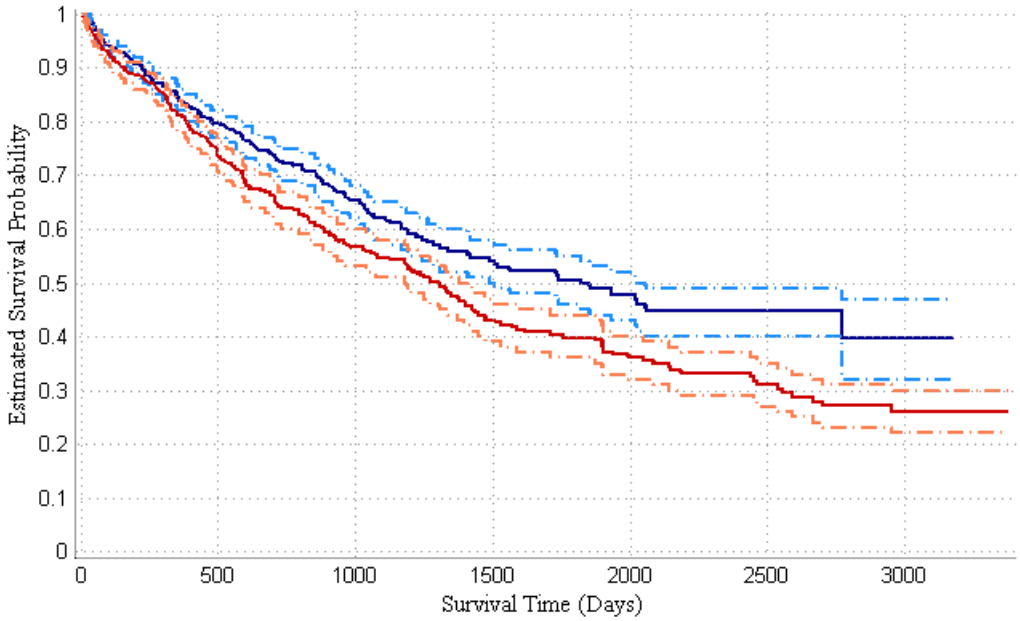
Vysoká škola báňská - Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra aplikované matematiky

# Biostatistika

## Cvičení - pracovní listy




Martina Litschmannová, Kateřina Janurová

5.května 2013

Jméno: .....

# 1. Kombinatorika

Kombinatorika je vstupní branou do teorie pravděpodobnosti. Zabývá se různými způsoby výběru prvků z daného souboru.

- **Kombinatorické pravidlo o součinu**

  Počet všech uspořádaných $k$-tic, jejichž první člen lze vybrat $n_1$ způsoby, druhý člen po výběru prvního členu $n_2$ způsoby atd. až $k$-tý člen po výběru všech předcházejících členů $n_k$ způsoby, je roven $n_1 \cdot n_2 \cdot \ldots \cdot n_k$.

- **Kombinatorické pravidlo součtu**

  Jsou-li $A_1, A_2, \ldots, A_n$ konečné množiny, které mají po řadě $p_1, p_2, \ldots, p_n$ prvků, a jsou-li každé dvě disjunktní, pak počet prvků množiny $A_1 \bigcup A_2 \bigcup \ldots \bigcup A_n$ je roven $p_1 + p_2 + \ldots + p_n$.

- **Typy výběrů**

| | | | |
|---|---|---|---|
| Uspořádané výběry | Bez opakování | Variace bez opakování | $V(n,k) = \frac{n!}{(n-k)!}$ |
| | | Permutace bez opakování | $P(n) = V(n,n) = n!$ |
| | S opakováním | Variace s opakováním | $V^*(n,k) = n^k$ |
| | | Permutace s opakováním | $P^*(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1! n_2! \ldots n_k!}$ |
| Neuspořádané výběry | Bez opakování | Kombinace bez opakování | $C(n,k) = \frac{n!}{(n-k)!k!}$ |
| | S opakováním | Kombinace opakováním | $C^*(n,k) = \frac{(n+k-1)!}{(n-1)!k!}$ |

- **Klasická definice pravděpodobnosti**

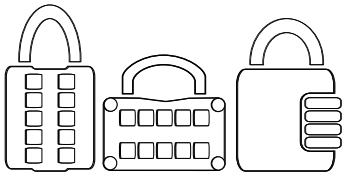
$$P(A) = \frac{m}{n}$$

kde: $m$ ... počet výsledků příznivých jevu $A$
$n$ ... počet všech možných výsledků

## 1.1 Příklady

*Příklady 1-10 jsou upravené příklady z přednášky prof. Adama Plocki (Kombinatorika a pravděpodobnost kolem nás a ve škole; Seminář matematiky pro středoškolské profesory a učitele základních škol; Ostrava; 2014)*

1. V prodejně mají k dispozici tři typy zámků. Pro otevření prvního zámku je nutno zmáčknout čtyři z deseti tlačítek označených číslicemi 0 až 9. (Na pořadí nezáleží - tlačítka zůstávají zmáčknuta.) Druhý zámek se otevře pokud zmáčkneme šest tlačítek z deseti. Pro otevření třetího zámku je nutno nastavit správnou kombinaci na čtyřech kotoučích. Který z těchto zámků nejlépe chrání před zloději?



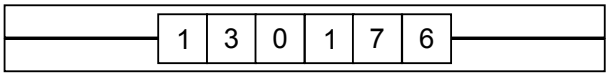
Řešení:

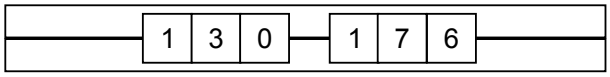
| | Počet možností | | Pravděpodobnost otevření | |
|---|---|---|---|---|
| 1. zámek: | $C(10,4)$ | $\binom{n}{k} = \binom{n}{n-k}$ | $1/\binom{10}{4} \cong 0,5\%$ | rovnocenné |
| 2. zámek: | $C(10,6)$ | | $1/\binom{10}{6} \cong 0,5\%$ | |
| 3. zámek: | $V^*(10,4)$ | | $1/(10^4) \cong 0,01\% \Rightarrow$ nejlepší | |

Nejlépe chrání před zloději třetí zámek.

[třetí]

2. V prodejně nabízejí dva druhy zamykání kufříku. První kufřík se zamyká šifrou, která se skládá z šesti číslic. Druhý kufřík se zamyká dvěma zámky, které se otevírají současně. Šifra každého z nich se skládá ze tří číslic. Určete pro každý kufřík pravděpodobnost otevření zlodějem při prvním pokusu. Který typ zámku je bezpečnější?




Řešení:

$\uparrow$
$V^*(10,6) = 10^6$
$P(\text{otevření}) = \frac{1}{10^6}$
Oba typy jsou stejně bezpečné.

$\uparrow$
$V^*(10,3) \cdot V^*(10,3) = 10^3 \cdot 10^3 = 10^6$
$P(\text{otevření}) = \frac{1}{10^6}$
Myšlenka: zmenšování mezery mezi zámky.

[oba jsou stejně bezpečné, pravděpodobnost otevření je $10^{-6}$]

3. V urně je 40 koulí - 2 červené a 38 bílých. Z urny náhodně vytáhneme 2 koule. S jakou pravděpodobností budou obě červené?

Řešení:

jev $A \ldots$ vytáhneme dvě červené koule

počet všech příznivých dvojic (dvě červené) $= 1 \quad \Rightarrow \quad P(A) = \frac{1}{780} \cong 0,0013 (\cong 0,13\%)$

počet všech možných dvojic: $C(40,2) = \binom{40}{2} = 780$

Dvě červené koule vytáhneme s pravděpodobností $0,13\%$, tzn. je „prakticky nemožné“, aby obě koule byly červené. (Jev „prakticky nemožný“... pravděpodobnost jevu $< 0,05$ $(5\%)$.)

$[0,0013]$

4. Student si měl ke zkoušce připravit odpovědi na 40 otázek. Na dvě otázky, které mu dal zkoušející, neuměl odpovědět a tak řekl „To mám smůlu! To jsou jediné dvě otázky, na které neumím odpovědět.“ S jakou pravděpodobností mluví pravdu?

   Řešení:

   Matematizace odpovídá př. 3 ⇒ Je „prakticky nemožné“, že student mluví pravdu.

   [0, 0013]

5. Test z chemie žák složí, pokud v seznamu 40 chemických sloučenin podtrhne jediné dva aldehydy, které v seznamu jsou. Jaká je pravděpodobnost, že test složí žák, který provede výběr sloučenin náhodně?

   Řešení:

   Matematizace odpovídá př. 3 ⇒ Je „prakticky nemožné“, aby žák složil test bez potřebných znalostí.

   [0, 0013]

6. Ze zahraničí se vracela skupina 40 turistů a mezi nimi byli 2 pašeráci. Na hranici celník 2 turisty vyzval k osobní prohlídce a ukázalo se, že oba dva jsou pašeráci. Zbylí turisté na to reagovali: „Celník měl opravdu štěstí!“, „Pašeráky někdo udal!“, .... Jak se postavit k těmto výrokům? Je oprávněné podezření, že pašeráky někdo udal?

   Řešení:

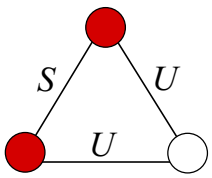
   Matematizace odpovídá př. 3 ⇒ Je „prakticky jisté“, že pašeráky někdo udal.

   [pravděpodobnost, že pašeráky někdo udal: 0, 9987 ]

7. Z urny se třemi koulemi, dvěma červenými a jednou bílou, budou současně vybrány dvě koule. Student a učitel uzavřou sázku. Pokud budou obě koule stejné barvy, vyhraje student. Pokud budou mít koule různou barvu, vyhraje učitel. Je hra férová? Jaké jsou pravděpodobnosti výhry učitele a studenta?

   Řešení:

   jev $S$ ... vyhraje student
   jev $U$ ... vyhraje učitel

   počet všech možností = 3
   počet možností příznivých pro studenta = 1
   počet možností příznivých pro učitele = 2
   $P(S) = 1/3, P(U) = 2/3$

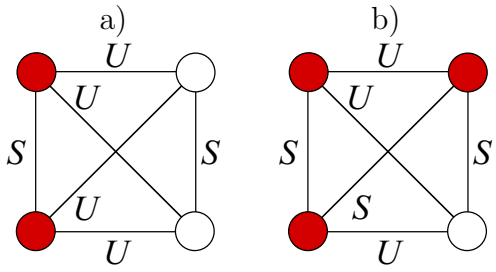
   Hra není férová, pravděpodobnost výhry studenta je 0,33, zatímco pravděpodobnost výhry učitele je 0,67.

   [hra není férová; $S : 0, 33; U : 0, 67$]

8. Hra popsaná v příkladu 7 nebyla férová. Jakou kouli (červenou nebo bílou) musíme do urny přidat, aby hra férová byla?

   Řešení:

   jev $S$ ... vyhraje student
   jev $U$ ... vyhraje učitel




| | a) | b) |
|---|---|---|
| počet všech možností | 6 | 6 |
| počet příznivých možností ($S$) | 2 | 3 |
| počet příznivých možností ($U$) | 4 | 3 |

   a) $P(U) = 2/3 = 2 \cdot 1/3 = 2 \cdot P(S)$
   b) $P(U) = P(S) = 1/2$

   Aby hra byla férová, musíme do urny přidat červenou kouli.

   [červenou]

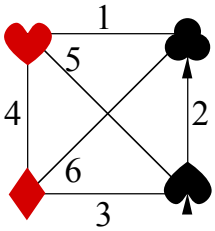
9. Chcete hrát Člověče nezlob se, ale ztratila se hrací kostka. Čím a jak lze nahradit hrací kostku, máte-li k dispozici hrací karty (balíček 32 karet) a 4 různobarevné kuličky?

   Řešení:

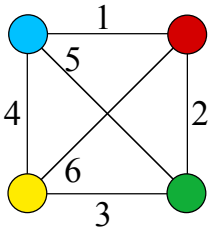
   např.

   a) losování 1 karty ze 6

   b) losování 2 karet ze 4

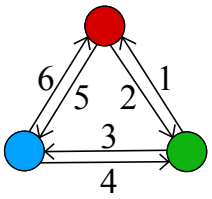



   c) losování 2 kuliček ze 4




10. Chcete hrát Člověče nezlob se, ale ztratila se hrací kostka. Jak lze nahradit hrací kostku, máte-li k dispozici 3 různobarevné kuličky?

    Řešení:




    Losování dvou kuliček ze tří s ohledem na pořadí.

11. V prodejně vozů Škoda mají v měsíci únoru prodejní akci. Ke standardnímu vybavení nabízejí 3 položky z nadstandardní výbavy zdarma. Nadstandardní výbava zahrnuje 7 položek:

    - tempomat, vyhřívání sedadel, zadní airbagy, xenonová světla, stropní okénko, bezpečnostní zámek převodovky, speciální odolný metalízový lak.

    Kolik možností má zákazník, jak zvolit 3 položky z nadstandardní výbavy?

    Řešení:

    Záleží na pořadí? Ne ⇒ kombinace
    Mohou se prvky ve výběru opakovat? Ne ⇒ kombinace bez opakování
    Z kolika prvků vybíráme? $n = 7$
    Kolik prvků vybíráme? $k = 3$

    $$C(7,3) = \binom{7}{3} = \frac{7!}{3!\cdot 4!} = \frac{7\cdot 6\cdot 5}{6} = 35$$

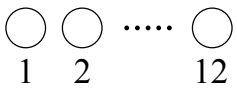
    Zákazník má 35 možností.

    [35]

12. Při zkoušce si do 5. řady sedlo 12 studentů. Zkoušející chce určit sám, jak tyto studenty v řadě rozesadit.

    a) Kolik je možností jak studenty rozesadit?

    Řešení:

    Záleží na pořadí? Ano ⇒ variace nebo permutace
    Z kolika prvků vybíráme? $n = 12$
    Kolik prvků vybíráme? $k = 12$ ⇒ permutace („přesmyčky“)
    Mohou se prvky ve výběru opakovat? Ne ⇒
    permutace bez opakování

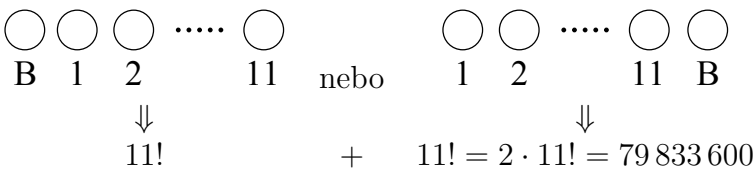
    $$P(12) = 12! = 479\,001\,600$$

    Je 479 001 600 možností jak rozesadit studenty.

    [479 001 600]

    b) Student Brahý žádá, aby mohl sedět na kraji a odejít dříve, aby stihl vlak. Kolik je možností jak studenty rozesadit, chce-li zkoušející vyhovět požadavku studenta Brahého?

    Řešení:

    $11!$ + $11! = 2 \cdot 11! = 79\,833\,600$

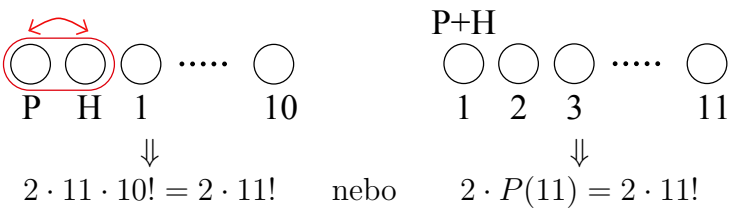
    Je 79 833 600 možností jak rozesadit studenty, sedí-li student Brahý na kraji.

    [79 833 600]

c) Kolik je možností jak studenty rozesadit, nesmějí-li Pažout a Horáček sedět vedle sebe?

Řešení:

Sedí vedle sebe:




$2 \cdot 11 \cdot 10! = 2 \cdot 11!$ nebo $2 \cdot P(11) = 2 \cdot 11!$

Nesedí vedle sebe: $12! - 2 \cdot 11! = 11!(12 - 2) = 11! \cdot 10 = 399\ 168\ 000$

Je 399 168 000 možností jak rozesadit studenty, nesmějí-li Pažout a Horáček sedět vedle sebe.

[399 168 000]

13. Kolik anagramů lze vytvořit ze slova STATISTIKA?

Řešení:

Záleží na pořadí? Ano ⇒ variace nebo permutace
Z kolika prvků vybíráme? $n = 10$
Kolik prvků vybíráme? $k = 10$ ⇒ permutace („anagram“ = „přesmyčka“)
Mohou se prvky ve výběru opakovat? Ano ⇒ permutace s opakováním

| S | T | A | I | K | celkem |
|---|---|---|---|---|---|
| 2 | 3 | 2 | 2 | 1 | 10 |

$$P^*(2,3,2,2,1) = \frac{10!}{2!3!2!2!1!} = \frac{10!}{48} = 75\,600$$

Ze slova STATISTIKA lze vytvořit 75 600 anagramů.

[75 600]

14. V Tescu dostali nové zboží – 6 druhů chlapeckých trik. Od každého druhu mají alespoň 7 kusů. Maminka chce synovi koupit 4 trika. Kolik je možností, jak je vybrat

a) mají-li být všechna různá?

Řešení:

Záleží na pořadí? Ne ⇒ kombinace
Mohou se prvky ve výběru opakovat? Ne ⇒ kombinace bez opakování
Z kolika prvků vybíráme? $n = 6$
Kolik prvků vybíráme? $k = 4$

$$C(6,4) = \binom{6}{4} = \frac{6!}{2! \cdot 4!} = \frac{6 \cdot 5}{2} = 15$$

Je 15 možností jak vybrat trika, mají-li být všechna různá.

[15]

b) připouští-li, že mohou být všechna stejná?

Řešení:

Záleží na pořadí? Ne ⇒ kombinace
Mohou se prvky ve výběru opakovat? Ano ⇒ kombinace s opakováním
Z kolika prvků vybíráme? $n = 6$
Kolik prvků vybíráme? $k = 4$

$$C^*(6,4) = \binom{6+4-1}{4} = \binom{9}{4} = \frac{9!}{5!\cdot 4!} = 126$$

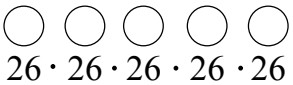
Je 126 možností jak vybrat trika, mohou-li být stejná.

[126]

15. Kolik hesel délky 5 můžeme vytvořit ze znaků abecedy

a) nejsou-li rozlišována velká a malá písmena?

Řešení:




Záleží na pořadí? Ano ⇒ variace nebo permutace
Z kolika prvků vybíráme? $n = 26$
Kolik prvků vybíráme? $k = 5$ ⇒ variace
Mohou se prvky ve výběru opakovat? Ano ⇒ variace s opakováním

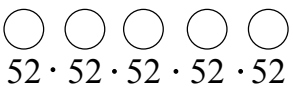
$$V^*(26,5) = 26^5 = 11\,881\,376$$

Nerozlišujeme-li velká a malá písmena, můžeme ze znaků anglické abecedy vytvořit celkem 11 881 376 hesel.

[11 881 376]

b) jsou-li rozlišována velká a malá písmena?

Řešení:




Záleží na pořadí? Ano ⇒ variace nebo permutace
Z kolika prvků vybíráme? $n = 52$
Kolik prvků vybíráme? $k = 5$ ⇒ variace
Mohou se prvky ve výběru opakovat? Ano ⇒ variace s opakováním

$$V^*(52,5) = 52^5 = 380\,204\,032$$

Rozlišujeme-li velká a malá písmena, můžeme ze znaků anglické abecedy vytvořit celkem 380 204 032 hesel, tj. 32 krát více než v předchozím případě.

[380 204 032]

# 2. Pravděpodobnost

**Náhodný pokus** je každý konečný děj, jehož výsledek není předem jednoznačně určen podmínkami, za nichž probíhá, a který je, alespoň teoreticky, neomezeně opakovatelný.

Množinu všech možných výsledků $\{\omega\}$ daného pokusu označujeme pojmem **základní prostor** a značíme jako $\Omega$

Prvky, popř. jednoprvkové podmnožiny, základního prostoru jsou **elementární jevy**.

**Jevem** $A$ je libovolná podmnožina základního prostoru. Pro náhodné jevy platí algebraické zákony a rovnosti stejné jako pro množiny.

**Úplná množina vzájemně disjunktních jevů** je množina po dvou disjunktních jevů $A_1, A_2, A_3, \ldots, A_n$, jejichž sjednocení tvoří množinu $\Omega$.

**Jevové pole** $\mathcal{A}$ je neprázdný systém podmnožin základního prostoru uzavřený vůči doplňku a vůči sjednocení ($\sigma$-algebra na $\Omega$).

**Vlastnosti pravděpodobnostního prostoru** jsou dány Kolmogorovovým axiomatickým systémem.

**Podmíněná pravděpodobnost** je pravděpodobnost výskytu jevu za podmínky, že nastal určitý jev, který není nemožný.

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

Jesliže platí $P(A \cap B) = P(A) \cdot P(B)$ nebo $P(B) = 0$, pak nazýváme **jevy** $A, B$ **nezávislé**.

**Věta o úplné pravděpodobnosti** nám dává návod, jak určit pravděpodobnost jevu $A$, o kterém je známo, že může nastat pouze současně s některým z jevů $B_1, B_2, \ldots, B_n$, které tvoří úplnou množinu disjunktních jevů.

$$P(A) = \sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)$$

**Bayesova věta** nám umožňuje spočítat podmíněné pravděpodobnosti jednotlivých jevů této úplné množiny za předpokladu, že nastal jev $A$.

$$P(B_k|A) = \frac{P(A|B_k) \cdot P(B_k)}{\sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)}$$

**Vlastnosti pravděpodobnosti:**

- $0 \leq P(A) \leq 1$
- $P(\emptyset) = 0$
- $P(\bar{A}) = 1 - P(A)$
- $A \subset B \Rightarrow P(A) \leq P(B)$
- $P(B - A) = P(B) - P(A \cap B)$, speciálně pro $A \subset B \Rightarrow P(B - A) = P(B) - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, speciálně pro $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$
- $P(A \cup B) = 1 - P(\overline{A \cup B}) = 1 - P(\bar{A} \cap \bar{B})$, z 1. de Morganova zákona
- $P(A \cap B) = P(A) \cdot P(B|A)$, speciálně pro nezávislé jevy $A, B \Rightarrow P(A \cap B) = P(A) \cdot P(B)$
- $P(A \cap B) = 1 - P(\overline{A \cap B}) = 1 - P(\bar{A} \cup \bar{B})$, z 2. de Morganova zákona

## 2.1 Test z teorie

1. Určete, která z následujících tvrzení jsou pravdivá.

   a) Klasická definice pravděpodobnosti vychází ze stability relativních četností.

   b) Kolmogorovovy axiomy pravděpodobnosti udávají návod ke stanovení pravděpodobnosti elementárních jevů.

