,0613	3,0035	2,9740	2,9441	2,8828
12	2,9633	2,9063	2,8478	2,8178	2,7874	2,7249
13	2,8372	2,7797	2,7204	2,6900	2,6590	2,5955
14	2,7324	2,6742	2,6142	2,5833	2,5519	2,4872
15	2,6437	2,5850	2,5242	2,4930	2,4611	2,3953
16	2,5678	2,5085	2,4471	2,4154	2,3831	2,3163
17	2,5020	2,4422	2,3801	2,3481	2,3153	2,2474
18	2,4445	2,3842	2,3214	2,2890	2,2558	2,1869
19	2,3937	2,3329	2,2696	2,2368	2,2032	2,1333
20	2,3486	2,2873	2,2234	2,1902	2,1562	2,0853
21	2,3082	2,2465	2,1819	2,1485	2,1141	2,0422
22	2,2718	2,2097	2,1446	2,1108	2,0760	2,0032
23	2,2389	2,1763	2,1107	2,0766	2,0415	1,9677
24	2,2090	2,1460	2,0799	2,0454	2,0099	1,9353
25	2,1816	2,1183	2,0516	2,0169	1,9811	1,9055
26	2,1565	2,0928	2,0257	1,9907	1,9545	1,8781
27	2,1334	2,0693	2,0018	1,9665	1,9299	1,8527
28	2,1121	2,0477	1,9797	1,9441	1,9072	1,8291
29	2,0923	2,0276	1,9591	1,9232	1,8861	1,8072
30	2,0739	2,0089	1,9400	1,9039	1,8664	1,7867
40	1,9429	1,8752	1,8028	1,7644	1,7242	1,6371
60	1,8152	1,7440	1,6668	1,6252	1,5810	1,4821
80	1,7523	1,6790	1,5987	1,5549	1,5079	1,3997
120	1,6899	1,6141	1,5299	1,4834	1,4327	1,3104
∞	1,5660	1,4835	1,3883	1,3329	1,2684	1,0000

Příloha B – Základní informace o programu STATISTICA 6

Systém má modulární stavbu. V multilicenci pro Masarykovu univerzitu jsou k dispozici moduly: Basic Statistics/Tables, Multiple Regression, ANOVA, Nonparametrics, Distribution Fitting, Advanced Linear / Nonlinear Models, Multivariate Explorartory Techniques, Industrial Statistics & Six Sigma.

Velké množství informací o systému STATISTICA lze najít na webové stránce společnosti StatSoft, která je jejím distributorem v České republice (internetová adresa je `www.statsoft.cz`). Z této stránky vede rovněž odkaz na elektronickou učebnici statistiky.

STATISTICA 6 má několik typů oken:

- **spreadsheet** (datové okno, má příponu sta, jeho obsah však lze exportovat i v jiných formátech). Do datového okna lze načítat datové soubory nejrůznějších typů (např. z tabulkových procesorů, databázové soubory, ASCII soubory).
- **workbook** (má příponu stw). Do workbooku ukládají výstupy, tj. tabulky a grafy. Skládá se ze dvou oken, v levém okně je znázorněna stromová struktura výstupů, v pravém jsou samotné výstupy. V levém okně se lze pohybovat myší nebo kurzorem, mazat, přesouvat, editovat apod. Výstupy mohou sloužit jako vstupy pro další analýzy a grafy.
- **report** (má příponu str, lze ho uložit i ve formátu rtf, txt či htm). Pokud požadujeme, aby se výstupy ukládaly nejen do workbooku, ale i do reportu, postupujeme takto: Tools - Options - Output Manager - zaškrtneme Also send to Report Window - OK. Report se podobně jako workbook skládá ze dvou oken. Do reportu můžeme vkládat vlastní text, vysvětlující komentáře, poznámky apod. Tabulky a grafy lze v reportu i workbooku dále upravovat.
- **okno grafů** (přípona stg, lze ho uložit i jako bmp, jpg, png a wmf). Získá se tak, že ve workbooku klikneme pravým tlačítkem na graf a vybereme Clone Graph.
- **programovací okno** (přípona svb). Slouží pro zápis programů v jazyku STATISTICA Visual Basic. Mezi jednotlivými typy oken se přepínáme pomocí položky Window v hlavním menu.

B.1. Bodové zpracování četností

1. Zapište do datového okna programu STATISTICA datový soubor, který bude obsahovat známky z matematiky, angličtiny a údaje o pohlaví dvaceti studentů (viz příklad 1.10).
   Návod: File – New – Number of variables 3, Number of cases 20, OK.

2. Znaky nazvěte X, Y, Z, vytvořte jim návěští (X – známka z matematiky, Y – známka z angličtiny, Z – pohlaví studenta) a popište, co znamenají jednotlivé varianty (u znaků X a Y: 1 – výborně, 2 – velmi dobře, 3 – dobře, 4 - neprospěl, u znaku Z: 0 – žena, 1 – muž). Soubor uložte pod názvem znamky.sta.
   Návod: Kurzor nastavíme na Var1 – 2× klikneme myší – Name X – Long Name známka z matematiky, Text label – 1 výborně, 2 velmi dobře, 3 dobře, 4 neprospěl, OK. U proměnné Y lze text label okopírovat z proměnné X – v Text Labels Editor zvolíme Copy from variable X.
   Přepínání mezi číselnými hodnotami a jejich textovým popisem se děje pomocí tlačítka s obrázkem štítku.

3. U znaků X a Y vypočtěte absolutní četnosti, relativní četnosti a relativní kumulativní četnosti. Návod: Statistics - Basic Statistics/Tables - Frequency tables - OK - Variables X, Y, OK - Summary. Všechny tři tabulky se uloží do workbooku a listovat v nich můžeme pomocí stromové struktury v levém okně.

4. Vytvořte sloupkový diagram absolutních četností znaků X a Y.
   Návod: Graphs – Histograms – Variables X, Y – OK – vypneme Normal fit – Advanced – zaškrtneme Breaks between Columns, OK.

   Vytvořte výsečový diagram absolutních četností znaků X a Y.
   Návod: Graphs – 2D Graphs – Pie Charts – Variables X, Y – OK – Advanced – Pie legend Text and Percent (nebo Text and Value) – OK.

   Vytvořte polygon absolutních četností znaků X a Y.
   Návod: ve workbooku vstoupíme do tabulky rozložení četností proměnné X. Pomocí Edit – Delete – Cases vymažeme řádek označený Missing. Nastavíme se kurzorem na Count – Graphs – Graphs of Block Data – Line Plot:Entire Columns. Vykreslí se polygon četností.

5. Vytvořte graf empirické distribuční funkce znaku X.
   Návod: Při tvorbě histogramu zadáme v Advanced volbu Showing Type Cumulative, Y axis % – 2× klikneme myší na pozadí grafu – otevře se okno All Options – vybereme Plot: Bars – Type Rectangles. V tomto grafu jsou však svislé čáry až k vodorovné ose. Lze použít i jiný typ grafu: vytvoříme nový datový soubor, který bude mít dvě proměnné a případů o dva víc než je počet variant znaku X. Do 1. proměnné zapíšeme do 1. řádku hodnotu o 1 menší než je 1. varianta znaku X, pak varianty znaku X a nakonec hodnotu o 1 větší než je poslední varianta znaku X. Do 2. proměnné zapíšeme 0, pak relativní kumulativní četnosti znaku X (v procentech) a nakonec 100. Graphs – Scatterplots –Variables V1, V2 – OK – vypneme Linear fit – OK – 2× klikneme na pozadí grafu – Plot:General – vypneme Markers, zaškrtneme Line – Line Type: Step – OK.

Vytvořte graf četnostní funkce znaku X.
Návod: Při tvorbě histogramu zadáme v Advanced Y axis % – 2× klikneme myší na pozadí grafu – vybereme Plot General – zaškrtneme Markers – vybereme Plot:Bars – Type Lines.