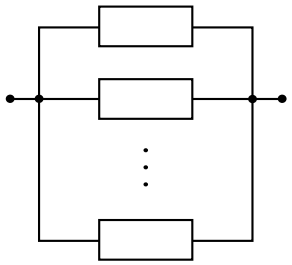
   c) Je-li pravděpodobnost jevu $A$ rovna 0,75, pak pravděpodobnost podjevu jevu $A$ je nejvýše 0,75.

   d) Jestliže pravděpodobnosti dvou jevů z jevového pole $\mathcal{A}$ jsou 0,7 a 0,5, pak tyto jevy nejsou disjunktní.

   e) Pravděpodobnost, že při deseti hodech mincí padne desetkrát po sobě „panna“ je menší než pravděpodobnost, že při deseti hodech klasickou kostkou padne desetkrát po sobě sudé číslo.

2. Pravděpodobnost poruchy každé součástky je $p$. Předpokládejme, že součástky pracují nezávisle na sobě. Určete pravděpodobnost poruchy bloku složeného z 10 paralelně zapojených součástek. (Je-li funkční alespoň jedna součástka, blok funguje.)

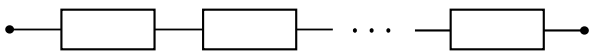
   a) $p/10$
   b) $10 \cdot p$
   c) $10/p$
   d) $p^{10}$
   e) $1 - p^{10}$
   f) $(1-p)^{10}$
   g) $1 - (1-p)^{10}$
   h) $(1-p)/10$



$\overline{S}$ ... systém nefunguje
$p$ ... součástka nefunguje
$P(S) = \prod_{i=1}^{10} p = p^{10}$

3. Pravděpodobnost poruchy každé součástky je $p$. Předpokládejme, že součástky pracují nezávisle na sobě. Určete pravděpodobnost poruchy bloku složeného z 10 sériově zapojených součástek. (Je-li porouchaná alespoň jedna součástka, blok nefunguje.)

   a) $p/10$
   b) $10 \cdot p$
   c) $10/p$
   d) $p^{10}$
   e) $1 - p^{10}$
   f) $(1-p)^{10}$
   g) $1 - (1-p)^{10}$
   h) $(1-p)/10$



$S$ ... systém funguje
$p$ ... součástka nefunguje
$P(S) = \prod_{i=1}^{10} (1-p) = (1-p)^{10}$ $P(\overline{S} = 1-(1-p)^{10})$

4. Podmíněná pravděpodobnost $P(A|B)$ je rovna:

   a) $P(A \cap B) \cdot P(B)$
   b) $\dfrac{P(A \cap B)}{P(A)}$, $P(A) \neq 0$
   c) $P(A \cap B) \cdot P(A)$
   d) $\dfrac{P(A \cap B)}{P(B)}$, $P(B) \neq 0$

5. Mějme jevy $A$ a $B$. Pravděpodobnost jevu $A$ je $P(A)$ a pravděpodobnost jevu $B$ je $P(B)$. Pravděpodobnost sjednocení jevů $A$ a $B$ je rovna:

   a) $P(A) + P(B)$
   b) $P(A) \cdot P(A)$
   c) $P(A) + P(B) - P(A \cap B)$
   d) $P(A|B) \cdot P(B)$

6. Mějme nezávislé jevy $A$ a $B$. Pravděpodobnost jevu $A$ je $P(A)$ a pravděpodobnost jevu $B$ je $P(B)$. Pravděpodobnost sjednocení jevu $A$ a $B$ je rovna:

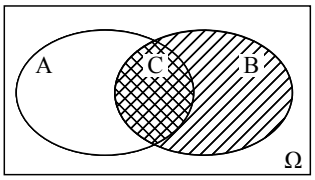
   a) $P(A) + P(B)$
   b) $P(A) \cdot P(A)$
   c) $P(A) + P(B) - P(A \cap B)$, (lze zjednodušit, jsou-li jevy disjunktní)
   d) $P(A|B) \cdot P(B)$

7. Mějme disjunktní jevy $A$ a $B$. Pravděpodobnost jevu $A$ je $P(A)$ a pravděpodobnost jevu $B$ je $P(B)$. Pravděpodobnost průniku jevu $A$ a $B$ je rovna:

   a) $P(A) + P(B)$
   b) $P(A) \cdot P(A)$
   c) $P(A) + P(B) - P(A \cap B)$
   d) $0$

8. Mějme jevy $A$ a $B$. Jev $C$ je průnikem jevů $A$ a $B$. Pravděpodobnost jevu $A$ je $P(A)$ a pravděpodobnost jevů $B$ je $P(B)$. Pravděpodobnost sjednocení jevu $B$ a $C$ vyjádřena pomocí pravděpodobností jevů $A$ a $B$ je rovna:

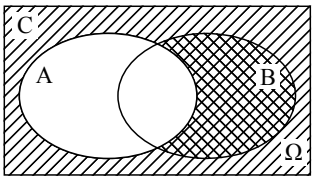
   a) $P(A)$
   b) $P(B)$
   c) $P(B) \cdot (1 + P(A))$
   d) $P(B) \cdot (1 - P(A))$
   e) $P(B) \cdot (1 + P(A|B))$
   f) $P(B) \cdot (1 - P(A|B))$



$C \subset B \Rightarrow C \cup B = B$

9. Mějme nezávislé jevy $A$ a $B$. Jev $C$ je doplněk jevu $A$. Pravděpodobnost jevu $A$ je $P(A)$ a pravděpodobnost jevu $B$ je $P(B)$. Pravděpodobnost průniku jevu $B$ a $C$ vyjádřena pomocí pravděpodobností jevů $A$ a $B$ je rovna:

   a) $P(A)$
   b) $P(B)$
   c) $P(B) \cdot (1 + P(A))$
   d) $P(B) \cdot (1 - P(A))$
   e) $P(B) \cdot (1 + P(A|B))$



$C \cap B = B - A$

10. Vyberte 3 Kolmogorovovy axiomy pravděpodobnosti.

    a) Pravděpodobnost každého jevu $A$ je nezáporné reálné číslo.
    b) Pravděpodobnost každého jevu $A$ je menší než 1.
    c) Pravděpodobnost jistého jevu $\Omega$ je rovna nule.
    d) Pravděpodobnost jistého jevu $\Omega$ je rovna jedné.
    e) Pravděpodobnost sjednocení konečného počtu vzájemně disjunktních jevů je rovna součtu jejich pravděpodobností.
    f) Pravděpodobnost sjednocení jevů je rovna součtu jejich pravděpodobností.

## 2.2 Příklady

1. Určete pravděpodobnost, že při hodu 20stěnnou spravedlivou (férovou) kostkou padne číslo větší než 14.

   Řešení:

   $A \ldots$ padne číslo větší než 14

   $$P(A) = \frac{6}{20} = 0,3$$

   Pravděpodobnost, že padne číslo větší než 14 je 0,3.

   [0,3]

2. Určete pravděpodobnost, že při hodu 20stěnnou kostkou padne číslo větší než 14, víte-li, že sudá čísla padají 2x častěji než lichá.

   Řešení:

   $S \ldots$ padne sudé číslo
   $L \ldots$ padne liché číslo

   $$\underbrace{\underbrace{20\text{x}}_{10 \text{ sudých čísel}} \quad \underbrace{10\text{x}}_{10 \text{ lichých čísel}}}_{30\text{x}=1 \Rightarrow x=\frac{1}{30}} \quad \Rightarrow \quad \begin{matrix} P(S) = 2\text{x} = \frac{2}{30} \\ P(L) = 1\text{x} = \frac{1}{30} \end{matrix}$$

   $A \ldots$ padne číslo větší než 14

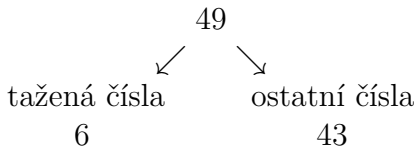
   $$P(A) = 3 \cdot P(S) + 3 \cdot P(L) = 3 \cdot \frac{2}{30} + 3 \cdot \frac{1}{30} = \frac{9}{30} = 0,3$$

   Pravděpodobnost, že padne číslo větší než 14 je 0,3.

   [0,3]

3. Určete pravděpodobnost, že ve sportce uhodnete 4 čísla. (Ve sportce se losuje 6 čísel ze 49.)

   Řešení:

   $A \ldots$ uhodnete 4 čísla z 6 tažených

   $$P(A) = \frac{C(6,4) \cdot C(43,2)}{C(49,6)} \cong 0,001$$

   Pravděpodobnost, že uhodnete 4 čísla z 6 tažených je 0,001. (Je téměř nemožné uhodnout 4 čísla ve sportce.)

   [0,001]

4. Z abecedního seznamu studentů zapsaných na dané cvičení vybere učitel prvních 12 a nabídne jim sázku: „Pokud se každý z Vás narodil v jiném znamení zvěrokruhu, dám každému z Vás 100 Kč. Pokud jsou však mezi Vámi alespoň dva studenti, kteří se narodili ve stejném znamení, dá mi každý z Vás 100 Kč.“ Vyplatí se studentům sázku přijmout? S jakou pravděpodobností studenti vyhrají?

   Řešení:

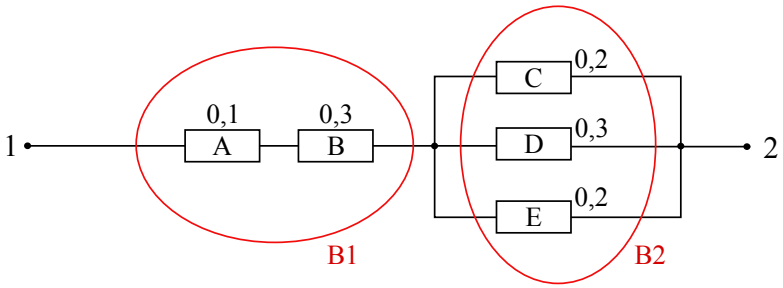
   $A \dots$ všichni studenti mají různá znamení

| | $P(A)$ | $P(\overline{A})$ |
|---|---|---|
| 1 student | $1$ | $0$ |
| 2 studenti | $1 \cdot \frac{11}{12}$ | $\frac{1}{12}$ |
| 3 studenti | $1 \cdot \frac{11}{12} \cdot \frac{10}{12}$ | $1 - 1 \cdot \frac{11}{12} \cdot \frac{10}{12}$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| 12 studentů | $1 \cdot \frac{11}{12} \cdot \frac{10}{12} \cdot \dots \cdot \frac{1}{12} = \frac{12!}{12^{12}} \cong 5 \cdot 10^{-5}$ | $\cong 1$ |

   Studentům se nevyplatí přijmout sázku, vyhrají s pravděpodobnostní $5 \cdot 10^{-5}$. (Je prakticky nemožné, že vyhrají.)

   $[5 \cdot 10^{-5}\ ]$

5. Spočtěte pravděpodobnost toho, že z bodu 1 do bodu 2 bude protékat elektrický proud, je-li část el. obvodu včetně pravděpodobnosti poruch jednotlivých součástek vyznačen na následujícím obrázku. (Poruchy jednotlivých součástek jsou na sobě nezávislé.)

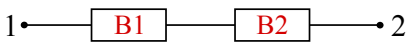



   Řešení:

   $B1 \dots$ blok B1 funguje
   $B2 \dots$ blok B2 funguje
   $A(B, \dots, E) \dots$ součástka A (B,...,E) funguje
   $S \dots$ systém funguje

   $P(B1) = P(A \cap B) = P(A) \cdot P(B) = (1 - 0,1) \cdot (1 - 0,3) = 0,9 \cdot 0,7 = 0,630$

   $P(\overline{B2}) = P(\overline{C} \cap \overline{D} \cap \overline{E}) = P(\overline{C}) \cdot P(\overline{D}) \cdot P(\overline{E}) = 0,2 \cdot 0,3 \cdot 0,2 = 0,012$



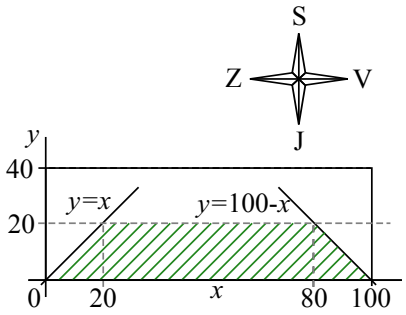

   $P(S) = P(B1 \cap B2) = P(B1) \cdot P(B2) = 0,630 \cdot (1 - 0,012) = 0,630 \cdot 0,988 \cong 0,622$

   Pravděpodobnost, že z bodu 1 do bodu 2 protéká elektrický proud je 0,622.

   [0,622]

6. Ohrada má obdélníkový tvar, východní a západní stěna mají délku 40 m, jižní a severní pak 100 m. V této ohradě běhá kůň. Jaká je pravděpodobnost, že je k jižní stěně blíž než ke zbývajícím třem?

   Řešení:




$y \ldots$ vzdálenost od jižní stěny
$x \ldots$ vzdálenost od západní stěny
$100 - x \ldots$ vzdálenost od východní stěny
$y < x \wedge y < 100 - x \wedge y < 20$
$A \ldots$ kůň bude blíže k jižní stěně než ke zbývajícím

$$P(A) = \frac{|A|}{|\Omega|} = \frac{20 \cdot 100 - 20 \cdot 20}{40 \cdot 100} = \frac{1\,600}{4\,000} = \frac{2}{5} = 0,4$$

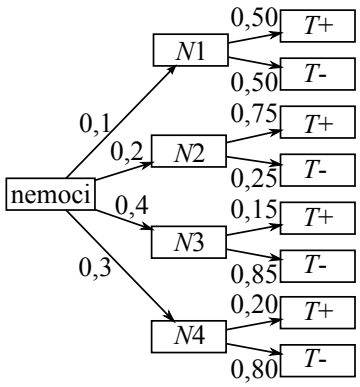
   Pravděpodobnost, že kůň je blíže k jižní stěně je 0,4.

   [0,4]

7. U pacienta je podezření na jednu ze čtyř vzájemně se vylučujících nemocí - $N1, N2, N3, N4$ s pravděpodobností výskytu $P(N1) = 0,1, P(N2) = 0,2, P(N3) = 0,4, P(N4) = 0,3$. Laboratorní zkouška $A$ je pozitivní v případě první nemoci v 50 % případů, u druhé nemoci v 75 % případů, u třetí nemoci v 15 % případů a u čtvrté v 20 % případů. Jaká je pravděpodobnost, že výsledek laboratorní zkoušky bude pozitivní?

   Řešení:

$N1(N2, \ldots, N4) \ldots$ pacient má nemoc $N1(N2, \ldots, N4)$
$T+ \ldots$ test vyšel pozitivní
$T- \ldots$ test vyšel negativní




$$P(T+) =$$
$$= P(N1 \cap T+) + P(N2 \cap T+) + P(N3 \cap T+) + P(N4 \cap T+) =$$
$$= P(T+|N1) \cdot P(N1) + \cdots + P(T+|N4) \cdot P(N4) =$$
$$= 0,1 \cdot 0,5 + 0,2 \cdot 0,75 + 0,4 \cdot 0,15 + 0,3 \cdot 0,2 =$$
$$= 0,05 + 0,15 + 0,06 + 0,06 = 0,32$$

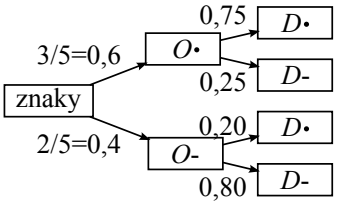
   Pravděpodobnost, že výsledek laboratorní zkoušky bude pozitivní je 0,32.

   [0,32]

8. Telegrafické znaky se skládají ze signálů „tečka“, „čárka“. Je statisticky zjištěno, že se zkomolí 25 % sdělení „tečka“ a 20 % signálů „čárka“. Dále je známo, že signály se používají v poměru 3:2. Určete pravděpodobnost, že byl přijat správně signál, jestliže byl přijat signál „tečka“.

   Řešení:

   $O\bullet$ ... byla odeslána tečka
   $O-$ ... byla odeslána čárka
   $D\bullet$ ... byla detekována tečka
   $D-$ ... byla detekována čárka




$$P(O\bullet|D\bullet) = \frac{P(O\bullet \cap D\bullet)}{D\bullet} = \frac{0,6\cdot 0,75}{0,6\cdot 0,75 + 0,4\cdot 0,2} = \frac{0,45}{0,53} \cong 0,849$$

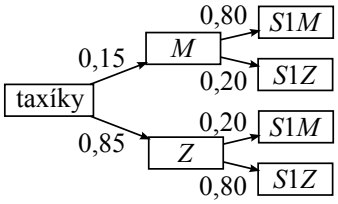
   Pravděpodobnost, že byl signál správně přijat, víme-li, že byla detekována tečka, je 0,849. [0,849]

9. V jednom městě jezdí 85 % zelených taxíků a 15 % modrých. Svědek dopravní nehody vypověděl, že nehodu zavinil řidič modrého taxíku, který pak ujel. Testy provedené za obdobných světelných podmínek ukázaly, že svědek dobře identifikuje barvu taxíku v 80 % případů a ve 20 % případů se mýlí.

   a) Jaká je pravděpodobnost, že viník nehody skutečně řídil modrý taxík?

      Řešení:

      $M$ ... viník řídil modrý taxík
      $Z$ ... viník řídil zelený taxík
      $S1M (S2M)$ ... první (druhý) svědek označil modrý taxík
      $S1Z (S2Z)$ ... první (druhý) svědek označil zelený taxík



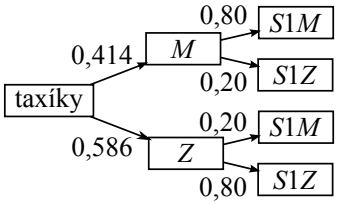

$$P(M|S1M) = \frac{0,15\cdot 0,8}{0,15\cdot 0,8 + 0,85\cdot 0,2} = \frac{0,12}{0,29} \cong 0,414$$

      Pravděpodobnost, že viník nehody skutečně řídil modrý taxík je 0,414. [0,414]

   b) Následně byl nalezen další nezávislý svědek, který rovněž tvrdí, že taxík byl modrý. Jaká je nyní pravděpodobnost, že viník nehody skutečně řídil modrý taxík?

      Řešení:

      Po první výpovědi: $P(M) = 0,414$

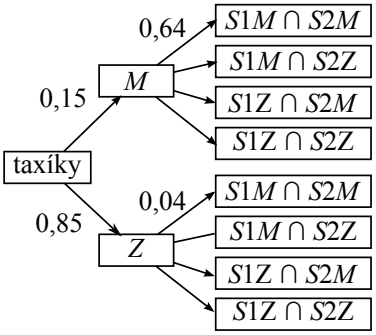



$$P(M|S2M) = \frac{0,414\cdot 0,8}{0,414\cdot 0,8 + 0,586\cdot 0,2} = \frac{0,331}{0,448} \cong 0,739$$

      Pravděpodobnost, že viník nehody řídil modrý taxík je 0,739. [0,739]

c) Ovlivní pravděpodobnost, že viník nehody skutečně řídil modrý taxík to, zda dva výše zmínění svědci vypovídali postupně nebo najednou?

Řešení:




$$P((S1M \cap S2M)|M) = 0,8 \cdot 0,8 = 0,64$$
$$P((S1M \cap S2M)|Z) = 0,2 \cdot 0,2 = 0,4$$

$$P(M|(S1M \cap S2M)) = \frac{0,15 \cdot 0,64}{0,15 \cdot 0,64 + 0,85 \cdot 0,04} =$$

$$\frac{0,48}{0,65} = 0,7384$$

Jsou-li svědci nezávislí, nezáleží na tom, zda vypovídají postupně nebo najednou. [Ne]

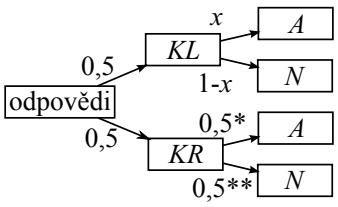
Úlohu prezentovali psychologové Kahneman a Tversky

(Anděl; Matematika náhody; 2007)

10. Potřebujeme zjistit odpověď na určitou citlivou otázku (např. „Podváděl(a) jste u zkoušky?“, „Jste homosexuál (lesba)?“, apod.). Jak odhadnout, kolik procent dotazovaných na otázku odpoví ANO a přitom všem respondentům zaručit naprostou anonymitu? Jedním z řešení je tzv. **dvojitě anonymní anketa**:
Necháme respondenty hodit korunou a dvojkorunou a ti, kterým padl na koruně líc napíšou na lísteček odpověď (ANO/NE) na citlivou otázku. Ostatní respondenti napíší, zda jim padl na dvojkoruně líc (ANO/NE). Jakým způsobem určíme podíl studentů, kteří na citlivou otázku odpověděli ANO?
Předpokládejme, že respondenti byli dotazování, zda podváděli u zkoušky. Z anketních lístků se zjistilo, že „ANO“ odpovědělo 120 respondentů a „NE“ odpovědělo 200 respondentů. Kolik procent studentů podvádělo u zkoušky?

Řešení:

$KL$ ... na koruně padne líc
$KR$ ... na koruně padl rub
$DL$ ... na dvojkoruně padl líc
$DR$ ... na dvojkoruně padl rub
$A$ ... respondent odpověděl „ano“
$N$ ... respondent odpověděl „ne“




$$*P(A|KR) = P(DL) = 0,5$$
$$**P(N|KR) = P(DR) = 0,5$$
$$P(A) = P(A \cap KL) + P(A \cap KR) = 0,5 \cdot x + 0,5 \cdot 0,5 =$$
$$= 0,5x + 0,25 \Rightarrow x = \frac{P(A) - 0,25}{0,5} = 2 \cdot P(A) - 0,5$$

Anketa: $P(A) = \frac{120}{320} = \frac{3}{8} \Rightarrow x = 2 \cdot \frac{3}{8} - 0,5 = 0,25$

Podvádělo přibližne 25 % studentů.

[přibližne 25 %]

# 3. Náhodná veličina

**Náhodná veličina** - funkce, která každému výsledku náhodného pokusu přiřadí reálné číslo. Je to matematický model popisující více či méně dobře realitu, který vytváříme, jestliže chceme zpracovávat výsledky náhodného pokusu.

**Hodnota náhodné veličiny** - výsledek náhodného pokusu vyjádřený reálným číslem.

**Rozdělení pravděpodobnosti** - pravidlo, které každé hodnotě (nebo intervalu hodnot) přiřazuje pravděpodobnost, že náhodná veličina nabude této hodnoty (nebo intervalu hodnot). Jestliže známe rozdělení pravděpodobnosti, je náhodná veličina z pravděpodobnostního hlediska úplně popsána. Nejčastějším způsobem popisu rozdělení pravděpodobnosti je distribuční funkce.

**Distribuční funkce** je reálná funkce definována jako $F(x) = P(X < x)$. Jde tedy o funkci, která každému reálnému číslu přiřazuje pravděpodobnost, že náhodná veličina nabývá hodnot menších než toto reálné číslo.

**Diskrétní náhodná veličina** je náhodná veličina, která může nabývat pouze konečného (výsledek hodu kostkou) nebo spočetně nekonečného (počet zákazníků snažících se dovolat do call centra během dne) množství hodnot. Diskrétní náhodnou veličinu popisujeme prostřednictvím **pravděpodobnostní funkce**, popř. distribuční funkce.

**Spojitá náhodná veličina** je náhodnou veličinou, která má spojitou distribuční funkci. Pro popis spojité náhodné veličiny používáme distribuční funkci a **hustotu pravděpodobnosti**.

Pravděpodobnost výskytu náhodné veličiny na nějakém intervalu určujeme na základě následujících vztahů.

$$P(X < a) = F(a)$$

$$P(X \geq b) = 1 - F(b)$$

$$P(a \leq X < b) = F(b) - F(a)$$

**Číselné charakteristiky diskrétní náhodné veličiny:**

- střední hodnota: $E(X) = \mu = \sum_{(i)} x_i \cdot P(x_i)$
- rozptyl - výpočetní vztah: $D(X) = E(X^2) - (E(X))^2$
- směrodaná odchylka: $\sigma = \sqrt{D(X)}$
- modus: $\hat{x}$ - taková hodnota DNV, v níž $P(x_i)$ nabývá svého maxima

**Číselné charakteristiky spojité náhodné veličiny:**

- střední hodnota: $E(X) = \mu = \int_{-\infty}^{\infty} x \cdot f(x)$
- rozptyl - výpočetní vztah: $D(X) = E(X^2) - (E(X))^2$
- směrodaná odchylka: $\sigma = \sqrt{D(X)}$
- modus: $\hat{x}$ - taková hodnota SNV, v níž $f(x)$ nabývá svého maxima
- p-kvantil: $x_p$ - taková hodnota, že pravděpodobnost, že SNV nabude menších hodnot než $x_p$ je 100p%: $P(X < x_p) = p \Rightarrow F(x_p) = p$

## 3.1 Test z teorie

1. Vytvořte dvojice pojem - příklad.

| | |
|---|---|
| a) náhodný pokus | 1) Doba přenosu testovacího datového souboru je delší než 30 s. |
| b) náhodný jev | 2) Měření doby přenosu testovacího datového souboru. |
| c) náhodná veličina | 3) Doba přenosu testovacího datového souboru. |

a-2, b-1, c-3

2. Určete pravdivost následujících výroků.

N a) Náhodnou veličinu chápeme jako výsledek náhodného pokusu.
[číselný výsledek ⇒ hodnota NV]

P b) Diskrétní náhodná veličina může nabývat konečného nebo spočetného množství hodnot.

P c) Distribuční funkce náhodné veličiny $X$ v bodě $t$ udává pravděpodobnost, že $X$ nabývá hodnot menších než $t$.

P d) Má-li náhodná veličina spojitou distribuční funkci, je spojitá.

P e) Je-li $X$ diskrétní náhodná veličina, pak $\sum_i P(X = x_i) = 1$.

N f) Oborem hodnot distribuční funkce jsou všechna reálná čísla.
[Pozor! Střední hodnota je jednoznačně definovaný pojem.]

N h) Nabývá-li funkce $f(x)$ hodnoty 1,3, nemůže jít o hustotu pravděpodobnosti.

P i) Rozdělení spojité náhodné veličiny můžeme popsat distribuční funkci a hustotou pravděpodobnosti.

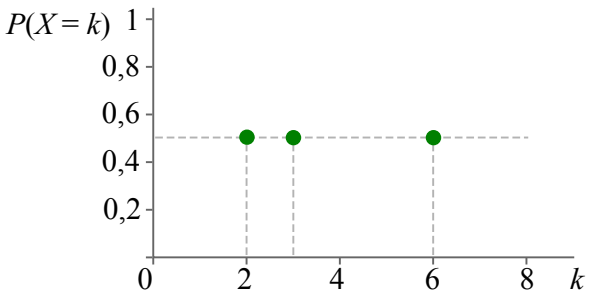
3. Určete, která ze zadaných funkcí nemůže představovat pravděpodobnostní funkci.

a) $P(X = k) = \begin{cases} \frac{1}{k} & k \in \{2; 3; 6\} \\ 0 & k \notin \{2; 3; 6\} \end{cases}$

b)

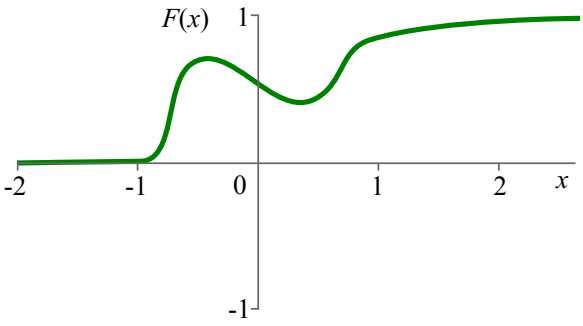
| $k$ | 2 | 3 | 6 |
|---|---|---|---|
| $P(X = k)$ | 0,2 | 0,4 | 0,4 |

c)

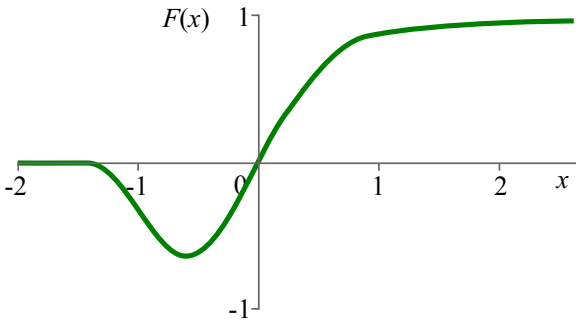



$\sum_{i}^{3} P(X = k_i) \neq 1$

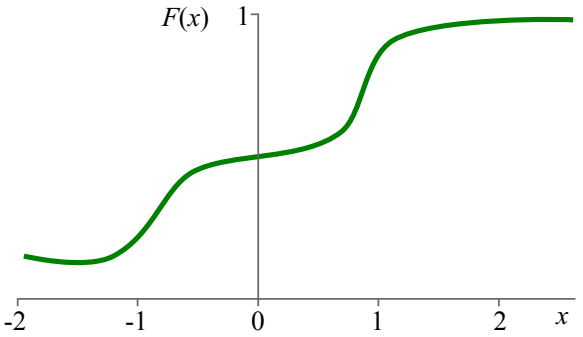
4. Určete, zda by grafy znázorněných funkcí mohly představovat distribuční funkci.



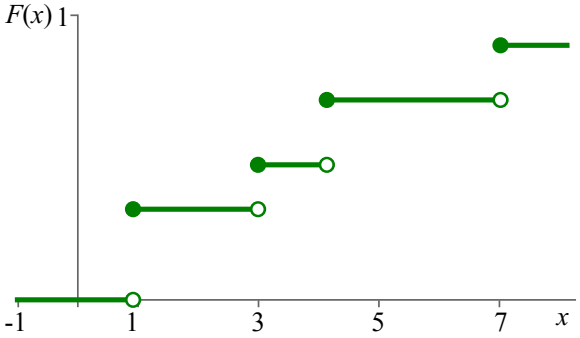

a)



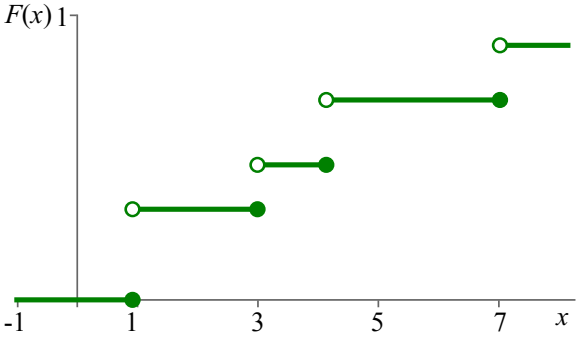

b)



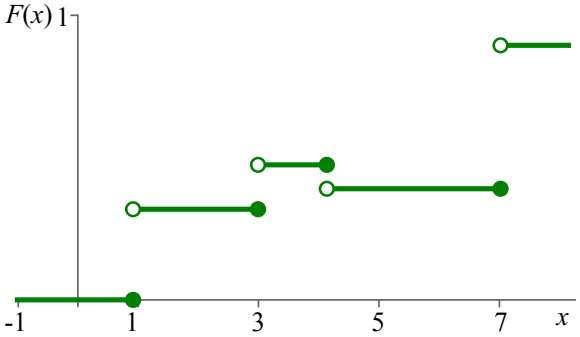

c)




d)

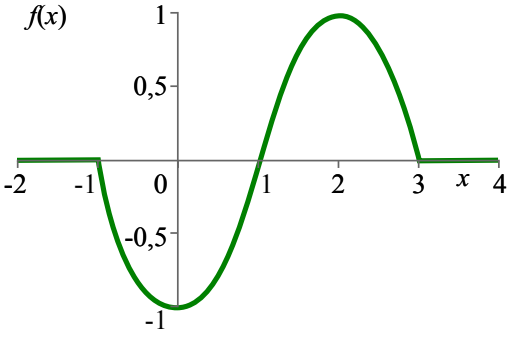



e)

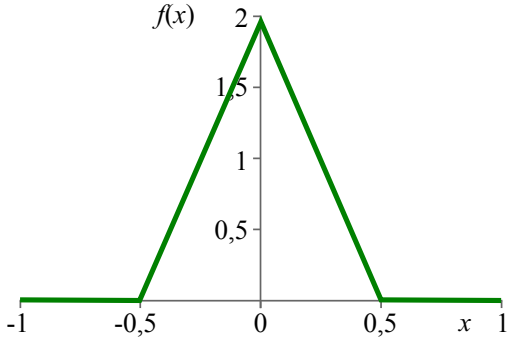



f)

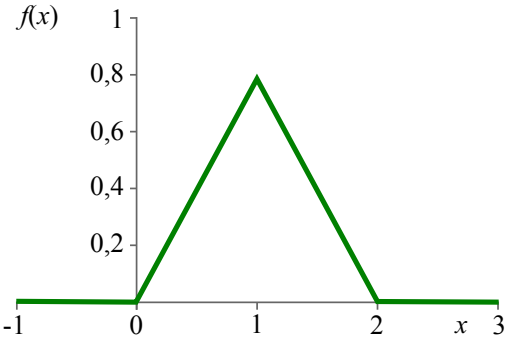
5. Určete, zda by grafy znázorněných funkcí mohly představovat hustotu pravděpodobnosti.



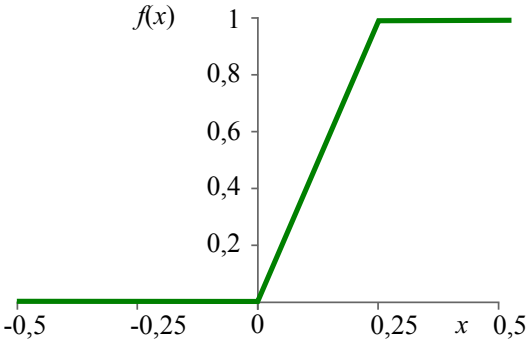

a)




b)




c)
plocha pod křivkou není rovna 1




d)

6. Nechť náhodná veličina $X$ představuje životnost (dobu do poruchy) monitorů na počítačové učebně E320. Určete pravdivost následujících výroků.

P a) $X$ je spojitou náhodnou veličinou.
N b) Rozdělení $X$ může být popsáno pravděpodobnostní funkcí.
P c) Pro popis $X$ lze použít intenzitu poruch.

7. Vyjádřete následující pravděpodobnosti pomocí distribuční funkce.

a) $P(X < 10) = F(10)$,
b) $P(X \geq 5) = 1 - F(5)$,
c) $P(5 \leq X < 10) = F(10) - F(5)$.

8. Nechť $X$ je diskrétní náhodná veličina. Vyjádřete co nejjednodušeji následující pravděpodobnosti pomocí $P(X = 10), P(X < 10), P(X > 10), P(X = 5), P(X < 5), P(X > 5)$.

a) $P(X \leq 10) = P(X < 10) + P(X = 10)$,
b) $P(X \geq 10) = 1 - P(X < 10) = P(X > 10) + P(X = 10)$,
c) $P(5 < X \leq 10) = P(X < 10) + P(X = 10) - P(X < 5) - P(X = 5)$,
d) $P(5 \leq X \leq 10) = P(X < 10) + P(X = 10) - P(X = 5)$.

9. Nechť $X$ je spojitá náhodná veličina. Vyjádřete co nejjednodušeji následující pravděpodobnosti pomocí $P(X = 10), P(X < 10), P(X > 10), P(X = 5), P(X < 5), P(X > 5)$.

a) $P(X \leq 10) = P(X < 10)$,
b) $P(X \geq 10) = P(X > 10)$,
c) $P(5 < X \leq 10) = P(X < 10) - P(X < 5)$,
d) $P(5 \leq X \leq 10) = P(X < 10) - P(X < 5)$.

10. Nechť $X$ je spojitá náhodná veličina. Vyjádřete následující pravděpodobnosti pomocí hustoty pravděpodobnosti.

a) $P(X \leq 10) = \int_{-\infty}^{10} f(x)dx$,
b) $P(X \geq 10) = \int_{10}^{\infty} f(x)dx$,
c) $P(5 < X \leq 10) = \int_{5}^{10} f(x)dx$,
d) $P(5 \leq X \leq 10) = \int_{5}^{10} f(x)dx$.

## 3.2 Diskrétní náhodná veličina - příklady

1. Majitel servisního střediska nabídl prodejně automobilů, která si zřídila autopůjčovnu své služby. Za každý automobil zapůjčený jeho prostřednictvím obdrží od autopůjčovny 500,- Kč. Zároveň se však zavázal, že každý den investuje do údržby zapůjčených automobilů 800,- Kč. Počet automobilů zapůjčených prostřednictvím servisního střediska za 1 den je popsán následující pravděpodobnostní funkcí:

| $x_i$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| $P(x_i)$ | 0,01 | 0,40 | 0,25 | 0,15 | 0,10 | | 0,03 |

a) Hodnota pravděpodobnostní funkce pro 5 automobilů byla špatně čitelná. Určete ji.

Řešení:

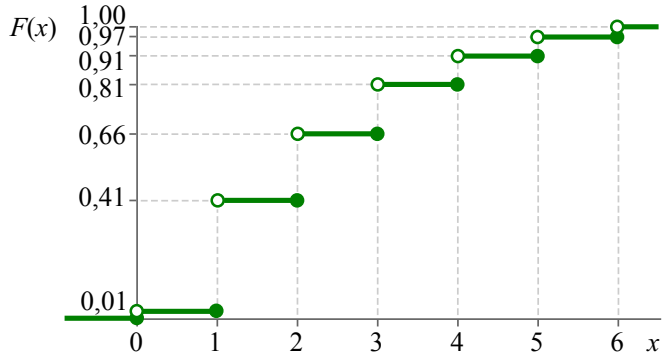
$P(X = 5) = 1 - (0,01 + 0,40 + 0,25 + 0,15 + 0,10 + 0,03) = 0,06$

Hodnota pravděpodobnostní funkce pro 5 automobilů je 0,06. [0,06]

b) Určete a zakreslete distribuční funkci náhodné veličiny $X$, která je definována jako počet zapůjčených automobilů.

Řešení:

| $x$ | $(-\infty; 0\rangle$ | $(0; 1\rangle$ | $(1; 2\rangle$ | $(2; 3\rangle$ | $(3; 4\rangle$ | $(4; 5\rangle$ | $(5; 6\rangle$ | $(6; \infty\rangle$ |
|---|---|---|---|---|---|---|---|---|
| $F(x)$ | 0,00 | 0,01 | 0,41 | 0,66 | 0,81 | 0,91 | 0,97 | 1,00 |




c) Určete střední hodnotu, rozptyl, směrodatnou odchylku a modus počtu zapůjčených automobilů během jednoho dne.

Řešení:

| $x_i$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | $\sum$ |
|---|---|---|---|---|---|---|---|---|
| $P(x_i)$ | 0,01 | 0,40 | 0,25 | 0,15 | 0,10 | 0,06 | 0,03 | 1 |
| $x_i \cdot P(x_i)$ | 0,00 | 0,40 | 0,50 | 0,45 | 0,40 | 0,30 | 0,18 | 2,23 |
| $x_i^2 \cdot P(x_i)$ | 0,00 | 0,40 | 1,00 | 1,35 | 1,60 | 1,50 | 1,08 | 6,93 |

$E(X) = \sum_{i=1}^{n} x_i \cdot P(x_i) \Rightarrow E(X) = 2,23$

$E(X^2) = \sum_{i=1}^{n} x_i^2 \cdot P(x_i)$

$D(X) = E(X^2) - (E(X))^2 = 6,93 - 2,23^2 = 1,96$

$\sigma(X) = \sqrt{D(X)} = \sqrt{1,96} = 1,4$

$\hat{x} = 1$ $[E(X) = 2,23;\ D(X) = 1,96;\ \sigma(X) = 1,4\ ;\hat{x} = 1]$

d) Určete pravděpodobnostní funkci a distribuční funkci náhodné veličiny $Y$, která je definována jako denní příjem majitele servisu.

Řešení:

$Y = 500 \cdot X \quad [\text{Kč}]$

| $y_i$ | 0 | 500 | 1000 | 1500 | 2000 | 2500 | 3000 |
|---|---|---|---|---|---|---|---|
| $P(y_i)$ | 0,01 | 0,40 | 0,25 | 0,15 | 0,10 | 0,06 | 0,03 |

$$F_Y(y) = \begin{cases} 0 & y \leq 0 \\ 0,01 & 0 < y \leq 500 \\ 0,41 & 500 < y \leq 1\,000 \\ 0,66 & 1\,000 < y \leq 1\,500 \\ 0,81 & 1\,500 < y \leq 2\,000 \\ 0,91 & 2\,000 < y \leq 2\,500 \\ 0,97 & 2\,500 < y \leq 3\,000 \\ 1 & y > 3\,000 \end{cases}$$

e) Určete střední hodnotu, směrodatnou odchylku a modus příjmu majitele servisu ze zapůjčených automobilů během jednoho dne.

Řešení:

$E(a \cdot X + b) = a \cdot E(X) + b \Rightarrow E(Y) = E(500 \cdot X) = 500 \cdot E(X) = 1\,115$ Kč

$D(a \cdot X + b) = a^2 \cdot D(X) \Rightarrow D(Y) = D(500 \cdot X) = 500^2 \cdot D(X) \Rightarrow \sigma(Y) =$
$= 500 \cdot \sigma(X) = 700$ Kč

$\hat{y} = 500$ Kč

$[E(Y) = 1\,115$ Kč; $\sigma(Y) = 700$ Kč; $\hat{y} = 500$ Kč$]$

f) Určete pravděpodobnost, že příjem majitele servisu (náhodná veličina $Y$) z půjčování automobilů převýší jeho výdaje.

Řešení:

$P(Y > 800) = 1 - P(Y \leq 800) = 1 - (P(Y = 0) + P(Y = 500)) = 0,59$

[0,59]

g) Určete střední hodnotu, směrodatnou odchylku a modus náhodné veličiny $Z$, která je definována jako zisk majitele servisu ze zapůjčených automobilů během jednoho dne.

Řešení:

$Z = 500 \cdot X - 800 = Y - 800$

$E(Z) = E(Y - 800) = E(Y) - 800 = 1\,115 - 800 = 315$ Kč

$D(Z) = D(Y - 800) = D(Y) \Rightarrow \sigma(Z) = \sigma(Y) = 700$ Kč

$\hat{z} = \hat{y} - 800 = -300$ Kč

$[E(Z) = 315$ Kč; $\sigma(Z) = 700$ Kč; $\hat{z} = -300$ Kč$]$

2. Pro distribuční funkci náhodné veličiny $X$ platí:

$$F(x) = \begin{cases} 0 & x \leq -1 \\ 0,3 & -1 < x \leq 0 \\ 0,7 & 0 < x \leq 1 \\ 1 & x > 1 \end{cases}$$

a) Určete pravděpodobnostní funkci náhodné veličiny $X$, její střední hodnotu a směrodatnou odchylku.

Řešení:

| $x_i$ | -1 | 0 | 1 | $\sum$ |
|---|---|---|---|---|
| $P(x_i)$ | 0,3 | 0,4 | 0,3 | 1 |

$$E(X) = \sum_{i=1}^{n} x_i \cdot P(x_i) = -0,3 + 0 + 0,3 = 0$$

$$E(X^2) = \sum_{i=1}^{n} x_i^2 \cdot P(x_i) = 0,3 + 0 + 0,3 = 0,6$$

$$D(X) = E(X^2) - (E(X))^2 = 0,6 - 0^2 = 0,6 \Rightarrow \sigma(X) = \sqrt{D(X)} = 0,77$$

$$[E(X) = 0;\ \sigma(X) = 0,77]$$

b) Náhodná veličina $Y = 1 - 3X$, určete $P(y), F(y), E(Y), D(Y)$.