6. Z datového souboru vyberte pouze ženy (pouze muže) a úkol 3 proveďte pro ženy (pro muže). Návod: Statistics - Basic Statistics/Tables - Frequency tables - OK - Variables X, Y, OK - Select Cases - zaškrtneme Selection Conditions - Include cases - zaškrtneme Specific, selected by Z = 0, OK.

7. Nadále pracujte s celým datovým souborem. Vytvořte kontingenční tabulku absolutních četností znaků X a Y a graf simultánní četností funkce.
   Návod: Statistics – Basic Statistics/Tables – Tables and banners – OK – Select cases – All – OK – Specify tables – List 1 X, List 2 Y, OK, Summary. Vytvoření grafu simultánní četnostní funkce: Návrat do Crosstabulation Tables Result – 3D histograms – vybereme Axis Scaling – Mode Manual – Minimum 0 (a totéž provedeme pro Axis Y) – dále vybereme Graph Layout – Type – Spikes – OK. Graf lze natáčet pomocí Point of View.

   Vytvořte kontingenční tabulku sloupcově a řádkově podmíněných relativních četností znaků X a Y.
   Návod: Návrat do Crosstabulation Tables Result – Options – zaškrtneme ve sloupci Compute tables volbu Percentages of column counts (resp. Percentages of row counts).

B.2. Intervalové zpracování četností

1. Zapište do datového okna programu STATISTICA datový soubor, který bude obsahovat údaje o mezi plasticity oceli a mezi pevnosti (viz příklad 2.13). Proměnným X a Y vytvořte návěští „mez plasticity“ a „mez pevnosti“. Soubor pak uložte pod názvem ocel.sta.
   Návod: viz 1. cvičení, bod 1.

2. Pro X a Y použijeme intervalové zpracování četností. Pro aplikaci Sturgersova pravidla potřebujeme znát počet variant proměnné X a Y.
   Návod: Zjištění absolutních četností – viz 1. cvičení, bod 3. Zjištění počtu variant: ve workbooku se nastavíme kurzorem na sloupec Count – 2× klikneme myší – vybereme Values/Stats – ve výstupní tabulce se objeví mj. N. Počet variant je N−1. (X má 50 variant, Y má 52 variant, v obou případech volíme 7 třídicích intervalů.) Dále musíme zjistit minimum a maximum, abychom vhodně stanovili třídicí intervaly.
   Návod: Statistics – Basic Statistics/Tables – Descriptive statistics – Variables X, Y – zaškrtneme Minimum & maximum – Summary. (Pro X je minimum 33 a maximum 160, tedy vhodná volba třídicích intervalů je $(30, 50\rangle$, $50, 70\rangle, \ldots, (150, 170\rangle$ – viz příklad 2.13, pro Y je minimum 52 a maximum 189, tedy třídicí intervaly zvolíme $(50, 70\rangle$, $70, 90, \ldots 170, 190\rangle$ – viz příklad 2.19.)

3. Vytvořte histogram pro X a pro Y.
   Návod: Graphs – Histograms – Variables X – vypneme Normal fit – Advanced – zaškrtneme Boundaries – Specify Boundaries – 50 70 90 110 130 150 170 OK – Y Axis %. 2× klikneme na pozadí grafu a ve volbě All Options můžeme měnit různé vlastnosti grafu.
   Upozornění: STATISTICA v histogramu znázorňuje relativní četnost výškou obdélníku, nikoliv jeho plochou, což není v souladu s definicí 2.14.

4. Proveďte zakódování hodnot proměnných X a Y do příslušných třídicích intervalů.
   Návod: Insert – Add Variables – 2 – After Y – OK – přejmenujeme je na RX a RY. Nastavíme se kurzorem na RX – Data – Recode – vyplníme podmínky pro všech 7 kategorií. (Pozor – podmínky se musí psát ve tvaru X>30 and X<=50 atd.). Pak klepneme na OK. Analogicky pro Y.