Řešení:

| $y_i$ | 4 | 1 | -2 | $\sum$ |
|---|---|---|---|---|
| $P(y_i)$ | 0,3 | 0,4 | 0,3 | 1 |

$\Rightarrow$

| $y_i$ | -2 | 1 | 4 | $\sum$ |
|---|---|---|---|---|
| $P(y_i)$ | 0,3 | 0,4 | 0,3 | 1 |

$$E(Y) = E(1 - 3 \cdot X) = 1 - 3 \cdot E(X) = 1 - 3 \cdot 0 = 1$$

$$D(Y) = D(1 - 3 \cdot X) = (-3)^2 \cdot D(X) = 9 \cdot D(X) = 9 \cdot 0,6 = 5,4$$

| $y$ | $(-\infty; -2\rangle$ | $(-2; 1\rangle$ | $(1; 4\rangle$ | $(4; \infty\rangle$ |
|---|---|---|---|---|
| $F(y)$ | 0 | 0,3 | 0,7 | 1 |

$$[E(Y) = 1; D(Y) = 5,4]$$

c) Náhodná veličina $W = 3X^2$, určete $P(w), F(w), E(W), D(W)$.

Řešení:

" | $w_i$ | 3 | 0 | 3 | $\sum$ | "

| $P(w_i)$ | 0,3 | 0,4 | 0,3 | 1 |
|---|---|---|---|---|

$\Rightarrow$

| $w_i$ | 0 | 3 | $\sum$ |
|---|---|---|---|
| $P(w_i)$ | 0,4 | 0,6 | 1 |

| $w$ | $(-\infty; 0\rangle$ | $(0; 3\rangle$ | $(3; \infty\rangle$ |
|---|---|---|---|
| $F(w)$ | 0 | 0,4 | 1 |

$$E(W) = \sum_{i=1}^{n} w_i \cdot P(w_i) = 0 + 1,8 = 1,8$$

$$E(W^2) = \sum_{i=1}^{n} w_i^2 \cdot P(w_i) = 0 + 5,4 = 5,4$$

$$D(W) = E(W^2) - (E(W))^2 = 5,4 - 1,8^2 = 2,16$$

$$[E(W) = 1,8; D(W) = 2,16]$$

3. V dílně jsou dva stroje pracující nezávisle na sobě. Pravděpodobnost poruchy prvního stroje je 0,2, pravděpodobnost poruchy druhého stroje je 0,3. Náhodná veličina $X$ je definována jako počet současně porouchaných strojů. Určete:

   a) pravděpodobnostní funkci náhodné veličiny $X$,

   Řešení:

   $\overline{S}_1 \ldots$ stroj 1 se porouchá
   $\overline{S}_2 \ldots$ stroj 2 se porouchá

   $P(X=0) = P(S_1) \cdot P(S_2) = 0,8 \cdot 0,7 = 0,56$
   $P(X=1) = P(S_1 \cap \overline{S}_2) + P(\overline{S}_1 \cap S_2) = 0,8 \cdot 0,3 + 0,2 \cdot 0,7 = 0,38$
   nebo $P(X=1) = 1 - (P(X=0) + P(X=2)) = 1 - (0,56 + 0,06) = 0,38$
   $P(X=2) = P(\overline{S}_1 \cap \overline{S}_2) = 0,2 \cdot 0,3 = 0,06$

| $x_i$ | 0 | 1 | 2 | $\sum$ |
|---|---|---|---|---|
| $P(x_i)$ | 0,56 | 0,38 | 0,06 | 1 |

   b) distribuční funkci náhodné veličiny $X$,

   Řešení:

| $x$ | $(-\infty; 0\rangle$ | $(0; 1\rangle$ | $(1; 2\rangle$ | $(2; \infty\rangle$ |
|---|---|---|---|---|
| $F(x)$ | 0,00 | 0,56 | 0,94 | 1,00 |

   c) střední hodnotu a rozptyl náhodné veličiny $X$.

   Řešení:

   $$E(X) = \sum_{i=1}^{n} x_i \cdot P(x_i) = 0 + 0,38 + 0,12 = 0,50$$

   $$E(X^2) = \sum_{i=1}^{n} x_i^2 \cdot P(x_i) = 0 + 0,38 + 0,24 = 0,62$$

   $$D(X) = E(X^2) - (E(X))^2 = 0,62 - 0,5^2 = 0,37$$

   $[E(X) = 0,50;\ D(X) = 0,37]$

## 3.3 Spojitá náhodná veličina - příklady

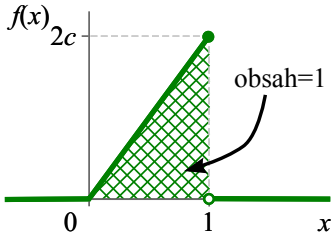
1. Náhodná veličina $X$ má distribuční funkci

$$F(x) = \begin{cases} 0 & x \leq 0 \\ cx^2 & 0 < x \leq 1 \\ 1 & x > 1 \end{cases}$$

Jaké hodnoty může nabývat konstanta $c$?

Řešení:

a) Pomocí hustoty pravděpodobnosti:




$$f(x) = \frac{dF(x)}{dx} \begin{cases} 2cx & x \in \langle 0;1 \rangle \\ 0 & x \notin \langle 0;1 \rangle \end{cases}$$

$$\int_{-\infty}^{\infty} f(x)dx = 1 \Rightarrow \int_{-\infty}^{\infty} 2cxdx = \frac{2c \cdot 1}{2} = 1 \Rightarrow c = 1$$

b) Pomocí distribuční funkce, která má být z definice spojitá $\Rightarrow$ v bodě $x = 1$ musí navazovat $\Rightarrow c \cdot 1^2 = 1$

Konstanta $c$ může nabývat pouze hodnoty 1.

[1]

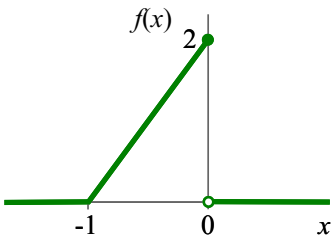
2. Rozdělení náhodné veličiny $X$ je dáno hustotou

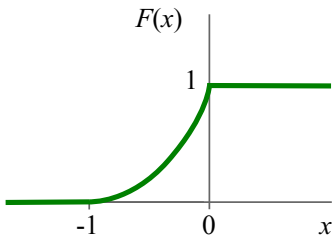
$$f(x) = \begin{cases} 2x + 2 & x \in \langle -1;0 \rangle \\ 0 & x \notin \langle -1;0 \rangle \end{cases}$$

Určete:

a) $F(x)$,

Řešení:




$$F(x) = \begin{cases} 0 & x \in (-\infty;-1) \\ \int_{-1}^{x} (2t+2)dt = [t^2+2t]_{-1}^{x} = x^2+2x+1 = (x+1)^2 & x \in \langle -1;0 \rangle \\ 1 & x \in (0;\infty) \end{cases}$$

b) $P(-2 \leq X \leq -0,5)$, $P(-2 \leq X \leq -1)$, $P(X > 0,5)$, $P(X = 0,3)$

Řešení:

$P(-2 \leq X \leq -0,5) = \int_{-1}^{-0,5} (2x+2)dx = F(-0,5) - F(-2) = (\frac{1}{2})^2 - 0 = 0,25$

$P(-2 \leq X \leq -1) = 0$ (viz graf hustoty pravděpodobnosti)

$P(X > 0,5) = 1 - P(X \leq 0,5) = 1 - 1 = 0$ (viz graf hustoty pravděpodobnosti)

$P(X = 0,3) = 0$ (jde o SNV)

[0,25; 0; 0; 0]

c) střední hodnotu, rozptyl a směrodatnou odchylku náhodné veličiny $X$.

Řešení:

$$E(X) = \int_{-\infty}^{\infty} f(x)dx = \int_{-1}^{0} x \cdot (2x+x)dx = \int_{-1}^{0} 2x^2 + 2xdx = \left[\frac{2}{3}x^3 + x^2\right]_{-1}^{0} = 0 - \frac{1}{3} = -\frac{1}{3}$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_{-1}^{0} x^2 \cdot (2x+x)dx = \left[\frac{x^4}{2} + \frac{2x^3}{3}\right]_{-1}^{0} = \frac{1}{6}$$

$$D(X) = E(X^2) - (E(X))^2 = \frac{1}{6} - \left(-\frac{1}{3}\right)^2 = \frac{1}{6} - \frac{1}{9} = \frac{3-2}{18} = \frac{1}{18}$$

$$\sigma(X) = \sqrt{D(X)} = \frac{\sqrt{18}}{18} = \frac{3 \cdot \sqrt{2}}{18} = \frac{\sqrt{2}}{6}$$

$\left[E(X) = -\frac{1}{3}; D(X) = \frac{1}{18}; \sigma(X) = \frac{\sqrt{2}}{6}\right]$

d) modus $\hat{x}$

Řešení:

$\hat{x} = 0$ viz graf hustoty pravděpodobnosti

(nebo hledání globálního maxima hustoty pravděpodobnosti)

[0]

e) medián $x_{0,5}$

Řešení:

$F(x_{0,5}) = 0,5 \Rightarrow x_{0,5} \in \langle -1; 0 \rangle$

$x_{0,5}^2 + 2 \cdot x_{0,5} + 1 = 0,5$

$x_{0,5}^2 + 2 \cdot x_{0,5} + 0,5 = 0$

$D = 4 - 2 = 2$

$$x_{0,5_{1,2}} = \frac{-2 \pm \sqrt{2}}{2} = \begin{array}{ll} \nearrow & \frac{-2+\sqrt{2}}{2} = x_{0,5} \\ \searrow & \frac{-2-\sqrt{2}}{2} \notin \langle -1; 0 \rangle \end{array}$$

$$\left[-1 + \frac{\sqrt{2}}{2}\right]$$

3. Náhodná veličina $Y$ je definována jako: $Y = 3X+1$, kde $X$ je náhodná veličina z předcházejícího příkladu. Určete:

   a) $F_Y(y)$

   Řešení:

$$F_Y(y) = P(Y < y) = P(3X + 1 < y) = P(X < \tfrac{y-1}{3}) = F_X(\tfrac{y-1}{3})$$

$$F_Y(y) = \begin{cases} 0 & \frac{y-1}{3} \in (-\infty; -1) \\ \left(\frac{y-1}{3} + 1\right)^2 & \frac{y-1}{3} \in \langle -1; 0\rangle \\ 1 & \frac{y-1}{3} \in (0; \infty) \end{cases}$$

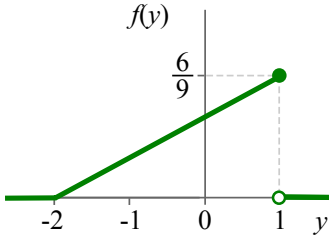
$$F_Y(y) = \begin{cases} 0 & y \in (-\infty; -2) \\ \frac{1}{9}(y+2)^2 & y \in \langle -2; 1\rangle \\ 1 & y \in (1; \infty) \end{cases}$$

   b) $f_Y(y)$

   Řešení:

$$f_Y(y) = \frac{dF_Y(y)}{dy}$$

$$f_Y(y) = \begin{cases} \frac{2}{9}(y+2) & y \in \langle -2; 1\rangle \\ 0 & y \notin \langle -2; 1\rangle \end{cases}$$




   c) $E(Y), D(Y), \sigma(Y)$

   Řešení:

$$E(Y) = E(3X + 1) = 3E(X) + 1 = 3 \cdot \left(-\frac{1}{3}\right) + 1 = 0$$

$$D(Y) = D(3X + 1) = 9D(X) = 9 \cdot \frac{1}{18} = \frac{1}{2}$$

$$\sigma(Y) = \sqrt{D(Y)} = \frac{\sqrt{2}}{2}$$

$$\left[E(Y) = 0; D(Y) = \tfrac{1}{2}; \sigma(Y) = \tfrac{\sqrt{2}}{2}\right]$$

# 4. Náhodný vektor

**Náhodným vektorem** rozumíme vektor $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ složený z náhodných veličin, který je charakterizován sdruženým rozdělením pravděpodobnosti. Ze sdruženého rozdělení pravděpodobnosti můžeme snadno najít marginální rozdělení pravděpodobnosti charakterizující jednotlivé složky náhodného vektoru.

**Podmíněné rozdělení pravděpodobnosti** určuje rozdělení NV $X$ za předpokladu, že NV $Y$ nabyla hodnoty $y$. Chápeme ho jako podíl sdruženého a marginálního rozdělení pravděpodobnosti (má-li tento podíl smysl), v souladu s definicí podmíněné pravděpodobnosti.
Podmíněná pravděpodobnostní funkce má tvar:

$$P(x|y) = \frac{p(x,y)}{P_Y(y)}, P_Y(y) \neq 0$$

$$P(y|x) = \frac{p(x,y)}{P_X(x)}, P_X(x) \neq 0$$

**Nezávislost** složek náhodného vektoru se projevuje tím, že jeho sdružená distribuční funkce (sdružená pravděpodobnostní funkce, resp. sdružená hustota pravděpodobnosti) se dá matematicky vyjádřit jako součin marginálních distribučních funkcí (marginálních pravděpodobností, resp. marginálních hustot pravděpodobnosti) jednotlivých náhodných veličin.

**Střední hodnota** náhodného vektoru $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ je dána jako vektor marginálních středních hodnot $E(\boldsymbol{X}) = (E(X_1), E(X_2), \ldots, E(X_n))^T$.

**Rozptyl** náhodného vektoru $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ je dán jako vektor marginálních rozptylů $D(\boldsymbol{X}) = (D(X_1), D(X_2), \ldots, D(X_n))^T$. Obdobně pro ostatní číselné charakteristiky.

**Kovariance** $cov(X,Y)$ je nejjednodušší ukazatel vztahu mezi dvěma náhodnými veličinami. Kladná hodnota kovariance znamená, že se zvětšením hodnoty $X$ se pravděpodobně zvýší i hodnota $Y$. Oproti tomu záporná hodnota kovariance informuje o tom, že se zvětšením hodnoty $X$ se pravděpodobně sníží hodnota $Y$. Kovarianci nelze použít jako míru závislosti, protože $cov(X,Y) \in (-\infty; \infty)$.

$$cov(X,Y) = E(X \cdot Y) - E(X) \cdot E(Y)$$

V praxi se často setkáváme s **kovarianční maticí**

$$var(X) = \begin{pmatrix} D(X) & cov(X,Y) \\ cov(Y,X) & D(Y) \end{pmatrix}$$

**Korelační koeficient** $\rho(X,Y)$ je mírou lineární závislosti dvou složek náhodného vektoru.

$$\rho(X,Y) = \begin{cases} \dfrac{cov(X,Y)}{\sqrt{D(X) \cdot D(Y)}}, & D(X), D(Y) \neq 0 \\ 0, & \text{jinak.} \end{cases}$$

**Korelační matice**

$$cor(X) = \begin{pmatrix} 1 & \rho(X,Y) \\ \rho(Y,X) & 1 \end{pmatrix}$$

## 4.1 Test z teorie

1. Určete, zda jsou pravdivé následující výroky.

N a) Náhodný vektor je definován jako dvourozměrný vektor, jehož složkami jsou náhodné veličiny. [$n$-rozměrný]

P b) Sdružené rozdělení popisuje rozdělení náhodného vektoru.

P c) Marginální rozdělení popisuje rozdělení jednotlivých složek náhodného vektoru.

N d) Je-li $\boldsymbol{X} = (X, Y)^T$, pak $E(\boldsymbol{X}) = E(XY)$.

N e) $E(XY) = E(X) \cdot E(Y)$.

N f) Marginální charakteristiky náhodného vektoru popisují vztah mezi náhodnými veličinami, které tvoří jeho složky.

N g) Kovariance je mírou závislosti náhodných veličin.

N h) Je-li $cov(X, Y) = 0$, pak jsou náhodné veličiny X a Y nezávislé.

P i) Je-li $cov(X, Y) = 0$, jsou náhodné veličiny X a Y nekorelované.

P j) Je-li $\rho(X, Y) = 0$, pak jsou náhodné veličiny X a Y nekorelované.

P k) Jsou-li náhodné veličiny X a Y nekorelované, jsou lineárně nezávislé.

N l) $cov(X, X) = 1$.

P m) $cov(X, Y) = cov(Y, X)$.

P n) $\rho(X, X) = 1$.

P o) $cov(X, Y) = E(XY) - E(X) \cdot E(Y)$.

## 4.2 Příklady

1. Náhodný vektor $\boldsymbol{Z} = (Y; X)^T$ má pravděpodobnostní funkci zadanou tabulkou:

| $X \backslash Y$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 3 | 0,01 | 0,02 | 0,03 | 0,25 |
| 5 | 0,04 | 0,16 | ? | 0,05 |
| 7 | 0,12 | 0,07 | 0,06 | 0,01 |

Určete:

a) chybějící hodnotu sdružené pravděpodobnostní funkce,

Řešení:

$$\sum_i \sum_j p(y_i, x_j) = 1 \Rightarrow p(Y = 3, X = 5) = 1 - (0,01 + 0,02 + 0,03 + 0,25 + 0,4 + 0,16 +$$

$$+0,05 + 0,12 + 0,07 + 0,06 + 0,01) = 0,18$$

[0,18]

b) $p(2;5)$,

Řešení:

$p(2;5) = p(Y = 2; X = 5) = 0,16$ [0,16]

c) $p(5;2)$,

Řešení:

$p(5;2) = p(Y = 5; X = 2) = 0$ [0]

d) $F(2,8;7,1)$,

Řešení:

$F(2,8;7,1) = p(Y < 2,8; X < 7,1) = p(1;3) + p(2;3) + p(1;5) + p(2;5) + p(1;7) + p(2;7) =$
$= 0,01 + 0,02 + 0,04 + 0,16 + 0,12 + 0,07 = 0,42$

[0,42]

e) $P(Y > 2,1; X < 5,3)$,

Řešení:

$P(Y > 2,1; X < 5,3) = p(3;3)+p(4;3)+p(3;5)+p(4;5) = 0,03+0,25+0,18+0,05 = 0,51$

[0,51]

f) $P(Y > 2,1|X < 5,3)$,

Řešení:

$$P(Y > 2,1|X < 5,3) = \frac{P(Y > 2,1 \wedge X < 5,3)}{P(X < 5,3)} = \frac{0,51}{P(X = 3) + P(X = 5)} =$$
$$= \frac{0,51}{0,01 + 0,02 + 0,03 + 0,25 + 0,04 + 0,16 + 0,18 + 0,05} = \frac{0,51}{0,74} = 0,69$$

[0,69]

g) marginální rozdělení NV $X$ (marginální pravděpodobnostní f-ci i marginální distribuční f-ci),

Řešení:

| $x_i$ | 3 | 5 | 7 | $\sum$ |
|---|---|---|---|---|
| $P(x_i)$ | 0,31 | 0,43 | 0,26 | 1 |

| $x$ | $(-\infty;3\rangle$ | $(3;5\rangle$ | $(5;7\rangle$ | $(7;\infty\rangle$ |
|---|---|---|---|---|
| $F(x)$ | 0,00 | 0,31 | 0,74 | 1,00 |

h) marginální rozdělení NV $Y$ (marginální pravděpodobnostní f-ci i marginální distribuční f-ci),

Řešení:

| $y_i$ | 1 | 2 | 3 | 4 | $\sum$ |
|---|---|---|---|---|---|
| $P(y_i)$ | 0,17 | 0,25 | 0,27 | 0,31 | 1 |

| $y$ | $(-\infty;1\rangle$ | $(1;2\rangle$ | $(2;3\rangle$ | $(3;4\rangle$ | $(4;\infty\rangle$ |
|---|---|---|---|---|---|
| $F(y)$ | 0,00 | 0,17 | 0,42 | 0,69 | 1,00 |

i) $F_X(5,3)$,

Řešení:

$$F_X(5,3) = P(X < 5,3) = 0,74$$ [0,74]

j) $F_Y(5,3)$,

Řešení:

$$F_Y(5,3) = P(Y < 5,3) = 1$$ [1]

k) $P(X = 5|Y = 1)$,

Řešení:

$$P(X = 5|Y = 1) = \frac{P(X = 5 \wedge Y = 1)}{P(Y = 1)} = \frac{0,04}{0,17} = 0,24$$

[0,24]

l) $P(Y = 5|X = 1)$,

Řešení:

$$P(Y = 5|X = 1) = \frac{P(Y = 5 \wedge X = 1)}{P(X = 1)} = \frac{0}{0} \Rightarrow \text{není definováno}$$