5. Vytvořte graf intervalové empirické distribuční funkce pro X.
   Návod: Vytvoříme Frequency table pro RX. Před 1. případ vložíme řádek, kde do Category napíšeme 0 a do Cumulative Count také 0. Nastavíme se kurzorem na Cumulative Percent – Graphs – Graphs of Block Data – Custom Graph from Block by Column – Line Plots (Variables) – OK. 2× klikneme na pozadí grafu – Plot: General – vypneme Markers – Axis: Scaling – Mode Manual – Minimum 1, Maximum 9 – Axis: Custom Units – Position 1, Text 30 atd až Position 9, Text 190 – OK.

6. Sestavte kontingenční tabulky absolutních četností (relativních četností, sloupcově a řádkově podmíněných relativních četností) dvourozměrných třídicích intervalů pro (X,Y).
   Návod: Viz úkol č. 6 ve cvičení 1, kde budeme pracovat s proměnnými RX a RY.

B.3. Výpočet číselných charakteristik jednorozměrného a dvourozměrného souboru, regresní přímka

1. Načtěte soubor znamky.sta. Pro známky z matematiky a angličtiny vypočtěte medián, dolní a horní kvartil a kvartilovou odchylku. Výsledky porovnejte s příkladem 3.5.
   Návod: Stastistics – Basic Statistics/Tables – Descriptive Statistics – OK – Variables X, Y, OK – zaškrtneme Median, Lower & upper quartiles, Quartile range – Summary.

2. Načtěte soubor ocel.sta. Pro mez plasticity a mez pevnosti vypočtěte aritmetické průměry, směrodatné odchylky a rozptyly. Výsledky porovnejte s příkladem 3.17.
   Návod: Návod: Stastistics – Basic Statistics/Tables – Descriptive Statistics – OK – Variables X, Y, OK – zaškrtneme Mean, Standard Deviation, Variance – Summary.
   Vysvětlení: Rozptyl a směrodatná odchylka vyjdou ve STATISTICE jinak než v příklad 3.17, protože STATISTICA ve vzorci pro výpočet rozptylu nepoužívá $1/n$, ale $1/(n-1)$ – bude objasněno později v matematické statistice.

3. Nakreslete dvourozměrný tečkový diagram pro (X,Y).
   Návod: Graphs – Scatterplots – Variables X,Y – OK – vypneme Linear fit – OK.

4. Vypočtěte kovarianci a koeficient korelace meze plasticity a meze pevnosti. Výsledky porovnejte s příkladem 3.17.
   Návod: Statistics – Multiple Regression – Variables Independent X, Dependent Y – OK – OK – Residuals/assumption-prediction – Descriptive statistics – Covariances. Pro získání korelačního koeficientu zvolíme Correlation místo Covariances.
   Vysvětlení: Kovariance vyjde ve STATISTICE jinak než v příkladu 3.17, protože ve STATISTICE se ve vzorci pro výpočet kovariance nepoužívá $1/n$, ale $1/(n-1)$ – bude objasněno později.

5. Určete koeficienty regresní přímky meze pevnosti na mez plasticity a stanovte index determinace. Určete regresní odhad meze pevnosti, je-li mez plasticity 110. Nakreslete regresní přímku do dvourozměrného tečkového diagramu.
   Návod: V tabulce Multiple Regression zvolíme Variables Independent X, Dependent Y – OK – Summary:Regression results. Ve výstupní tabulce najdeme koeficient b_0 ve sloupci B na řádku označeném Intercept, koeficient b_1 ve sloupci B na řádku označeném X, index determinace pod označením R2. Pro výpočet predikované hodnoty zvolíme Residuals/assumption/prediction Predict dependent variable X:110 – OK. Ve výstupní tabulce je hledaná hodnota označena jako Predictd.
   Nakreslení regresní přímky: Návrat do Multiple Regression – Residuals / assumption / prediction – Perform residuals analysis – Scatterplots – Bivariate correlation – X, Y – OK. Jiný způsob: Do dvourozměrného tečkového diagramu nakreslíme regresní přímku tak, že v tabulce 2D Scatterplots zvolíme Fit Linear, OK.

B.4. Výpočty pravděpodobností s využitím distribuční funkce binomického rozložení

Označme X náhodnou veličinu. Její distribuční funkci zavedeme vztahem $\Phi(x) = P(X \leq x)$. Pokud náhodná veličina X nabývá pouze konečně nebo spočetně mnoha hodnot, lze pomocí $\Phi(x)$ vyjádřit následující pravděpodobnosti:

a) $P(X = x) = P(X \leq x) - P(X \leq x - 1) = \Phi(x) - \Phi(x - 1)$;
b) $P(X \geq x) = 1 - P(X < x) = 1 - P(X \leq x - 1) = 1 - \Phi(x - 1)$;
c) $P(x_1 \leq X \leq x_2) = P(x_1 - 1 < X \leq x_2) = \Phi(x_2) - \Phi(x_1 - 1)$.

STATISTICA poskytuje hodnoty distribučních funkcí mnoha rozložení. Omezíme se na **binomické rozložení** (funkce IBinom(x, p, n), kde x ... počet úspěchů, p ... pravděpodobnost úspěchu v jednom pokusu, n ... celkový počet pokusů).

Vzorový příklad na binomické rozložení: Pojišťovna zjistila, že 12% pojistných událostí je způsobeno vloupáním. Jaká je pravděpodobnost, že mezi 30 náhodně vybranými pojistnými událostmi bude způsobeno vloupáním a) nejvýše 6, b) aspoň 6, c) právě 6, d) od dvou do pěti?

Řešení:
X ... počet pojistných událostí způsobených vloupáním , $n = 30$, $p = 0{,}12$.
ad a) $P(X \leq 6) = \Phi(6) = 0{,}9393$,
ad b) $P(X > 6) = 1 - P(X \leq 5) = 1 - \Phi(5) = 0{,}1431$,
ad c) $P(X = 6) = \Phi(6) - \Phi(5) = 0{,}0825$,
ad d) $P(2 \leq X \leq 5) = \Phi(5) - \Phi(1) = 0{,}7469$.

Postup ve STATISTICE: Otevřeme nový datový soubor se čtyřmi proměnnými a o jednom případu.

Řešení:
Do Long Name 1. proměnné napíšeme =IBinom(6;0,12;30).
Do Long Name 2. proměnné napíšeme =1-IBinom(5;0,12;30).
Do Long Name 3. proměnné napíšeme =IBinom(6;0,12;30)-IBinom(5;0,12;30).
Do Long Name 4. proměnné napíšeme =IBinom(5;0,12;30)-IBinom(1;0,12;30).
(Do Lange Name proměnné vstoupíme tak, že v datovém okně 2× klikneme myší na název proměnné.)

Kreslení grafů distribuční funkce a pravděpodobnostní funkce binomického rozložení

Vzorový příklad: Nakreslete graf distribuční funkce a pravděpodobnostní funkce náhodné veličiny $X \sim Bi(12; 0{,}3)$.

Postup ve STATISTICE: Vytvoříme nový datový soubor o 3 proměnných a 13 případech. První proměnnou nazveme X a uložíme do ní hodnoty 0, 1,..., 12 (do Long Name napíšeme =v0−1). Druhou proměnnou nazveme DF a uložíme do ní hodnoty distribuční funkce (do Long Name napíšeme příkaz =IBinom(x;0,3;12)). Třetí proměnnou nazveme PF a uložíme do ní hodnoty pravděpodobnostní funkce (do Long Name napíšeme příkaz =Binom(x;0,3;12)).

Graf distribuční funkce: Graphs – Scatterplots – Variables X, DF – OK – vypneme Linear fit – OK – 2× klikneme na pozadí grafu – Plot: General – zaškrtneme Line – Line Type: Step – OK.

Graf pravděpodobnostní funkce: Graphs – Scatterplots – Variables X, PF – OK – vypneme Linear fit – OK.

Podle tohoto návodu nakreslete grafy distribučních a pravděpodobnostních funkcí binomického rozložení pro různá n a p, např. $n = 5$, $p = 0{,}5$ (resp. 0,75) apod. Sledujte vliv parametrů na vzhled grafů.