[není definováno]

m) podmíněnou pravděpodobnostní funkci $P(x|y)$,

Řešení:

$$P(x|y) = \frac{p(y,x)}{P_Y(y)}$$

$$P(3|1) = \frac{P(X=3,Y=1)}{P_Y(1)} = \frac{0,01}{0,17} = \frac{1}{17} \qquad P(3|2) = \frac{P(X=3,Y=2)}{P_Y(2)} = \frac{0,02}{0,25} = \frac{2}{25}$$
$$P(5|1) = \frac{P(X=5,Y=1)}{P_Y(1)} = \frac{0,04}{0,17} = \frac{4}{17} \qquad P(5|2) = \frac{P(X=5,Y=2)}{P_Y(2)} = \frac{0,16}{0,25} = \frac{16}{25}$$
$$P(7|1) = \frac{P(X=7,Y=1)}{P_Y(1)} = \frac{0,12}{0,17} = \frac{12}{17} \qquad P(7|2) = \frac{P(X=7,Y=2)}{P_Y(2)} = \frac{0,07}{0,25} = \frac{7}{25}$$

$$P(3|3) = \frac{P(X=3,Y=3)}{P_Y(3)} = \frac{0,03}{0,27} = \frac{3}{27} \qquad P(3|4) = \frac{P(X=3,Y=4)}{P_Y(4)} = \frac{0,25}{0,27} = \frac{25}{31}$$
$$P(5|3) = \frac{P(X=5,Y=3)}{P_Y(3)} = \frac{0,18}{0,27} = \frac{18}{27} \qquad P(5|4) = \frac{P(X=5,Y=4))}{P_Y(4)} = \frac{0,05}{0,27} = \frac{5}{31}$$
$$P(7|3) = \frac{P(X=7,Y=3)}{P_Y(3)} = \frac{0,06}{0,27} = \frac{6}{27} \qquad P(7|4) = \frac{P(X=7,Y=4)}{P_Y(4)} = \frac{0,01}{0,27} = \frac{1}{31}$$

| $X\backslash Y$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 3 | $\frac{1}{17}$ | $\frac{2}{25}$ | $\frac{3}{27}$ | $\frac{25}{31}$ |
| 5 | $\frac{4}{17}$ | $\frac{16}{25}$ | $\frac{18}{27}$ | $\frac{5}{31}$ |
| 7 | $\frac{12}{17}$ | $\frac{7}{25}$ | $\frac{6}{27}$ | $\frac{1}{31}$ |
| $\sum$ | 1 | 1 | 1 | 1 |

n) podmíněnou pravděpodobnostní funkci $P(y|x)$,

Řešení:

$$P(y|x) = \frac{p(y,x)}{P_X(x)}$$

$$P(1|3) = \frac{P(Y=1,X=3)}{P_Y(3)} = \frac{0{,}01}{0{,}31} = \frac{1}{31}$$
$$P(2|3) = \frac{P(Y=2,X=3)}{P_Y(3)} = \frac{0{,}02}{0{,}31} = \frac{2}{31}$$
$$P(3|3) = \frac{P(Y=3,X=3)}{P_Y(3)} = \frac{0{,}13}{0{,}31} = \frac{16}{31}$$
$$P(4|3) = \frac{P(Y=4,X=3)}{P_Y(3)} = \frac{0{,}25}{0{,}31} = \frac{25}{31}$$

$$P(1|5) = \frac{P(Y=1,X=5)}{P_Y(5)} = \frac{0{,}04}{0{,}43} = \frac{4}{143}$$
$$P(2|5) = \frac{P(Y=2,X=5)}{P_Y(5)} = \frac{0{,}16}{0{,}43} = \frac{16}{43}$$
$$P(3|5) = \frac{P(Y=3,X=5)}{P_Y(5)} = \frac{0{,}18}{0{,}43} = \frac{18}{43}$$
$$P(4|5) = \frac{P(Y=4,X=5)}{P_Y(5)} = \frac{0{,}05}{0{,}43} = \frac{5}{43}$$

$$P(1|7) = \frac{P(Y=1,X=7)}{P_Y(7)} = \frac{0{,}12}{0{,}26} = \frac{12}{26}$$
$$P(2|7) = \frac{P(Y=2,X=7)}{P_Y(7)} = \frac{0{,}07}{0{,}26} = \frac{7}{26}$$
$$P(3|7) = \frac{P(Y=3,X=7)}{P_Y(7)} = \frac{0{,}06}{0{,}26} = \frac{6}{26}$$
$$P(4|7) = \frac{P(Y=4,X=7)}{P_Y(7)} = \frac{0{,}01}{0{,}26} = \frac{1}{26}$$

| $X\backslash Y$ | 1 | 2 | 3 | 4 | $\sum$ |
|---|---|---|---|---|---|
| 3 | $\frac{1}{31}$ | $\frac{2}{31}$ | $\frac{3}{31}$ | $\frac{25}{31}$ | 1 |
| 5 | $\frac{4}{43}$ | $\frac{16}{43}$ | $\frac{18}{43}$ | $\frac{5}{43}$ | 1 |
| 7 | $\frac{12}{26}$ | $\frac{7}{26}$ | $\frac{6}{26}$ | $\frac{1}{26}$ | 1 |

o) základní číselné charakteristiky NV $X$, tj. $E(X), D(X), \sigma(X), \hat{X}$,

Řešení:

| $x_i$ | $P(x_i)$ | $x_i \cdot P(x_i)$ | $x_i^2 \cdot P(x_i)$ |
|---|---|---|---|
| 3 | 0,31 | 0,93 | 2,79 |
| 5 | 0,43 | 2,15 | 10,75 |
| 7 | 0,26 | 1,82 | 12,74 |
| $\sum$ | 1 | 4,90 | 26,28 |

$E(X) = 4,9$

$D(X) = E(X^2) - (E(X))^2 = 26,28 - 4,9^2 = 2,27 \cong 2,3$

$\sigma_X = \sqrt{D(X)} = \sqrt{2,27} = 1,50 \cong 1,5$

$\hat{X} = 5$

$[E(X) = 4,9; D(X) = 2,3; \sigma_X = 1,5; \hat{X} = 5]$

p) základní číselné charakteristiky NV $Y$, tj. $E(Y), D(Y), \sigma(Y), \hat{Y}$,

Řešení:

| $y_i$ | $P(y_i)$ | $y_i \cdot P(y_i)$ | $y_i^2 \cdot P(y_i)$ |
|---|---|---|---|
| 1 | 0,17 | 0,17 | 0,17 |
| 2 | 0,25 | 0,50 | 1,00 |
| 3 | 0,27 | 0,81 | 2,43 |
| 4 | 0,31 | 1,24 | 4,96 |
| $\sum$ | 1 | 2,72 | 8,56 |

$E(Y) = 2,72 \cong 2,7$

$D(Y) = E(Y^2) - (E(Y))^2 = 8,56 - 2,72^2 = 1,16 \cong 1,2$

$\sigma_Y = \sqrt{D(Y)} = \sqrt{1,16} = 1,08 \cong 1,1$

$\hat{Y} = 4$

$[E(Y) = 2,7; D(Y) = 1,2; \sigma_Y = 1,1; \hat{Y} = 4]$

q) základní číselné charakteristiky náhodného vektoru $\boldsymbol{Z} = (Y; X)^T, tj. E(\boldsymbol{Z}), D(\boldsymbol{Z}), \sigma(\boldsymbol{Z})$,

Řešení:

$E(\boldsymbol{Z}) = (E(Y), E(X)) = (2,72; 4,90) \cong (2,7; 4,9)$

$D(\boldsymbol{Z}) = (D(Y), D(X)) = (1,16; 2,27) \cong (1,2; 2,3)$

$\sigma(\boldsymbol{Z}) = (\sigma(Y), \sigma(X)) = (1,08; 1,50) \cong (1,1; 1,5)$

$[E(\boldsymbol{Z}) = (2,7; 4,9), D(\boldsymbol{Z}) = (1,2; 2,3), \sigma(\boldsymbol{Z}) = (1,1; 1,5)]$

r) $E(X|Y = 2)$,

Řešení:

$E(X|Y = 2) = \sum_i x_i \cdot P(x_i|Y = 2) = 3 \cdot \frac{2}{25} + 5 \cdot \frac{16}{25} + 7 \cdot \frac{7}{25} = \frac{6+80+49}{25} = \frac{135}{25} = 5,4$

[5,4]

s) kovarianci $cov(Y; X)$,

Řešení:

$cov(Y; X) = E(Y \cdot X) - E(Y) \cdot E(X)$

$E(Y \cdot X) = \sum_i \sum_j x_i \cdot y_j \cdot p(y_i, x_j) = 1 \cdot 3 \cdot 0,01 + 2 \cdot 3 \cdot 0,02 + 3 \cdot 3 \cdot 0,03 + 4 \cdot 3 \cdot 0,25 +$

$1 \cdot 5 \cdot 0,04 + 2 \cdot 5 \cdot 0,16 + 3 \cdot 5 \cdot 0,18 + 4 \cdot 5 \cdot 0,05 + 1 \cdot 7 \cdot 0,12 + 2 \cdot 7 \cdot 0,07 + 3 \cdot 7 \cdot 0,06 + 4 \cdot 7 \cdot 0,01 = 12,28$

$cov(Y; X) = E(Y \cdot X) - E(Y) \cdot E(X) = 12,28 - 2,7 \cdot 4,9 \cong -1,1$

$[-1,1]$

t) koeficient korelace $\rho(Y; X)$.

Řešení:

$$\rho(Y; X) = \frac{cov(Y; X)}{\sqrt{D(Y) \cdot D(X)}} = \frac{cov(Y, X)}{\sigma(Y) \cdot \sigma(X)} = \frac{-1,1}{1,1 \cdot 1,5} \cong -0,7$$

$[-0,7]$

u) Jsou NV $X$ a $Y$ nezávislé?

Řešení:

Nejsou, např. pro $Y = 1 \wedge X = 3 : P_Y(1) \cdot P_X(3) \neq p(1,3) : 0,17 \cdot 0,31 \neq 0,01$

[nejsou]

v) Jsou NV $X$ a $Y$ lineárně nezávislé ?

Řešení:

Není, protože $\rho(Y,X) \neq 0$

[nejsou]

2. Náhodný vektor $\boldsymbol{Z} = (X;Y)^T$ nabývá hodnot $(0;1)^T$ s pravděpodobností 1/2, hodnoty $(0;2)^T$ s pravděpodobností 1/3 a hodnoty $(1;1)^T$ s pravděpodobností 1/6.

a) Určete korelační koeficient.

Řešení:

| $X\backslash Y$ | 1 | 2 | $P_X(x_i)$ | $x_i \cdot P_X(x_i)$ | $x_i^2 \cdot P_X(x_i)$ |
|---|---|---|---|---|---|
| 0 | $\frac{1}{2}$ | $\frac{1}{3}$ | $\frac{5}{6}$ | 0 | 0 |
| 1 | $\frac{1}{6}$ | 0 | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ |
| $P_Y(y_i)$ | $\frac{4}{6}$ | $\frac{2}{6}$ | 1 | $\frac{1}{6}$ | $\frac{1}{6}$ |
| $y_i \cdot P_Y(y_i)$ | $\frac{4}{6}$ | $\frac{4}{6}$ | $\frac{8}{6}$ | | |
| $y_i^2 \cdot P_Y(y_i)$ | $\frac{4}{6}$ | $\frac{8}{6}$ | $\frac{12}{6}$ | | |

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma(X) \cdot \sigma(Y)} = \frac{E(Y \cdot X) - E(Y) \cdot E(X)}{\sqrt{D(Y) \cdot D(X)}}$$

$$E(X \cdot Y) = 0 \cdot 1 \cdot \tfrac{1}{2} + 0 \cdot 2 \cdot \tfrac{1}{3} + 1 \cdot 1 \cdot \tfrac{1}{6} + 1 \cdot 2 \cdot 0 = \tfrac{1}{6}$$

$$D(X) = \tfrac{1}{6} - \tfrac{1}{36} = \tfrac{5}{36}$$

$$D(Y) = \tfrac{12}{6} - \tfrac{64}{36} = \tfrac{8}{36}$$

$$\rho(X,Y) = \frac{\frac{1}{6} - \frac{1}{6} \cdot \frac{8}{6}}{\sqrt{\frac{5}{36} \cdot \frac{8}{36}}} = \frac{\frac{6-8}{36}}{\frac{\sqrt{40}}{36}} = \frac{-2}{\sqrt{40}} = \frac{-2}{2\sqrt{10}} = \frac{-1}{\sqrt{10}} = -0,32$$

[$-0,32$]

b) Rozhodněte, zda jsou NV $X$ a $Y$ nezávislé.

Řešení:

Nejsou, např. pro $X = 0 \wedge Y = 1 : P_X(0) \cdot P_Y(1) \neq p(0,1) : \frac{4}{6} \cdot \frac{5}{6} \neq \frac{1}{2}$

[nejsou]

c) Rozhodněte, zda jsou NV $X$ a $Y$ lineárně nezávislé.

Řešení:

Není, protože $\rho(X,Y) \neq 0$

[nejsou]

# 5. Vybraná rozdělení diskrétní náhodné veličiny

| Název NV $X$ | Popis | Pravděpodobnostní funkce | $E(X)$ | $D(X)$ |
|---|---|---|---|---|
| **Binomická** $Bi(n,\pi)$ | počet úspěchů v $n$ Bernoulliho pokusech | $P(X=k)=\binom{n}{k}\pi^k(1-\pi)^k$ | $n\pi$ | $n\pi(1-\pi)$ |
| **Hypergeometrická** $H(N,M,n)$ | počet úspěchů v $n$ **závislých** pokusech | $P(X=k)=\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$ | | |
| **Alternativní** $A(\pi)$ | počet úspěchů v 1 pokusu | $P(X=1)=\pi$<br>$P(X=0)=1-\pi$ | $\pi$ | $\pi(1-\pi)$ |
| **Geometrická** $Ge(\pi)$ | počet pokusů do 1. úspěchu (včetně) | $P(X=n)=\pi(1-\pi)^{n-1}$ | $\frac{1}{\pi}$ | $\frac{1-\pi}{\pi^2}$ |
| **Negativně binomická (Pascalova)** $NB(k,\pi)$ | počet pokusů do k. úspěchu (včetně) | $P(X=n)=\binom{n-1}{k-1}\pi^k(1-\pi)^{n-k}$ | $\frac{k}{\pi}$ | $\frac{k(1-\pi)}{\pi^2}$ |
| **Poissonova** $Po(\lambda t)$ | počet událostí v Poissonově procesu v uzavřené oblasti (v čase, na ploše, v objemu) | $P(X=k)=\frac{(\lambda t)^k}{k!}e^{-\lambda t}$ | $\lambda t$ | $\lambda t$ |

| Popis | Podmínky | | Název NV $X$ |
|---|---|---|---|
| počet úspěchů v $n$ pokusech | nezávislé pokusy | $n=1$ | **Alternativní** - $A(\pi)$ |
| | | $n\geq 1$ | **Binomická** - $Bi(n,\pi)$ |
| | závislé pokusy | | **Hypergeometrická** - $H(N,M,n)$ |
| počet pokusů do $k$. úspěchu (včetně) | nezávislé pokusy | $k=1$ | **Geometrická** - $Ge(\pi)$ |
| | | $k\geq 1$ | **Negativně binomická** - $NB(k,\pi)$ |
| počet události v uzavřené oblasti (v čase, na ploše, v objemu) | ordinarita, stacionarita, beznáslednost procesu | | **Poissonova** - $Po(\lambda t)$ |

## 5.1 Test z teorie

1. Určete pravdivost následujících tvrzení.

N a) Rozdělení pravděpodobnosti diskrétní náhodné veličiny může být dáno výhradně pravděpodobnostní funkcí.
[nebo distribuční funkcí]

P b) Posloupnost nezávislých pokusů majících pouze dva možné výsledky se stejnou pravděpodobnosti úspěchu nazýváme Bernoulliho pokusy.

N c) Počet úspěchů v $n$ pokusech lze popsat binomickou náhodnou veličinou.
[nebo hypergeometrickou NV]

P d) Geometrické rozdělení je speciálním případem negativně binomického rozdělení.

P e) Pascalovo rozdělení je pouze jiný název pro negativně binomické rozdělení.

N f) Jistý supermarket má otevřeno 24h denně. Počet zákazníků v supermarketu během otevírací doby lze popsat náhodnou veličinou s Poissonovým rozdělením.
[nesplněn předpoklad stacionarity]

2. Charakterizujte rozdělení náhodné veličiny popisující

a) počet studentů, kteří úspěšně ukončí kurz STA1 v tomto semestru (z minulých let víme, že pravděpodobnost, že student úspěšně dokončí kurz STA1 je 0,63; do kurzu je v tomto semestru přihlášeno 248 studentů), $[X \sim Bi(248; 0,63)]$

b) počet vadných mikroprocesorů na chipu (na chipu je průměrně 1 vadný mikroprocesor), $[X \sim Po(1)]$

c) počet hodů poctivou kostkou nutných k padnutí šestky, $[X \to G(\frac{1}{6})$, nebo $X \sim NB(1; \frac{1}{6})]$

d) počet řidičů obsloužených na čerpací stanici za půl hodiny (během 1h je na čerpací stanici obslouženo průměrně 72 řidičů), $[X \to Po(36)]$

e) počet řidičů obsloužených do chvíle, kdy 1. řidič ujede bez placení (průměrně ujíždí bez placení 1 z 50 řidičů), $[X \sim G(\frac{1}{50})$, nebo $X \to NB(1; \frac{1}{50})]$

f) počet týdnů v roce (52 týdnů), v nichž neujede žádný řidič z čerpací stanice bez placení (během týdne je na čerpací stanici obslouženo průměrně 4 000 řidičů, z nichž cca 2 % ujedou bez placení), $[X \sim Bi(52; 0)]$
[Y... počet řidičů, kteří ujedou bez zaplacení z 4000 $\Rightarrow Y \sim Bi(4000; 0,2), P(Y < 1) = 0$]

g) počet řidičů obsloužených do chvíle, kdy 4. řidič ujede bez placení (průměrně ujíždí bez placení 1 z 50 řidičů). $[X \sim NB(4; \frac{1}{50})]$

## 5.2 Příklady

1. Bridž se hraje s 52 bridžovými kartami, které se rozdají mezi 4 hráče. Vždy 2 hráči hrají spolu. Při rozdávání (13 karet) jste dostali do rukou 2 esa. Jaká je pravděpodobnost, že váš partner bude mít zbývající dvě esa?

   Řešení:

   39 karet (všechny karty kromě mých)
   ↙ ↘
   2 esa 37 ostatních karet

   $$P(\text{2esa}) = \frac{C(2,2) \cdot C(37,11)}{C(39,13)} = 0,105$$

   nebo

   $$X \ldots \text{počet es spoluhráče ve výběru 13 karet z 39}$$

   $$X \sim H(39,2,13)$$

   $P(X = 2) = 0,105$

   Pravděpodobnost, že spoluhráč má zbylá dvě esa je 0,11. [0,11]

2. Pokusy se zjistilo, že radioaktivní látka vyzařuje během 7,5 s průměrně 3,87 $\alpha$-částice. Určete pravděpodobnost toho, že za 1 sekundu vyzáří tato látka alespoň jednu $\alpha$-částici.

   Řešení:

   $$X \ldots \text{počet vyzářených } \alpha\text{-částic za 1 s}$$

   $$X \sim Po(0,516)$$

   $t = 1 \text{ s} \Rightarrow \lambda t = \frac{3,87}{7,5} = 0,516$

   $P(X \geq 1) = P(X > 0) = 0,403$

   Pravděpodobnost, že za 1 s vyzáří tato látka alespoň 1 $\alpha$-částici je 0,40. [0,40]

3. Kamarád vás pošle do sklepa, abyste donesl(a) 4 lahvová piva - z toho dvě desítky a dvě dvanáctky. Nevíte, kde rozsvítit, proto vezmete z basy poslepu 4 láhve. S jakou pravděpodobností jste vyhověl(a), víte-li, že v base bylo celkem 10 desítek a 6 dvanáctek?

   Řešení:

   16 piv
   ↙ ↘
   10 x 10° 6 x 12°

   $$P(2\ 10°) = \frac{C(10,2) \cdot C(6,2)}{C(16,4)} = 0,3709$$

   nebo

   $$X \ldots \text{počet } 10° \text{ ve výběru 4 lahví ze 16 lahví (jako } X \text{ lze volit i počet } 12° \text{ piv)}$$

   $$X \sim H(16,10,4)$$

   $P(X = 2) = 0,3709$

   Pravděpodobnost, že jsem donesl(a) dvě 10°a dvě 12° je 0,37.

   [0,37]

4. V jednom mililitru určitého dokonale rozmíchaného roztoku se v průměru nachází 15 určitých mikroorganismů. Určete pravděpodobnost, že při náhodném výběru vzorku o objemu 1/2 mililitru bude ve zkumavce méně než 5 těchto mikroorganismu.

   Řešení:

   $$X \ldots \text{počet mikroorganismů v } \tfrac{1}{2} \text{ ml roztoku}$$

   $$X \sim Po(7,5)$$

   $t = \frac{1}{2}$ ml $\Rightarrow \lambda t = \frac{15}{2} = 7,5$

   $P(X < 5) = P(X > 0) = 0,132$

   Pravděpodobnost, že ve zkumavce bude méně než 5 mikroorganismů je 0,13.

   [0,13]

5. Na stůl vysypeme 15 mincí. Jaká je pravděpodobnost, že počet mincí ležících lícem nahoře, je od 8 do 15?

   Řešení:

   $$X \ldots \text{počet mincí ležících lícem nahoře}$$

   $$X \sim Bi(15; 0,5)$$

   $P(8 \leq X \leq 15) = P(X > 7) = 0,5$

   Pravděpodobnost, že počet mincí ležících lícem nahoře je od 8 do 15 je 0,5.

   [0,50]

6. Pravděpodobnost, že se dovoláme do studia rozhlasové stanice, která právě vyhlásila telefonickou soutěž je 0,08. Jaká je pravděpodobnost, že se dovoláme nejvýše na 4. pokus?

   Řešení:

   $$X \ldots \text{počet pokusů, než se dovoláme do rozhlasové stanice}$$

   $$X \sim NB(1; 0,08) \ (G(0,08))$$

   $P(X \leq 4) = P(X < 5) = 0,284$

   Pravděpodobnost, že se dovoláme nejvýše na 4. pokus je 0,28.

   [0,28]

7. V továrně se vyrobí denně 10 % vadných součástek. Jaká je pravděpodobnost, že vybereme-li třicet součástek z denní produkce, tak nejméně dvě budou vadné?

   Řešení:

   $$X \ldots \text{počet vadných součástek ze 30 vybraných}$$

   $$X \sim Bi(30; 0,1)$$

   $P(X \geq 2) = 1 - P(X < 2) = P(X > 1) = 0,816$

   Pravděpodobnost, že nejméně dvě součástky budou vadné je 0,82.

   [0,82]

8. Ve skladu je 200 součástek. 10 % z nich je vadných. Jaká je pravděpodobnost, že vybereme-li ze skladu třicet součástek, tak nejméně dvě budou vadné?

   Řešení:

   $$X \ldots \text{počet vadných součástek ve výběru 30 součástek z 200}$$

   $$X \sim H(200; 20; 30)$$

   (Proč $X \sim H$ ? Protože se mění pravděpodobnost úspěchu, 200 je malý počet$\Rightarrow$ závislé pokusy, např. pro $N = 10\,000 \Rightarrow P(X \geq 2) = 0,817$)

   $P(X \geq 2) = 1 - P(X < 2) = 0,839$

   Pravděpodobnost, že nejméně dvě součástky ve výběru 30 součástek z 200 budou vadné je 0,84.