B.5. Grafy hustot a distribučních funkcí, výpočet kvantilů

STATISTICA umí kreslit grafy hustot a distribučních funkcí mnoha spojitých rozložení a počítat kvantily těchto rozložení. Slouží k tomu Probability Calculator v menu Statistics. Zaměříme se na rozložení uvedená definici 8.6.

1. *Rovnoměrné spojité rozložení* $Rs(0,1)$
   Statistics – Probability Calculator – Distributions – Beta – shape 1 – napíšeme 1, shape 2 – napíšeme 1. STATISTICA vykreslí graf hustoty a distribuční funkce. Hodnotu α-kvantilu zjistíme tak, že do okénka označeného p napíšeme dané α a po kliknutí na Compute se v okénku Beta objeví hodnota tohoto kvantilu.

2. *Exponenciální rozložení* $Ex(\lambda)$
   Ve volbě Distributions vybereme Exponential a do okénka lambda napíšeme patřičnou hodnotu. Hodnotu α-kvantilu zjistíme tak, že do okénka označeného p napíšeme dané α a po kliknutí na Compute se v okénku exp objeví hodnota tohoto kvantilu.

3. *Normální rozložení* $N(\mu, \sigma^2)$
   Ve volbě Distributions vybereme Z (Normal), do okénka mean napíšeme hodnotu μ a do okénka st. dev. napíšeme hodnotu σ. Hodnotu α-kvantilu zjistíme tak, že do okénka označeného p napíšeme dané α a po kliknutí na Compute se v okénku X objeví hodnota tohoto kvantilu.

4. *Pearsonovo rozložení chí-kvadrát s n stupni volnosti* $\chi^2(n)$
   Ve volbě Distributions vybereme Chi 2 a do okénka df napíšeme patřičný počet stupňů volnosti. Hodnotu α-kvantilu zjistíme tak, že do okénka označeného p napíšeme dané α a po kliknutí na Compute se v okénku Chi 2 objeví hodnota tohoto kvantilu.

5. *Studentovo rozložení s n stupni volnosti* $t(n)$ Ve volbě Distributions vybereme t (Student) a do okénka df napíšeme patřičný počet stupňů volnosti. Hodnotu α-kvantilu zjistíme tak, že do okénka označeného p napíšeme dané α a po kliknutí na Compute se v okénku t objeví hodnota tohoto kvantilu.

6. *Fisherovo-Snedecorovo rozložení s* n_1 *a* n_2 *stupni volnosti* $F(n_1, n_2)$
   Ve volbě Distributions vybereme F (Fisher) a do okének df1 a df2 napíšeme počet stupňů volnosti čitatele a jmenovatele. Hodnotu α-kvantilu zjistíme tak, že do okénka označeného p napíšeme dané α a po kliknutí na Compute se v okénku F objeví hodnota tohoto kvantilu.

B.6. Intervaly spolehlivosti pro parametry normálního rozložení

1. Interval spolehlivosti pro střední hodnotu, když neznáme rozptyl: pro tuto situaci umí STATISTICA vypočítat meze intervalu spolehlivosti sama.

   Příklad: Při kontrole pěti balíčků cukru o deklarované hmotnosti 1000 g byly zjištěny tyto odchylky: -3, 2, -2, 0, 1. Odchylky považujeme za realizace náhodného výběru rozsahu 5 z rozložení $N(\mu, \sigma^2)$. Sestrojte 90% interval spolehlivosti pro μ.
   Návod: Vytvoříme nový datový soubor o jedné proměnné a pěti případech. Zapíšeme do něj uvedené odchylky. Statistics – Basic Statistics/Tables – Descriptive statistics – OK – Advanced - Variables v1, OK, zaškrtněte Conf. limits for mean – Interval 90%, Summary.

2. Ve všech ostatních případech postupujeme podle vzorců uvedených ve větách 12.9 a 12.13. Uveďme postup pro situaci, kdy hledáme interval spolehlivosti pro rozdíl středních hodnot dvou nezávislých normálně rozložených náhodných výběrů, když neznáme rozptyly, ale víme, že jsou shodné.