   [0,84]

9. V určité firmě bylo zjištěno, že na 33 % počítačů je nainstalován nějaký nelegální software. Určete pravděpodobnostní a distribuční funkci počtu počítačů s nelegálním softwarem mezi třemi kontrolovanými počítači.

   Řešení:

   $$X \ldots \text{počet počítačů s nelegálním softwarem}$$

   $$X \sim Bi(3; 0,33)$$

   $P(X = k) = \binom{3}{k} \cdot 0,33^k \cdot (1 - 0,33)^{3-k} = \dfrac{3!}{(3-k)!k!} \cdot 0.33^k \cdot 0.67^{3-k}$

   $P(X = 0) = \frac{3!}{3!0!} \cdot 0,33^0 \cdot 0,67^3 = 0,67^3 = 0,30$

   $P(X = 1) = \frac{3!}{2!1!} \cdot 0,33^1 \cdot 0,67^2 = 3 \cdot 0,33 \cdot 0,67^3 = 0,44$

   $P(X = 2) = \frac{3!}{1!2!} \cdot 0,33^2 \cdot 0,67^1 = 3 \cdot 0,33^2 \cdot 0,67 = 0,22$

   $P(X = 3) = \frac{3!}{0!3!} \cdot 0,33^3 \cdot 0,67^0 = 0,33^3 = 0,04$

   (Pozor - možná chyba vlivem zaokrouhlení)

   nebo: $P(X = 3) = 1 - (P(X = 0) + P(X = 1) + P(X = 2))$

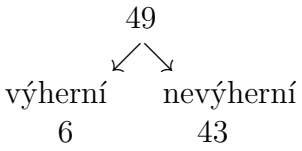
| $x_i$ | 0 | 1 | 2 | 3 | $\sum$ |
|---|---|---|---|---|---|
| $P(x_i)$ | 0,30 | 0,44 | 0,22 | 0,04 | 1,00 |

| $x$ | $(-\infty; 0\rangle$ | $(0; 1\rangle$ | $(1; 2\rangle$ | $(2; 3\rangle$ | $(3; \infty\rangle$ |
|---|---|---|---|---|---|
| $F(x)$ | 0,00 | 0,30 | 0,74 | 0,96 | 1,00 |

10. Sportka je loterijní hra, v níž sázející tipuje šest čísel ze čtyřiceti devíti, která očekává, že padnou při budoucím slosování. K účasti ve hře je nutné zvolit alespoň jednu kombinaci 6 čísel (vždy 6 čísel na jeden sloupec) a pomocí křížků tato čísla označit na sázence společnosti Sazka a.s. do sloupců, počínaje sloupcem prvním. Sázející vyhrává v případě, že uhodne alespoň tři čísla z tažené šestice čísel. Jaká je pravděpodobnost, že proto, aby sázející vyhrál, bude muset vyplnit:

a) právě tři sloupce,

Řešení:




$Y \ldots$ počet uhádnutých čísel v 6 ze 49

$$Y \sim H(49; 6; 6)$$

$$\pi = P(Y \geq 3) = 0,019$$

$X \ldots$ počet vyplněných sloupců nutných k výhře (1. úspěch)

$$X \sim NB(1; \pi)$$

$$X \sim NB(1; 0,019)$$

$P(X = 3) = 0,018$

Pravděpodobnost, že proto, aby sázející vyhrál, bude muset vyplnit právě tři sloupce je 0,02.

[0,02]

b) alespoň 5 sloupců,

Řešení:

$P(X \geq 5) = P(X > 4) = 0,928$

Pravděpodobnost, že proto, aby sázející vyhrál, bude muset vyplnit alespoň 5 sloupců je 0,93.

[0,93]

c) méně než 10 sloupců,

Řešení:

$P(X < 10) = 0,156$

Pravděpodobnost, že proto, aby sázející vyhrál, bude muset vyplnit méně než 10 sloupců je 0,16.

[0,16]

d) více než 5 a nejvýše 10 sloupců?

Řešení:

$P(5 < X \leq 10) = P(6 \leq X < 11) = P(X < 11) - P(X < 6) = 0,082$

Pravděpodobnost, že proto, aby sázející vyhrál, bude muset vyplnit více než 5 a nejvýše 10 sloupců je 0,08.

[0,08]

# 6. Vybraná rozdělení spojité náhodné veličiny

| **Rozdělení NV** | **Popis** | **Hustota pravděpodobnosti<br>Distribuční funkce<br>Intenzita poruch** | $E(X)$ | $D(X)$ |
|---|---|---|---|---|
| **Rovnoměrné**<br>$Ro(a,b)$ | $f(x)$ je na $(a;b)$ konstantní, jinde nulová | $f(x) = \begin{cases} \frac{1}{b-a} & x \in \langle a;b\rangle \\ 0 & \text{jinde} \end{cases}$ | $\frac{a+b}{2}$ | $\frac{(a-b)^2}{12}$ |
| **Exponenciální**<br>$Exp(\lambda)$ | doba do 1. události, doba mezi událostmi (pouze v období stabilního života) | $f(t) = \lambda \cdot e^{-\lambda t}; t>0; \lambda>0$<br>$F(t) = 1 - e^{-\lambda t}; t>0; \lambda>0$<br>$\lambda(t) = \lambda = \text{konst.}; t>0; \lambda>0$ | $\frac{1}{\lambda}$ | $\frac{1}{\lambda^2}$ |
| **Erlangovo**<br>$Erlang(k;\lambda)$ | doba do k-té události | $f(t) = \lambda \cdot e^{-\lambda t} \cdot \frac{(\lambda t)^{k-1}}{(k-1)!}; t>0$<br>$F(t) = 1 - e^{-\lambda t} \cdot \sum_{j=0}^{k-1} \frac{(\lambda t)^j}{j!}$<br>$\lambda(t) = \frac{\lambda}{(k-1)! \sum_{j=0}^{k-1} \frac{1}{(k-1-j)!(\lambda t)^j}}$ | $\frac{k}{\lambda}$ | $\frac{k}{\lambda^2}$ |
| **Weibullovo**<br>$Wb(\Theta,\beta)$ | doba do 1. události (poruchy)(vhodná volba $\beta$ umožuje použití v libovolném období intenzity poruch) | $f(t) = \frac{\beta}{\Theta}\left(\frac{t}{\Theta}\right)^{\beta-1} e^{-\left(\frac{t}{\Theta}\right)^\beta}$<br>$F(t) = 1 - e^{-\left(\frac{t}{\Theta}\right)^\beta}$<br>$\lambda(t) = \frac{\beta}{\Theta}\left(\frac{t}{\Theta}\right)^{\beta-1} t>0; \Theta>0; \beta>0$ | | |
| **Normované normální**<br>$N(0,1)$ | hodnoty distribuční funkce jsou tabelovány, hustota pravděpodobnosti je sudá funkce - Gaussův klobouk | $\phi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}; -\infty < x < \infty$<br>$\Phi(x) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$ | 0 | 1 |
| **Normální**<br>$N(\mu;\sigma^2)$ | distribučí funkci určjeme pomocí standardizace normální náhodné veličiny | $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2}$<br>$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \int_{-\infty}^{x} e^{-\left(\frac{t-\mu}{\sqrt{2}\sigma}\right)^2} dt$ | $\mu$ | $\sigma^2$ |
| **Logaritmicko -normální**<br>$LN(\mu;\sigma^2)$ | distribuční funkci určujeme převodem na dis. funkci norm. normálního rozdělení | $f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \cdot e^{-\left(\frac{ln(x)-\mu}{\sqrt{2}\sigma}\right)^2}; x>0$ | $e^{\mu+\frac{\sigma^2}{2}}$ | $e^{2\mu+\sigma^2} \cdot$<br>$\cdot (e^{\sigma^2-1})$ |

## 6.1 Test z teorie

1. Určete pravdivost následujících tvrzení.

N a) Intenzita poruch (hazardní funkce) je neklesající funkce.

N b) Exponenciální rozdělení používáme k modelování životnosti výrobků nacházejících se v období stárnutí.

[období stárnutí]

P c) Exponenciální rozdělení je speciálním případem Weibullova rozdělení.

P d) Weibullovo rozdělení lze použít k modelování životnosti výrobků nacházejících se v libovolném období života.

N e) Normální rozdělení má právě jeden parametr.

[2 parametry]

N f) Hustota pravděpodobnosti normální náhodné veličiny je sudá funkce.

[pouze pro normované normální rozdělení]

N g) Distribuční funkce normální náhodné veličiny je tabelována.

P h) Má-li náhodná veličina normální rozdělení, pak (střední hodnota = medián = modus).

N i) Má-li náhodná veličina normální rozdělení se střední hodnotou $\mu$ a sm. odchylkou $\sigma$, pak přibližně 5 % hodnot náhodné veličiny leží mimo interval $\langle \mu - 3\sigma; \mu + 3\sigma \rangle$.

2. Doplňte:

a) Intenzitu poruch lze použít k popisu ..... spojitých náhodných veličin.

[nezáporných]

b) Exponenciální rozdělení používáme k modelování životnosti výrobků nacházejících se v období .....

[stabilního života]

c) Pro modelování životnosti výrobku, který má lineárně rostoucí intenzitu poruch lze použít Weibullovo rozdělení s parametrem tvaru $\beta =$ .....

[2]

d) Gaussova křivka je grafem ..... normálního rozdělení.

[hustoty pravděpodobnosti]

## 6.2 Příklady

1. Výška v populaci chlapců ve věku 3,5-4 roky má normální rozdělení se střední hodnotou 102 cm a směrodatnou odchylkou 4,5 cm. Určete, jaké procento chlapců v uvedeném věku má výšku menší nebo rovnou 93 cm.

   Řešení:

   $$X \ldots \text{výška v populaci chlapců ve věku 3,5-4 roky}$$

   $$X \sim N(102 \text{ cm}; 4,5^2 \text{ cm}^2)$$

   $P(X \leq 93) = P(X < 93) = 0,0228$

   2,28 % chlapců v uvedeném věku má výšku menší nebo rovnu 93 cm.

   [2,28 %]

2. Průměrná životnost strojní součástky je 30 000 hodin. Předpokládejme, že součástka je v období stabilního života. Určete:

   a) pravděpodobnost, že součástka nevydrží více než 2 000 hodin,

   Řešení:

   $$X \ldots \text{životnost strojní součástky}$$

   $$X \sim Exp(\lambda)$$

   $\lambda$: průměrná životnost 30 000 h $\Rightarrow E(X) = 30\,000 \quad E(X) = \frac{1}{\lambda} \Rightarrow \lambda = \frac{1}{30\,000}$ h$^{-1}$

   $$X \sim Exp(30\,000^{-1}\text{h}^{-1})$$

   $P(X \leq 2\,000) = 0,065$

   Pravděpodobnost, že součástka nevydrží více než 2 000 hodin je 0,06. [0,06]

   b) pravděpodobnost, že součástka vydrží více než 35 000 hodin,

   Řešení:

   $P(X > 35\,000) = 0,311$

   Pravděpodobnost, že součástka vydrží více než 35 000 hodin je 0,31. [0,31]

   c) dobu, do níž se porouchá 95 % součástek.

   Řešení:

   $$\begin{array}{ll} P(X < T) = 0,95 & ln(e^{-\lambda T}) = ln(0,05) \\ F(T) = 0,95 & -\lambda T = ln(0,05) \\ 1 - e^{-\lambda T} = 0,95 & \\ 0,05 = e^{-\lambda T} \quad /ln & T = \dfrac{ln(0,05)}{-\lambda} \end{array}$$

   $$T = -ln(0,05) \cdot 30\,000 = 89\,872 \text{ h}$$

   nebo:

   $F(T) = 0,95 \Rightarrow T = x_{0,95} \Rightarrow$ dopočet přes software

   Doba, do níž se porouchá 95 % součástek je 89 872 hodin. [89 872 h]

3. Výrobní zařízení má poruchu v průměru jednou za 2000 hodin. Veličina $Y$ představující dobu čekání na poruchu má exponenciální rozdělení. Určete dobu $T_0$ tak, aby pravděpodobnost, že přístroj bude pracovat delší dobu než $T_0$, byla 0,99.

   Řešení:

   $$Y \ldots \text{doba čekání na poruchu}$$

   $$Y \sim Exp(\lambda)$$

   $\lambda : E(Y) = 2\,000, E(Y) = \frac{1}{\lambda} \Rightarrow \lambda = \frac{1}{2\,000}\ \text{h}^{-1}$

   $P(Y > T_0) = 0,99$ /odečíst od 1
   $1 - P(Y > T_0) = 1 - 0,99$
   $P(Y \leq T_0) = 0,01$
   $P(Y < T_0) = 0,01$
   $F(T_0) = 0,01$

   $1 - e^{-\lambda T_0} = 0,01$
   $e^{-\lambda T_0} = 0,99$ /ln
   $ln(e^{-\lambda T_0}) = ln(0,99)$
   $-\lambda T_0 = ln(0,99)$
   $T_0 = \dfrac{ln(0,99)}{-\lambda}$

   $$T_0 = ln(0,99) \cdot (-2\,000) = 20,1\ \text{h}$$

   nebo:

   $F(T_0) = 0,01 \ \Rightarrow \ T_0 = x_{0,01} \ \Rightarrow$ dopočet přes software

   Přístroj bude pracovat s pravděpodobností 0,99 delší dobu než 20,1 hodin.

   [20,1 h]

4. Výsledky měření jsou zatíženy jen normálně rozdělenou chybou s nulovou střední hodnotou a se směrodatnou odchylkou 3 mm. Jaká je pravděpodobnost, že při 3 měřeních bude alespoň jednou chyba v intervalu (0 mm; 2,4mm)?

   Řešení:

   $$X \ldots \text{počet měření s chybou v intervalu } (0; 2,4) \text{ mm ve 3 měřeních}$$

   $$X \sim Bi(3; \pi = ?)$$

   $$Y \ldots \text{velikost chyby měření}$$

   $$Y \sim N(0\ \text{mm}; 9\ \text{mm}^2)$$

   $\pi = P(0 < Y < 2,4) = 0,288$

   $$X \sim Bi(3; \pi = 0,288)$$

   $P(X \geq 1) = P(X > 0) = 0,639$

   Pravděpodobnost, že při 3 měřeních bude alespoň 1 chyba v intervalu $(0; 2,4)$ mm je 0,64.

   [0,64]

5. Ve velké počítačové síti se průměrně přihlašuje 25 uživatelů za hodinu. Určete pravděpodobnost, že:

   a) se nikdo nepřihlásí během 14:30 - 14:36,

   Řešení:

   $$X \dots \text{počet přihlášených uživatelů za 6 minut}$$
   $$X \sim Po(\lambda t)$$

   $\lambda t$ : 25 za 60 minut $\Rightarrow$ 2,5 za 6 minut $\Rightarrow X \sim Po(2,5)$

   $P(X = 0) = 0,082$

   Pravděpodobnost, že se nikdo nepřihlásí během 14:30 - 14:36 je 0,08.

   [0,08]

   b) do dalšího přihlášení uběhnou 2-3 minuty.

   Řešení:

   $$Y \dots \text{doba do dalšího přihlášení}$$
   $$Y \sim Exp(\lambda)$$

   $\lambda$ : 25 za 60 minut $\Rightarrow$ za 1 minutu: $\frac{25}{60} = 0,42$ (trojčlenka)
   $$Y \sim Exp(0,42)$$

   $P(2 \leq Y \leq 3) = 0,148$

   Pravděpodobnost, že do dalšího přihlášení uběhnou 2-3 minuty je 0,15.

   [0,15]

   c) Určete maximální délku časového intervalu tak, aby pravděpodobnost, že se nikdo nepřihlásí byla alespoň 0,90.

   Řešení:

   $$Z \dots \text{počet přihlášených uživatelů za } t \text{ minut}$$
   $$Z \sim Po(\lambda t = 0,42t)$$

   $$P(Z = 0) \geq 0,9$$
   $$\frac{(\lambda t)^0 e^{-\lambda t}}{0!} \geq 0,9$$
   $$e^{-\lambda t} \geq 0,9 \qquad /ln$$
   $$ln(e^{-\lambda t}) \geq ln(0,9)$$
   $$-\lambda t \geq ln(0,9)$$

   $$\lambda t \leq 0,105$$
   $$t \leq \frac{0,105}{\lambda}$$
   $$t \leq \frac{0,105}{0,42}$$
   $$t \leq 0,25$$

   Maximální délka časového intervalu je 0,25 min (15 s).

   [0,25 min]

6. Náhodná veličina $X$ má normální rozdělení $\mathrm{N}(\mu;\sigma^2)$. Určete:

   a) $P(\mu-2\sigma<X<\mu+2\sigma)$,

   Řešení:

   $$P(\mu-2\sigma<X<\mu+2\sigma)=F(\mu+2\sigma)-F(\mu-2\sigma)=\Phi\left(\frac{\mu+2\sigma-\mu}{\sigma}\right)-\Phi\left(\frac{\mu-2\sigma-\mu}{\sigma}\right)=$$
   $$=\Phi(2)-\Phi(-2)=\Phi(2)-(1-\Phi(2))=2\Phi(2)-1=0,955$$

   [0,955]

   b) $k\in\mathbb{Z}$ tak, aby $P(\mu-k\sigma<X<\mu+k\sigma)>0,99$.

   Řešení:

   $$P(\mu-k\sigma<X<\mu+k\sigma)=F(\mu+k\sigma)-F(\mu-k\sigma)=\Phi\left(\frac{\mu+k\sigma-\mu}{\sigma}\right)-\Phi\left(\frac{\mu-k\sigma-\mu}{\sigma}\right)=$$
   $$=\Phi(k)-\Phi(-k)=\Phi(k)-(1-\Phi(k))=2\Phi(k)-1>0,99$$

   $2\Phi(k)>1,99$

   $\Phi(k)>0,995$

   $\Phi(k)=0,995\Rightarrow k=2,58\Rightarrow k>2,58$

   Pozn. V praxi volíme kvůli snadnějšímu výpočtu $k=3$

   [2,58]

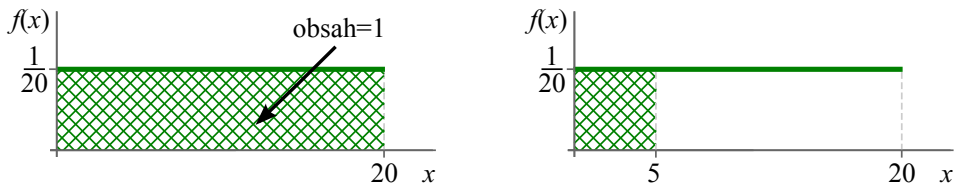
7. Na prohlídce výstavy je promítán doprovodný film o životě autora vystavovaných děl. Jeho projekce začíná každých 20 minut. Určete pravděpodobnost, že pokud náhodně přijdete do promítacího sálu,

   a) nebudete na začátek filmu čekat víc než 5 minut,

   Řešení:

   $X\ldots$ doba čekání na začátek filmu

   $$X\sim Ro(0;20)$$

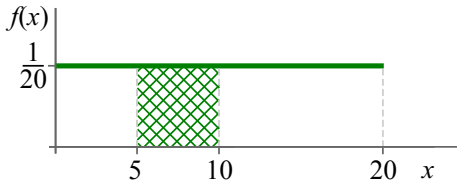
   $P(X\leq 5)=\frac{1}{20}\cdot 5=0,25$

   Pravděpodobnost, že na začátek filmu nebudeme čekat více než 5 minut je 0,25.

   [0,25]

b) budete čekat mezi 5 a 10 minutami,

Řešení:




$P(5 \leq X \leq 10) = \frac{1}{20} \cdot (10 - 5) = 0,25$

Pravděpodobnost, že na začátek filmu budeme čekat mezi 5 a 10 minutami je 0,25.

[0,25]

c) střední hodnotu a směrodatnou odchylku doby čekání na začátek filmu.

Řešení:

$$E(X) = \frac{a+b}{2} = \frac{0+20}{2} = 10,0 \text{ min}$$

$$D(X) = \frac{(a-b)^2}{12} = \frac{(0-20)^2}{12} = \frac{400}{12} \text{ min}^2$$

$$\sigma(X) = \sqrt{D(X)} = \sqrt{\frac{400}{12}} = \frac{20}{2\sqrt{3}} = \frac{10}{\sqrt{3}} = 5,8 \text{ min}$$

[$E(X)$ =10,0 min; $\sigma$ =5,8 min]

8. Při kontrole jakosti přebíráme součástku pouze tehdy, jestliže se její rozměr pohybuje v mezích 26-27 mm. Rozměry součástek mají normální rozdělení se střední hodnotou 26,4 mm a směrodatnou odchylkou 0,2 mm. Jaká je pravděpodobnost, že rozměr součástky náhodně vybrané ke kontrole bude v požadovaných mezích?

Řešení:

$X \ldots$ rozměr součástky

$$X \sim N(26,4 \text{ mm}; 0,2^2 \text{ mm}^2)$$

$P(26 \leq X \leq 27) = P(X < 27) - P(X < 26) = 0,976$

Pravděpodobnost, že rozměr součástky náhodně vybrané ke kontrole bude v požadovaných mezích je 0,98.

[0,98]

9. Délka skoků sportovce Jakuba měřená v cm má normální rozdělení $\mathrm{N}(\mu_1; \sigma_1^2)$, kde $\mu_1 = 690$ a $\sigma_1 = 10$. Délka skoků sportovce Aleše měřená v cm má také normální rozdělení $\mathrm{N}(\mu_2; \sigma_2^2)$, kde $\mu_2 = 705$ a $\sigma_2 = 15$. Na závody se kvalifikuje ten, kdo ze dvou skoků alespoň jednou skočí více než 700 cm.

   a) S jakou pravděpodobností se oba dva kvalifikují na závody?

   Řešení:

   $$X_1 \ldots \text{délka skoku Jakuba } X_1 \sim N(690; 100)$$
   $$X_2 \ldots \text{délka skoku Aleše } X_2 \sim N(705; 225)$$

   $J \ldots$ Jakub se kvalifikuje
   $A \ldots$ Aleš se kvalifikuje
   $O \ldots$ oba se kvalifikují
   $JS \ldots$ Jakub skočí více než 700 cm
   $AS \ldots$ Aleš skočí více než 700 cm

   $P(JS) = P(X_1 > 700) = 0,16 \Rightarrow P(\overline{JS}) = 1 - P(JS) = 0,84$

   $P(J) = P(JS \cap \overline{JS}) + P(\overline{JS} \cap JS) + P(JS \cap JS) = 0,16 \cdot 0,84 + 0,84 \cdot 0,16 + 0,16 \cdot 0,16 = 0,29$

   $P(AS) = P(X_2 > 700) = 0,63 \Rightarrow P(\overline{AS}) = 1 - P(AS) = 0,37$

   $P(A) = P(AS \cap \overline{AS}) + P(\overline{AS} \cap AS) + P(AS \cap AS) = 0,63 \cdot 0,37 + 0,37 \cdot 0,63 + 0,63 \cdot 0,63 = 0,86$

   $P(O) = P(J \cap A) = P(J) \cdot P(A) = 0,29 \cdot 0,86 = 0,249$

   jiný způsob výpočtu pro $P(J)$ a $P(A)$ (vhodný např. při vyšším poču pokusů než 2):

   $SJ \ldots$ počet úspěšných pokusů Jakuba ze 2 $\Rightarrow$ SJ $\sim Bi(n = 2; \pi = P(JS) = 0,16)$
   $SA \ldots$ počet úspěšných pokusů Aleše ze 2 $\Rightarrow$ AJ $\sim Bi(n = 2; \pi = P(AS) = 0,63)$

   $P(J) = P(SJ \geq 1)$

   $P(A) = P(SA \geq 1)$

   Pravděpodobnost, že se oba kvalifikují je 0,25.

   [0,25 ]

   b) S jakou pravděpodobností se kvalifikuje Aleš, ale Jakub ne?

   Řešení:

   $$P(A \cap \overline{J}) = P(A) \cdot P(\overline{J}) = P(A) \cdot (1 - P(J)) = 0,86 \cdot (1 - 0,29) = 0,611$$

   Pravděpodobnost, že se kvalifikuje pouze Aleš, ale Jakub ne, je 0,61.

   [0,61]

# 7. Explorační analýza dat

Kvantitativní - Numerická proměnná

- Míry polohy

| | |
|---|---|
| Průměr | $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ |
| Kvantily | dolní kvartil, horní kvartil, medián |

- Míry variability

| | |
|---|---|
| Variační rozpětí | $x_{max} - x_{min}$ |
| Interkvartilové rozpětí | $IQR = x_{0,75} - x_{0,25}$ |
| Výběrový rozptyl | $s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$ |
| Výběrová směrodatná odchylka | $s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$ |
| Variační koeficient | $V_x = \frac{s}{\overline{x}}$, popř. $V_x = \frac{s}{\overline{x}} \cdot 100$ (%) |

- Míry šikmosti a špičatosti

| | |
|---|---|
| Výběrová šikmost | $\alpha = \frac{n}{(n-1)(n-2)} \cdot \frac{\sum_{i=1}^{n} (x_i - \overline{x})^3}{s^3}$ |
| Výběrová špičatost | $\beta = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \cdot \frac{\sum_{i=1}^{n} (x_i - \overline{x})^4}{s^4} - 3\frac{(n-1)^2}{(n-2)(n-3)}$ |

- Identifikace odlehlých pozorování

| | |
|---|---|
| Vnitřní hradby | dolní mez: $h_D = x_{0,25} - 1,5 \cdot IQR$<br>horní mez: $h_H = x_{0,75} + 1,5 \cdot IQR$ |
| Z – souřadnice | z-skóre $= \frac{x_i - \overline{x}}{s}$ |
| Mediánová souřadnice | $x_{0,5}$-skóre $= \frac{x_i - x_{0,5}}{1,483 \cdot MAD}$ |

## 7.1 Test z teorie

1. Test ze Statistiky píše velké množství studentů. Představte si, že každý z nich odpoví správně přesně na polovinu otázek. V tomto případě bude směrodatná odchylka počtu správných odpovědí

   a) rovna průměru,
   b) rovna mediánu,
   c) rovna nule,
   d) směrodatnou odchylku nelze určit bez dalších informací.

2. Největší kumulativní absolutní četnost v množině čísel se rovná

   a) součtu všech absolutních četností,
   b) 1,
   c) dvojnásobku průměru,
   d) dvojnásobku mediánu,
   e) dvojnásobku módu.

3. Několik studentů píše test ze Statistiky s 10 otázkami. Nejhorší výsledek jsou 3 správné odpovědi, nejlepší výsledek je 10 správných odpovědí. Jakou hodnotu má medián?

   a) 7 = (10-3),
   b) 6,5 = $\frac{3+10}{2}$,
   c) medián nelze určit, pokud neznáme konkrétní výsledky jednotlivých žáků.

4. Představte si, že jste absolvovali normovaný test (např. SCIO test) a že Vám sdělili, že patříte do 91. percentilu. To znamená, že

   a) 90 žáků, kteří se podrobili stejnému testu, dosáhlo vyšších výsledků než vy,
   b) 90 žáků, kteří se podrobili stejnému testu, dosáhlo nižších výsledků než vy,
   c) 90 % žáků, kteří se podrobili stejnému testu, dosáhlo vyšších výsledků než vy,
   d) 90 % žáků, kteří se podrobili stejnému testu, dosáhlo nižších výsledků než vy.

5. Průměrná mzda je 60% kvantil mzdy. Lze tedy říci, že

   a) medián mzdy je vyšší než průměrná mzda,
   b) medián mzdy je nižší než průměrná mzda,
   c) medián mzdy je stejný jako průměrná mzda,
   d) o vztahu mezi mediánem mzdy a průměrnou mzdou nelze rozhodnout.

6. Průměrná mzda je 60% kvantil mzdy. Lze tedy říci, že

   a) mzdy mají kladnou šikmost,
   b) mzdy mají zápornou šikmost,
   c) mzdy mají kladnou špičatost,
   d) mzdy mají zápornou špičatost,
   e) vztah mezi průměrem a 60% kvantilem nevypovídá nic o šikmosti ani o špičatosti dat.

7. Lékař Petře sdělil, že patří do 3. percentilu ohledně BMI (Body mass index - poměr váhy (kg) ke kvadrátu výšky (m)). Petra má pravděpodobně

   a) podváhu,
   b) normální váhu,
   c) nadváhu,
   d) bez dalších informací nelze usuzovat na Petřinu váhu.

8. Představte si, že jste absolvovali normovaný test (např. SCIO test). Měl(a) jste lepší výsledek než 85 studentů ze 100. To znamená, že

   a) patříte do 99. decilu,
   b) patříte do 95. decilu,
   c) patříte do 10. decilu,
   d) patříte do 9. decilu,
   e) patříte do 2. kvartilu.

9. Pro srovnání variability váhy a výšky je možné použít

   a) průměr,
   b) rozptyl,
   c) směrodatnou odchylku,
   d) variační koeficient,
   e) šikmost.

10. Zvýšíme-li každému zaměstnanci ve firmě plat o 100,- Kč, průměrný plat ve firmě se zvýší

    a) o 100,- Kč,
    b) o 1 000,- Kč,
    c) průměrný plat se nezmění.

11. Zvýšíme-li každému zaměstnanci ve firmě plat dvojnásobně, průměrný plat ve firmě se zvýší

    a) dvojnásobně,
    b) čtyřnásobně,
    c) průměrný plat se nezmění.

12. Zvýšíme-li každému zaměstnanci ve firmě plat o 20%, průměrný plat ve firmě se zvýší

    a) o 20%,
    b) o 400%,
    c) o 40%,
    d) o 44%,
    e) průměrný plat se nezmění.

13. Zvýšíme-li každému zaměstnanci ve firmě plat o 100,- Kč, rozptyl platů ve firmě se zvýší

    a) o 100,- Kč,
    b) o 1 000,- Kč,
    c) rozptyl platů se nezmění.

14. Zvýšíme-li každému zaměstnanci ve firmě plat dvojnásobně, rozptyl platů ve firmě se zvýší

a) dvojnásobně,
b) čtyřnásobně,
c) průměrný plat se nezmění.

15. Zvýšíme-li každému zaměstnanci ve firmě plat o 20%, rozptyl platů ve firmě se zvýší

a) o 20%,
b) o 400%,
c) o 40%,
d) o 44%,
e) průměrný plat se nezmění.

16. Největší kumulativní relativní četnost se rovná

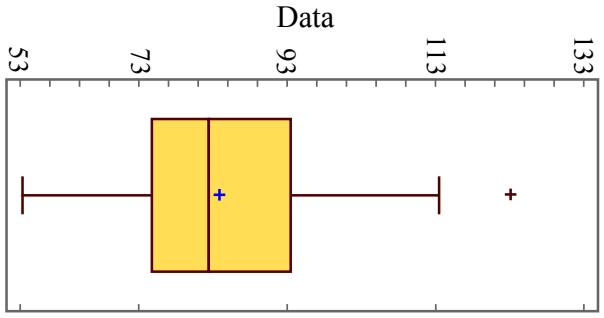
a) dvojnásobku průměru,
b) dvojnásobku mediánu,
c) dvojnásobku módu,
d) součtu všech jednotlivých hodnot absolutních četností,
e) 1.

17. Určete, zda jsou následující tvrzení pravdivá.

P a) Geometrický průměr je definován pro proměnné, které nabývají pouze kladných hodnot.
N b) Jedna čtvrtina hodnot je větší než 25% kvantil, zatímco tři čtvrtiny hodnot jsou menší.
P c) Mají-li dvě proměnné stejný průměr a stejný rozptyl, mají stejný variační koeficient.
P d) Mzdy v ČR mají kladnou šikmost. (V ČR mají zhruba 2/3 lidí podprůměrný plat.)
N e) Nejčetnější hodnota v souboru se nazývá medián. [modus]
N f) Rozptyl má vždy kladnou hodnotu. [nebo nulovou]

18. V grafu na obrázku, modrý křížek označuje

a) medián,
b) průměr,
c) modus,
d) interkvartilové rozpětí (IQR).

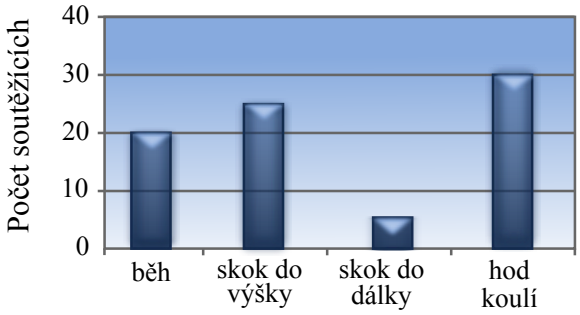



19. Určete, zda jsou následující tvrzení pravdivá. Proměnná znázorněna na předcházejícím obrázku

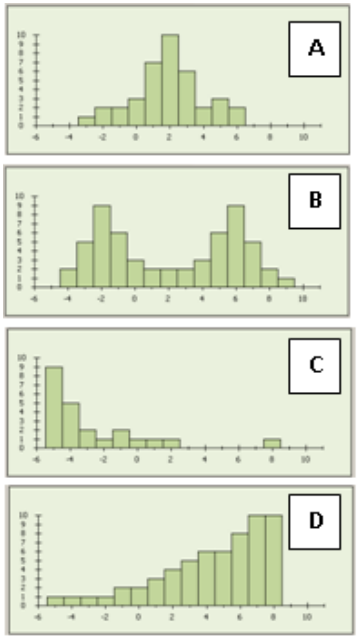
N a) neobsahuje odlehlá pozorování,
P b) má kladnou šikmost,
P c) je kladná,
N d) má více než polovinu hodnot větších než 83.

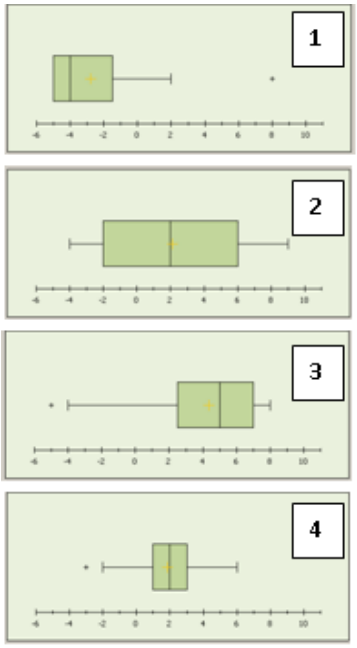
20. Na atletických závodech mládeže žáci soutěžili ve 4 kategoriích. Určete pravdivost výroků.

N a) Na obrázku je znázorněn histogram a nejméně soutěžících bylo ve skoku do dálky. [sloupcový graf]
P a) Celkem ve čtyřech kategoriích soutěžilo 80 žáků.
P a) Modus = hod koulí.
N d) Modus = 30.




21. Spárujte histogramy s odpovídajícími krabicovými grafy.




a-4, b-2, c-1, d-3

## 7.2 Příklady

1. V jistém supermarketu byla ve stejné chvíli na 8 pokladnách měřena doba, během které pokladní ověří platnost platební karty zákazníka v bance. U pěti zákazníků trvalo ověření 2 minuty, u zbývajících tří to byly 3 minuty. Určete průměrnou dobu potřebnou k ověření platnosti karty.

   Řešení:

   Jedná se o úlohu o společné práci ⇒ harmonický průměr.

   $$\overline{x}_H = \frac{n}{\sum\limits_{i=1}^{n} \frac{1}{x_i}} = \frac{8}{\frac{5}{2} + \frac{3}{3}} = \frac{48}{21} = 2,29$$

   Průměrná doba potřebná k ověření platnosti karty je 2,29 minut.

   [2,29 minut]

2. Při sledování proměnné $x$ byl určen aritmetický průměr 110 a rozptyl 800. Dodatečně byly zjištěny chyby u dvou údajů. Místo 85 mělo být správně 95 a místo 120 má být 150. Ostatních 18 údajů bylo správných. Opravte vypočítané charakteristiky (průměr a rozptyl).

   Řešení:

   a) Průměr:

   Původně:

   $$\left.\begin{array}{l} x_1 = ? \\ x_2 = ? \\ \vdots \\ x_{18} = ? \\ x_{19} = 85 \\ x_{20} = 120 \\ \overline{x} = 110 \end{array}\right\} \begin{array}{l} \sum\limits_{i=1}^{20} x_i = n \cdot \overline{x} = 20 \cdot 110 = \\ 2\,200 \\ \\ \sum\limits_{i=1}^{18} x_i = \sum\limits_{i=1}^{20} x_i - x_{19} - x_{20} = \\ 2\,200 - 85 - 120 = 1\,995 \end{array}$$

   Opraveno:

   $$\left.\begin{array}{l} y_1 = x_1 \\ y_2 = x_2 \\ \vdots \\ y_{18} = x_{18} \\ y_{19} = 95 \\ y_{20} = 150 \end{array}\right\} \begin{array}{l} \sum\limits_{i=1}^{18} y_i = \sum\limits_{i=1}^{18} x_i = 1\,995 \\ \\ \sum\limits_{i=1}^{20} y_i = \sum\limits_{i=1}^{18} y_i + y_{19} + y_{20} = \\ 1\,995 + 95 + 150 = 2\,240 \end{array}$$

   $$\overline{y} = \frac{\sum\limits_{i=1}^{20} y_i}{n} = \frac{2\,240}{20} = 112$$

   Správná hodnota aritmetického průměru je 112.

   [112]

   b) Rozptyl:

   $$s^2 = \frac{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2}{n-1} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i^2 - 2\overline{x}x_i + \overline{x}^2) = \frac{1}{n-1} \cdot \left( \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2\overline{x}x_i + \sum_{i=1}^{n} \overline{x}^2 \right) =$$

   $$\frac{1}{n-1} \cdot \left( \sum_{i=1}^{n} x_i^2 - 2\overline{x}n\overline{x} + n\overline{x}^2 \right) = \frac{\sum\limits_{i=1}^{n} x_i^2 - n\overline{x}^2}{n-1}$$

$$\sum_{i=1}^{n} x_i^2 = s^2(n-1) + n\overline{x}^2$$

Původně:

$$\left.\begin{array}{l} x_1 = ? \\ x_2 = ? \\ \vdots \\ x_{18} = ? \\ x_{19} = 85 \\ x_{20} = 120 \\ \overline{x} = 110 \\ s_x^2 = 800 \end{array}\right\} \quad \begin{array}{l} \sum_{i=1}^{20} x_i^2 = 800 \cdot 19 + 20 \cdot 110^2 \\ 257\,200 \\ \sum_{i=1}^{18} x_i^2 = \sum_{i=1}^{20} x_i^2 - x_{19}^2 - x_{20}^2 = \\ 257\,200 - 85^2 - 120^2 = \\ 235\,575 \end{array}$$

Opraveno:

$$\left.\begin{array}{l} y_1 = x_1 \\ y_2 = x_2 \\ \vdots \\ y_{18} = x_{18} \\ y_{19} = 95 \\ y_{20} = 150 \\ \overline{y} = 112 \end{array}\right\} \quad \begin{array}{l} \sum_{i=1}^{18} y_i^2 = \sum_{i=1}^{18} x_i^2 = 235\,575 \\ \sum_{i=1}^{20} y_i^2 = \sum_{i=1}^{18} y_i^2 + y_{19}^2 + \\ y_{20}^2 = 235\,575 + 95^2 \\ +150^2 = 267\,100 \end{array}$$

$$s_y^2 = \frac{\sum_{i=1}^{n} y_i^2 - n_y \overline{y}^2}{n_y - 1} = \frac{267\,100 - 20 \cdot 112^2}{19} \cong 853,7$$

Správná hodnota rozptylu je 853,7.

[853,7]

3. Ze čtyřiceti hodnot byl vypočítán aritmetický průměr 7,50 a rozptyl 2,25. Při kontrole bylo zjištěno, že chybí dvě hodnoty proměnné: $-3,8$ a 7. Opravte uvedené charakteristiky.

   Řešení:

   a) Průměr:

Původně:

$$\left.\begin{array}{l} x_1 = ? \\ x_2 = ? \\ \vdots \\ x_{40} = ? \\ \overline{x} = 7,5 \\ n_x = 40 \end{array}\right\} \quad \begin{array}{l} \sum_{i=1}^{40} x_i = n_x \cdot \overline{x} = 40 \cdot 7,5 = \\ 300 \end{array}$$

Opraveno:

$$\left.\begin{array}{l} y_1 = x_1 \\ y_2 = x_2 \\ \vdots \\ y_{40} = x_{40} \\ y_{41} = -3,8 \\ y_{42} = 7 \\ n_y = 42 \end{array}\right\} \quad \begin{array}{l} \sum_{i=1}^{42} y_i = \sum_{i=1}^{40} x_i + y_{41} + y_{42} = \\ = 300 + (-3,8) + 7 = \\ 303,2 \end{array}$$

$$\overline{y} = \frac{\sum_{i=1}^{42} y_i}{n_y} = \frac{303,2}{42} \cong 7,2$$

Správná hodnota aritmetického průměru je 7,2

[7,2]

b) Rozptyl:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1} = \frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i^2-2\overline{x}x_i+\overline{x}^2) = \frac{1}{n-1}\cdot\left(\sum_{i=1}^{n}x_i^2-\sum_{i=1}^{n}2\overline{x}x_i+\sum_{i=1}^{n}\overline{x}^2\right) =$$
$$\frac{1}{n-1}\cdot\left(\sum_{i=1}^{n}x_i^2-2\overline{x}n\overline{x}+n\overline{x}^2\right) = \frac{\sum_{i=1}^{n}x_i^2-n\overline{x}^2}{n-1}$$

$$\sum_{i=1}^{n}x_i^2 = s^2(n-1)+n\overline{x}^2$$

Původně:

$$\left.\begin{array}{l} x_1 = ? \\ x_2 = ? \\ \vdots \\ x_{40} = ? \\ \overline{x} = 7,5 \\ s_x^2 = 2,25 \\ n_x = 40 \end{array}\right\} \sum_{i=1}^{40}x_i^2 = 2,25\cdot 39 + 40\cdot 7,5^2 = 2\,337,75$$

Opraveno:

$$\left.\begin{array}{l} y_1 = x_1 \\ y_2 = x_2 = ? \\ \vdots \\ y_{40} = x_{40} \\ y_{41} = -3,5 \\ y_{42} = 7 \\ \overline{y} = 7,2 \\ n_y = 42 \end{array}\right\} \begin{array}{l} \sum_{i=1}^{42}y_i^2 = \sum_{i=1}^{40}x_i^2 + y_{41}^2 + \\ y_{42}^2 = 2\,337,75 + (-3,8)^2 \\ +7^2 = 2\,401,19 \end{array}$$

$$s_y^2 = \frac{\sum_{i=1}^{n}y_i^2 - n_y\overline{y}^2}{n_y-1} = \frac{2\,401,19 - 42\cdot 7,2^2}{41} \cong 5,5$$

Správná hodnota rozptylu je 5,5.

[5,5]

4. V důsledku výstavby satelitního městečka poklesl průměrný věk obyvatel vesnice o 19%, rozptyl věku vzrostl o 21%. Jak se změnil variační koeficient?

Řešení:

$\overline{x}_S$ ... starý průměr
$s_S^2$ ... starý rozptyl
$V_S$ ... starý variační koeficient
$\overline{x}_N$ ... nový průměr
$s_N^2$ ... nový rozptyl
$V_N$ ... nový variační koeficient

$$V_S = \frac{\sqrt{s_S^2}}{\overline{x}_S}$$

$$\overline{x}_N = 0,81\cdot\overline{x}_S$$

$$s_N^2 = 1,21\cdot s_N^2$$

$$V_N = \frac{\sqrt{s_N^2}}{\overline{x}_N} = \frac{\sqrt{1,21\cdot s_S^2}}{0,81\cdot\overline{x}_S} = \frac{\sqrt{1,21}}{0,81}\cdot\frac{\sqrt{s_S^2}}{\overline{x}_S} = 1,358\cdot V_S$$

Variační koeficient vzrostl o 35,8 %.