   Příklad: Na jisté velké americké univerzitě bylo v r. 1969 náhodně vybráno 5 profesorek a nezávisle na tom 5 profesorů a byl zjištěn jejich roční příjem v tisících dolarů. Ženy: 9 12 8 10 16, muži: 16 19 12 11 22. Předpokládáme, že uvedené hodnoty jsou realizace dvou nezávislých náhodných výběrů, první z rozložení $N(\mu_1, \sigma^2)$, druhý z rozložení $N(\mu_2, \sigma^2)$. Sestrojte 95% interval spolehlivosti pro rozdíl středních hodnot.
   Návod: Vytvoříme nový datový soubor o čtyřech proměnných (Plat, Sex, HorniMez, DolniMez) a 10 případech. Do proměnné Plat napíšeme příjmy žen, pak příjmy mužů. Do proměnné Sex napíšeme 5× jedničku a 5× dvojku (1=žena, 2=muž). Pomocí Descriptive statistics zjistíme průměry a rozptyly platů žen a mužů. (Výběr žen či mužů: viz cvičení 1, úkol 5.).
   Výsledky: $m_1 = 11$, $s_1^2 = 10$, $n_1 = 5$, $m_2 = 16$, $s_2^2 = 21{,}5$, $n_2 = 5$. Do Long Name proměnné DolniMez napíšeme vzorec pro dolní mez (viz věta 12.13 (b)):
   =11-16-sqrt((4*10+4*21,5)/8)*sqrt(1/5+1/5)*VStudent(0,975;8)
   Do proměnné DolniMez se 10× uloží hodnota $-10{,}79$. Do Long Name proměnné HorniMez napíšeme vzorec pro horní mez (viz věta 12.13 (b)):
   =11-16+sqrt((4*10+4*21,5)/8)*sqrt(1/5+1/5)*VStudent(0,975;8)
   Do proměnné HorniMez se 10× uloží hodnota 0,79. Znamená to, že s pravděpodobností aspoň 0,95 leží rozdíl středních hodnot platů žen a mužů v intervalu $(-10{,}79; 0{,}79)$. Tento výsledek však nemá praktický význam, protože rozsahy obou výběrů byly příliš malé.

Příklad: Vyřešte pomocí STATISTIKY příklad 12.16.
Návod: Vytvoříme nový datový soubor o třech proměnných (Leva, Prava, Rozdil) a šesti případech. Do prvních dvou proměnných zapíšeme zjištěné hodnoty. Do LongName proměnné Rozdil napíšeme =Leva - Prava a nyní postupujeme stejně jako v úkolu 1.

B.7. Testování hypotéz o parametrech normálního rozložení

Jednovýběrový *t*-test

Příklad: Při kontrole balicího automatu, který má plnit cukrem balíčky o hmotnosti 1000 g, byly při přesném převážení pěti balíčků zjištěny tyto odchylky (v gramech) od požadované hodnoty: 3, -2, 2, 0, 1. Na hladině významnosti 0,05 testujte hypotézu, že automat nemá systematickou odchylku od požadované hodnoty.

Návod pro provedení *t*-testu: Vytvořte soubor o jedné proměnné X a pěti případech. Do X zapište naměřené hodnoty. V menu Basic Statistics/Tables vyberte volbu t-test, single sample, OK, Variables X, zaškrtněte Test all means agains 0, Summary. Ve výstupní tabulce najdete hodnotu testového kritéria a p-hodnotu. Pokud p-hodnota nabude hodnoty $\leq \alpha$, pak se nulovou hypotézu zamítá na hladině významnosti α.

Dvouvýběrový *t*-test

Příklad: Na jisté velké americké univerzitě bylo v r. 1969 náhodně vybráno 5 profesorů a nezávisle na tom 5 profesorek a byl zjištěn jejich roční příjem v tisících dolarů.

Ženy:	9	12	8	10	16
Muži:	16	19	12	11	22

Na hladině významnosti 0,05 testujte hypotézu, že střední hodnota příjmu žen je stejná jako střední hodnota příjmu mužů.

Návod: Vytvořte soubor o dvou proměnných (Plat a Sex) a 10 případech. Do proměnné Plat napište příjmy žen a mužů a do proměnné Sex dejte 5× jedničku a 5× dvojku. V menu Basic Statistics/Tables vyberte volbu t-test, independent, by groups, OK, Variables – Grouping Sex, Dependent Plat, OK, Summary T-tests. Ve výstupní tabulce se nejprve podívejte na p-hodnotu pro test homogenity rozptylů. Je-li větší než zvolená hladinu významnosti, zjistěte hodnotu testového kritéria a p-hodnotu pro test shody středních hodnot. V opačném případě zaškrtněte v Options volbu t-test with separate variance estimates.

Párový *t*-test

Příklad: Na hladině významnosti 0,05 rozhodněte, zda se u osobního vozu určité značky při správném seřízení geometrie vozu sjíždějí obě přední pneumatiky stejně rychle. Bylo vybráno šest nových vozů a po určité době bylo zjištěno, o kolik mm se sjely jejich levé a pravé přední pneumatiky.

číslo automobilu	1	2	3	4	5	6
pravá pneumatika	1,8	1,0	2,2	0,9	1,5	1,6
levá pneumatika	1,5	1,1	2,0	1,1	1,4	1,4

Návod: Vytvořte soubor o dvou proměnných (Leva a Prava) a šesti případech. V menu Basic Statistics/Tables vyberte volbu t-test, dependent samples, OK, Variables Leva, Prava – Summary.

Závěr

Závěr

Učební text, který jste právě dočetli, byl určen k prvnímu seznámení s matematickou disciplinou nazývanou statistika. Autorským záměrem bylo ukázat vám, že statistika ve své popisné formě dokáže pomoci několika výstižných charakteristik zpřehlednit informace obsažené ve velkých datových souborech, zatímco ve své induktivní formě založené na počtu pravděpodobnosti slouží především jako nástroj rozhodování v situacích ovlivněných náhodou, kdy na základě znalosti náhodného výběru z určitého rozložení pravděpodobnosti usuzuje na vlastnosti tohoto rozložení.

V současnosti je statistika velice rozvinutá a důležitá věda, která se neustále doplňuje a rozšiřuje o nové poznatky. Z tohoto důvodu může být tento učební text jen značně omezeným úvodem, který však má dostatečnou oporu v obecných statistických principech. V seznamu literatury samozřejmě najdete knihy, které vám poslouží při prohlubování a rozšiřování vašich statistických znalosti, bez nichž se dnes neobejde žádný absolvent ekonomicky zaměřené vysoké školy. Od ekonoma se totiž očekává, že bude rozhodovat nejenom na základě svých zkušenosti, ale především na základě matematických a statistických analýz. Proto musí být schopen sám provést jednodušší analýzy a u těch složitějších najít společnou řeč se statistiky, aby jim mohl zadávat úkoly a správně interpretovat výsledky těchto analýz.

Jak jste již zjistili, použiti statistického programového systému STATlSTICA osvobozuje uživatele od namáhavých úkonů, jako je vyhledávání v datech, jejich třídění, sumarizace a grafické znázornění. Dbejte však na to, aby data byla do počítače vkládána pečlivě a vždy byla podrobena kontrole. Např. je užitečné pro každou proměnnou vypočítat minimum, maximum, medián, kvartilovou odchylku, vykreslit sloupkový diagram, dvourozměrný tečkový diagram apod. Při zpracování dat rozhodně používejte jen ty metody, kterým dobře rozumíte a jejichž výsledky umíte interpretovat. Systém STATlSTICA obsahuje velké množství metod, jejichž neadekvátní aplikace může vést k zavádějícím či dokonce chybným závěrům.

Po úspěšném zvládnuti předmětu „Statistika“ se před vámi otevírají značné možnosti, jak efektivně získávat informace obsažené v datech a využívat je ve své každodenní práci.