[35,8 %]

# 8. Výběrové charakteristiky

Mějme náhodný výběr $\boldsymbol{X}$ z normálního rozdělení:

$$\boldsymbol{X} = (X_1, \ldots, X_n), \forall i = 1, \ldots, n : X_i \to N(\mu, \sigma).$$

| Výběrová charakteristika | Rozdělení pravděpodobnosti | Poznámka |
|---|---|---|
| $\frac{\overline{X} - \mu}{\sigma}\sqrt{n}$ | $N(0;1)$ | viz CLV |
| $\frac{\overline{X} - \mu}{S}\sqrt{n}$ | $t_{n-1}$ | viz vlastnosti Studentova rozdělení |
| $\frac{S^2}{\sigma^2}(n-1)$ | $\chi^2_{n-1}$ | viz vlastnosti $\chi^2$ - rozdělení |
| $\frac{p - \pi}{\sqrt{\pi(1-\pi)}}\sqrt{n}$ | $N(0;1)$ | viz vlastnosti realtivní četnosti, předpoklad: $n > \frac{9}{p(1-p)}$ |

Mějme dva nezávislé výběry z normálního rozdělení:

$$\forall i = 1, 2, \ldots, n_1, \text{ kde } n_1 \text{ je rozsah prvního výběru: } X_{1i} \to N(\mu_1, \sigma_1^2)$$
$$\forall j = 1, 2, \ldots, n_2, \text{ kde } n_2 \text{ je rozsah druhého výběru: } X_{1j} \to N(\mu_2, \sigma_2^2)$$

| Výběrová charakteristika | Rozdělení pravděpodobnosti | Poznámka |
|---|---|---|
| $\frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ | $N(0;1)$ | viz CLV |
| $\frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{S_1^2(n_1-1) + S_2^2(n_2-1)}}\sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}$ | $t_{n_1+n_2-2}$ | viz vlastnosti Studentova rozdělení předpoklad: $\sigma_1^2 = \sigma_2^2$ |
| $\frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$ | $t_\nu$ $\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{S_1^2}{n_1}\right)^2 \frac{1}{n_1+1} + \left(\frac{S_2^2}{n_2}\right)^2 \frac{1}{n_2+1}} - 2$ | viz vlastnosti Studentova rozdělení předpoklad: $\sigma_1^2 \neq \sigma_2^2$ |
| $\frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}}$ | $F_{n_1-1;n_2-1}$ | viz vlastnosti Fisher-Snedecorova rozdělení |
| $\frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}}$ | $N(0;1)$ | viz CLV předpoklad: $\left(n_1 > \frac{9}{p_1(1-p_1)}\right) \wedge \left(n_2 > \frac{9}{p_2(1-p_2)}\right)$ |

## 8.1 Test z teorie

1. Střední hodnota pevně zvolené náhodné veličiny je
   a) náhodná veličina,
   b) konstanta,
   c) náhodný jev,
   d) výběrová charakteristika.

2. Výběrový průměr je
   a) náhodná veličina,
   b) konstanta,
   c) náhodný jev,
   d) výběrová charakteristika.

3. S rostoucím rozsahem výběru se obvykle rozptyl průměru
   a) snižuje,
   b) zvyšuje,
   c) nemění.

4. Statistická indukce je
   a) experiment,
   b) metoda, která umožňuje odhadnout vlastnosti výběru na základě znalostí vlastností populace,
   c) zobecnění statistických výsledků získaných zpracováním výběru na celou populaci,
   d) metoda sběru dat.

5. Zákon velkých čísel v důsledku říká, že při dostatečném rozsahu výběru
   a) má průměr normální rozdělení,
   b) má průměr Studentovo rozdělení,
   c) se střední hodnota přibližuje teoretické hodnotě průměru,
   d) se relativní četnost přibližuje teoretické hodnotě pravděpodobnosti.

6. Pro modelování průměru výběru dostatečně velkého rozsahu je vhodné použít rozdělení
   a) normální,
   b) Pearsonovo ($\chi^2$ ),
   c) Studentovo,
   d) Fisherovo-Snedecorovo.

7. Pro modelování průměru výběru malého rozsahu je vhodné použít rozdělení
   a) normální,
   b) Pearsonovo ($\chi^2$ ),
   c) Studentovo,
   d) Fisherovo-Snedecorovo.

8. Pro modelování relativní četnosti ve výběru o dostatečném rozsahu je vhodné použít rozdělení

   a) normální,
   b) Pearsonovo ($\chi^2$ ),
   c) Studentovo,
   d) Fisherovo-Snedecorovo.

9. Pro modelování rozptylu výběru z normálního rozdělení je vhodné použít rozdělení

   a) normální,
   b) Pearsonovo ($\chi^2$ ),
   c) Studentovo,
   d) Fisherovo-Snedecorovo.

10. Pro modelování poměru rozptylů dvou výběrů z normálního rozdělení je vhodné použít rozdělení

    a) normální,
    b) Pearsonovo ($\chi^2$ ),
    c) Studentovo,
    d) Fisherovo-Snedecorovo.

## 8.2 Příklady

1. Zatížení letadla s 64 místy nemá překročit 6 000 kg. Jaká je pravděpodobnost, že při plném obsazení bude tato hodnota překročena, má-li hmotnost cestujícího střední hodnotu 90 kg a směrodatnou odchylku 10 kg?

   Řešení:

   $$X_i \ldots \text{hmotnost } i\text{-tého cestujícího}$$

   $$E(X_i) = 90 \text{ kg } D(X_i) = 100 \text{ kg}^2$$

   $$X \ldots \text{hmotnost 64 cestujících}$$

   $$X = \sum_{i=1}^{n} X_i, n = 64 \qquad \Rightarrow \qquad X \sim N(64 \cdot 90; 64 \cdot 100)$$

   $P(X > 6\,000) \cong 0,001$

   Pravděpodobnost, že při plném obsazení bude hodnota 6 000 kg překročena je 0,001.

   [0,001]

2. Zásilka obsahuje 300 výrobků určitého typu. Je známo, že pravděpodobnost zhotovení vadného výrobku tohoto typu je 0,04.

   a) Odhadněte pravděpodobnost, že absolutní odchylka podílu vadných výrobků v zásilce od pravděpodobnosti vyrobení vadného výrobku bude menší než 1 %.

   Řešení:

   $$p \ldots \text{podíl vadných součástek}$$

   $$\pi \ldots \text{pravděpodobnost zhotovení vadného výrobku} = 0,04$$

   $P(|p - \pi| < 0,01) = ?$

   neznáme rozdělení $|p - \pi| \Rightarrow$ je nutné převést na veličinu, jejíž rozdělení známe

   $$X = \frac{p - \pi}{\sqrt{\pi(1 - \pi)}}\sqrt{n} \sim N(0, 1)$$

   $$P(|p - \pi| < 0,01) = P(-0,01 < p - \pi < 0,01) =$$

   $$= P\left(\frac{-0,01}{\sqrt{\pi(1 - \pi)}}\sqrt{n} < \frac{p - \pi}{\sqrt{\pi(1 - \pi)}}\sqrt{n} < \frac{-0,01}{\sqrt{\pi(1 - \pi)}}\sqrt{n}\right) =$$

   $$= P\left(\frac{-0,01}{\sqrt{0,04(1 - 0,04)}}\sqrt{300} < \frac{p - \pi}{\sqrt{\pi(1 - \pi)}}\sqrt{n} < \frac{-0,01}{\sqrt{0,04(1 - 0,04)}}\sqrt{300}\right) =$$

   $$= P\left(-0,883 < \frac{p - \pi}{\sqrt{\pi(1 - \pi)}}\sqrt{n} < 0,883\right) = P(-0,883 < X < 0,883) = 0,623$$

   Pravděpodobnost, že absolutní odchylka podílu vadných výrobků v zásilce od pravděpodobnosti vyrobení vadného výrobku bude menší než 1 % je 0,623.

   [0,623]

b) Jak se změní výsledek, jestliže zásilka bude obsahovat 3 000 výrobků?

Řešení:

$$P\left(\frac{-0,01}{\sqrt{0,04(1-0,04)}}\sqrt{3\,000} < \frac{p-\pi}{\sqrt{\pi(1-\pi)}}\sqrt{n} < \frac{-0,01}{\sqrt{0,04(1-0,04)}}\sqrt{3\,000}\right) =$$

$$= P\left(-2,795 < \frac{p-\pi}{\sqrt{\pi(1-\pi)}}\sqrt{n} < 2,795\right) = 0,994$$

Pravděpodobnost, že absolutní odchylka podílu vadných výrobků v zásilce od pravděpodobnosti vyrobení vadného výrobku bude menší než 1 % je 0,994.

[0,994]

3. Cestující pravidelně jezdí do zaměstnání a zpět MHD. Je známo, že doba čekání na příjezd MHD se pohybuje v mezích od 0 do 3 minut. Jaká je pravděpodobnost, že celková doba čekání zaměstnance na příjezd MHD během 23 pracovních dnů bude kratší než 80 minut?

Řešení:

$X_i \ldots$ doba $i$-tého čekání na MHD

$X_i \sim R(0;3)$

$E(X_i) = \frac{0+3}{2} = 1,5$ min $\qquad D(X_i) = \frac{(0-3)^2}{12} = 0,75$ min$^2$

$X \ldots$ celková doba čekání během 23 dnů (cesta tam a zpět $\Rightarrow$ 46 čekání)

$X = \sum_{i=1}^{n} X_i, n = 46 \quad \Rightarrow \quad X \sim N(46 \cdot 1,5; 46 \cdot 0,75)$, tj. $\quad \Rightarrow \quad X \sim N(69; 34,5)$

$P(X < 80) = 0,970$

Pravděpodobnost, že celková doba čekání zaměstnance bude kratší než 80 minut je 0,970.

[0,970]

4. Předpokládejme, že průměrná spotřeba elektrické energie domácností v určitém městě v lednu je 120 kWh a směrodatná odchylka spotřeby je 100 kWh. Určete pravděpodobnost, že průměrná spotřeba 100 náhodně vybraných domácností bude větší než 140 kWh.

Řešení:

$X_i \ldots$ spotřeba $i$-té domácnosti

$E(X_i) = 120$ kWh $D(X_i) = 100$ kWh$^2$

$X \ldots$ průměrná spotřeba 100 domácností

$$X = \frac{\sum_{i=1}^{100} X_i}{n}, n = 100 \qquad \Rightarrow \qquad X \sim N\left(120; \frac{100^2}{100}\right)$$

$P(X > 140) = 0,023$

nebo

$$Y \dots \text{celková spotřeba 100 domácností}$$

$$Y = \sum_{i=1}^{100} X_i, n = 100 \quad \Rightarrow \quad Y \sim N(100 \cdot 120; 100 \cdot 100^2), \text{ tj.} \quad \Rightarrow \quad Y \sim N(12\,000; 10^6)$$

$$P(X > 14\,000) = 0,023$$

Pravděpodobnost, že průměrná spotřeba 100 náhodně vybraných domácností bude větší než 140 kWh je 0,023.

[0,023]

5. Společnost Acme Battery Company vyvinula nový typ baterie mobilních telefonů. V průměru vydrží baterie 60 minut na jedno nabití. Směrodatná odchylka této doby je 4 minuty. Předpokládejme, že výrobní oddělení po 6 měsících spustí test kontroly kvality. Provedli dva náhodné výběry o rozsahu 10 baterií a v obou zjistili směrodatnou odchylku výdrže baterií větší než 6 minut. S jakou pravděpodobností takový výsledek mohli očekávat?

Řešení:

$$S \dots \text{výběrová směrodatná odchylka životnosti baterií}$$

$$J \dots \text{v 1 výběru bude } S > 6$$

$$O \dots \text{v obou výběrech bude } S > 6$$

$P(S > 0,6) = ?$

neznáme rozdělení $S \Rightarrow$ je nutné převést na veličinu, jejíž rozdělení známe

$$X = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}, \text{ tj. } \chi^2_9$$

$$P(J) = P(S > 6) = P\left(\frac{(n-1)S^2}{\sigma^2} > \frac{(n-1)6^2}{\sigma^2}\right) = P\left(\frac{(n-1)S^2}{\sigma^2} > \frac{(10-1)6^2}{4^2}\right) =$$

$$= P(X > 20,25) = 0,016$$

$$P(O) = P(J \cap J) = P(J) \cdot P(J) = 0,016 \cdot 0,016 = 0,000\,256$$

Výsledek mohli očekávat s pravděpodobností $0,000\,256$.

$[0,000\,256]$

6. Z úmrtnostních tabulek vyplývá pravděpodobnost 0,99, že se 35 - letý muž dožije dalšího roku. Roční pojistné této věkové skupiny činí 2 000 Kč, v případě úmrtí pojišťovna vyplatí 100 000 Kč. Jaká je pravděpodobnost, že zisk z 500 pojištěných mužů ve věku 35 let bude alespoň 500 000 Kč? (Řešte dvěma způsoby - pomocí binomického rozdělení a pomoci aproximace binomického rozdělení rozdělením normálním.)

$$X \dots \text{počet mužů z 500, kteří se nedožijí dalšího roku}$$

a) Řešení:

$$X \sim Bi(500; 0,01)$$

$$Z \ldots \text{zisk}$$

$$Z = 500 \cdot 2\,000 - X \cdot 100\,000$$

$$P(Z \geq 500\,000) = P(500 \cdot 2\,000 - X \cdot 100\,000 \geq 500\,000) = P\left(-X \geq \frac{500\,000 - 1\,000\,000}{100\,000}\right) =$$

$$P(-X \geq -5) = P(X \leq 5) = P(X < 6) = 0,616$$

Pravděpodobnost, že zisk pojišťovny bude alespoň 500 000 Kč, je 0,616.

[0,616]

b) Řešení:

$$X \sim Bi(500; 0,01) \approx N(500 \cdot 0,01; 500 \cdot 0,01 \cdot (1 - 0,01))$$

$$\text{tj. } X \approx N(5; 4,95)$$

$$P(Z \geq 500\,000) = P(500 \cdot 2\,000 - X \cdot 100\,000 \geq 500\,000) = P\left(-X \geq \frac{500\,000 - 1\,000\,000}{100\,000}\right) =$$

$$P(X \leq 5) \overset{\text{oprava na spojitost}}{=} P(X < 5,5) = 0,589$$

Pravděpodobnost, že zisk pojišťovny bude alespoň 500 000 Kč, je 0,589.

[0,589]

7. Předpokládejme, že v populaci má přibližně 60 % mladých mužů vyšší než doporučenou hladinu cholesterolu v séru. S jakou pravděpodobností bude mít v náhodném výběru 200 mladých mužů více než 120 z nich vyšší než doporučenou hladinu cholesterolu v séru?

$X \ldots$ počet mladých mužů z 200 s vyšší než doporučenou hladinou cholesterolu v séru

a) Řešení:

$$X \sim Bi(200; 0,6)$$

$$P(X > 120) = 0,470$$

S pravděpodobnostní 0,470 bude mít více než 120 mladých mužů z 200 vyšší než doporučenou hladinu cholesterolu.

[0,470]

b) Řešení:

$$X \approx N(200 \cdot 0,6; 200 \cdot 0,6 \cdot 0,4), \text{ tj. } X \approx N(120; 48)$$

$P(X > 120) = 0,5$ (lze určit zpaměti)

S pravděpodobnostní 0,5 bude mít více než 120 mladých mužů z 200 vyšší než doporučenou hladinu cholesterolu.

[0,5]

# 9. Intervalové odhady

viz applet

## 9.1 Test z teorie

1. Chceme-li najít nejlepší možný odhad směrodatné odchylky vybrané vlastnosti nekonečné populace, měli bychom

   a) použít co možná největší výběrový soubor,
   b) použít co možná nejmenší výběrový soubor,
   c) zjistit hodnotu sledované vlastnosti u všech prvků populace,
   d) použít výběrový soubor o rozsahu nejvýše 10 000 prvků populace.

2. Chceme-li najít nejlepší možný odhad směrodatné odchylky vybrané vlastnosti populace o rozsahu 50 000 jednotek (prvků), pak by rozsah výběru neměl překročit

   a) 49 999 jednotek,
   b) 10 000 jednotek,
   c) 5 000 jednotek,
   d) 2 500 jednotek, $[n \leq 0,05N]$
   e) 1 000 jednotek.

3. Doplňte:

   a) Průměr je (*náhodná veličina*, *konstanta*).

   b) Střední hodnota je (*výběrová*, *populační*) charakteristika.

   c) Odhadujeme-li populační charakteristiku jedním číslem, hovoříme o (*bodovém*, *intervalovém*) odhadu.

   d) Řekneme, že odhad je (*nestranný*, *vydatný*, *konzistentní*), jestliže se jeho střední hodnota rovná hledanému parametru.

   e) Nestranný odhad, jehož rozptyl je (*nejmenší*, *největší*) mezi rozptyly všech nestranných odhadů příslušného parametru, se nazývá nejlepší nestranný odhad.

   f) Mějme náhodný výběr. S rostoucí spolehlivostí odhadu $1-\alpha$ se obvykle intervalové odhady populačních parametrů (*zužují*, *rozšiřují*).

   g) S rostoucí spolehlivostí odhadu $1-\alpha$ (*roste*, *klesá*) hladina významnosti $\alpha$.

   h) Při dané spolehlivosti odhadu $1-\alpha$ se obvykle intervalové odhady populačních parametrů s rostoucím rozsahem výběru (*zužují*, *rozšiřují*).

   i) V technické praxi se obvykle volí spolehlivost odhadu $1-\alpha$ rovna (*0,80*; *0,90*; *0,95*;*0,99*; *0,20*; *0,10*; *0,05*; *0,01*).

   j) V technické praxi se obvykle volí hladina významnosti $\alpha$ rovna (*0,80*; *0,90*; *0,95*;*0,99*; *0,20*; *0,10*; *0,05*; *0,01*).

   k) Horní mez pravostranného intervalového odhadu je (*stejná*, *menší než*, *větší než*) horní mez příslušného oboustranného odhadu.

4. Výběrová charakteristika (náhodná veličina), která nabývá hodnot „blízkých“ hledanému parametru, se nazývá

   a) bodový odhad hledaného parametru,
   b) nestranný odhad hledaného parametru,
   c) konzistentní odhad hledaného parametru,
   d) vydatný odhad hledaného parametru.

5. Interval, v němž skutečná hodnota hledaného parametru leží s pravděpodobností $1 - \alpha$ ($\alpha$ je hladina významnosti), se nazývá

   a) interval spolehlivosti, [intervalový odhad je realizací IS]
   b) intervalový odhad.

6. Hladina významnosti $\alpha$ je pravděpodobnost toho, že skutečná hodnota hledaného parametru

   a) leží uvnitř intervalu spolehlivosti,
   b) neleží uvnitř intervalu spolehlivosti.

7. Spolehlivost odhadu $1-\alpha$ je pravděpodobnost toho, že skutečná hodnota hledaného parametru

   a) leží uvnitř intervalu spolehlivosti,
   b) neleží uvnitř intervalu spolehlivosti.

## 9.2 Příklady

1. Při kontrolních zkouškách 16 žárovek byl stanoven odhad střední hodnoty $\overline{x}_{16} = 3\,000$ hodin a směrodatné odchylky $s_{16} = 20$ hodin jejich životnosti. Za předpokladu,že životnost žárovky má normální rozdělení, určete 90% intervalový odhad pro parametry $\mu$ a $\sigma$.

   *Odhadované parametry*: střední hodnota životnosti žárovek - $\mu$,
   směrodatná odchylka životnosti žárovek - $\sigma$

   *Ověření předpokladů*: normální rozdělení životnosti žárovek $\Rightarrow$ ze zadání splněno

   *Intervalový odhad parametru 1*:

   $$\overline{x} = 3\,000 \text{ hod} \qquad \text{n} = 16$$

   $$\text{s} = 20 \text{ hod} \qquad 1 - \alpha = 0,9$$

   $$P\left(\overline{x} - t_{1-\frac{\alpha}{2};n-1} \cdot \frac{s}{\sqrt{n}} < \mu < \overline{x} + t_{1-\frac{\alpha}{2};n-1} \cdot \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

   $$P(2\,991 < \mu < 3\,009) = 0,9$$

   90% intervalový odhad střední hodnoty životnosti je $(2\,991; 3\,009)$ h.

   *Intervalový odhad parametru 2*:

   $$\text{s} = 20 \qquad \text{n} = 16 \qquad 1 - \alpha = 0,9$$

   $$P\left(\sqrt{\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2};n-1}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2};n-1}}}\right) = 1 - \alpha$$

   $$P(15 < \sigma < 29) = 0,9$$

   90% intervalový odhad směrodatné odchylky životnosti je $(15; 29)$ h.

   [$(2\,991; 3\,009)$ h, $(15; 29)$ h]

2. Hloubka moře se měří přístrojem, jehož systematická chyba je rovna nule a náhodné chyby mají normální rozdělení se směrodatnou odchylkou 20 m. Kolik nezávislých měření je třeba provést, aby s pravděpodobností 95 % stanovila hloubku s chybou menší než 10 m?

   *Odhad*: rozsahu výběru

   *Ověření předpokladů*: normální rozdělení náhodné chyby $\Rightarrow$ ze zadání splněno

   *Minimální počet nez. měření nutných pro stanovení střední hloubky moře s chybou menší než 10 m:*

   $$\sigma = 20 \text{ m} \qquad \Delta = 10 \text{ m} \qquad 1 - \alpha = 0,95$$

   $$n \geq \left(\frac{\sigma}{\Delta} \cdot z_{1-\frac{\alpha}{2}}\right)^2$$

   $$n \geq 15,4$$

   Minimální počet nez. měření nutných pro stanovení střední hloubky moře s chybou menší než 10 m je 16.

   [16]

3. Úkolem je určit průměrnou hladinu cholesterolu v séru v určité populaci mužů. V náhodném výběru (pocházejícím z normálního rozdělení ) 25 mužů je výběrový průměr 6,3 mmol/l a výběrová směrodatná odchylka 1,3 mmol/l.

   *Odhadovaný parametr:* střední hodnota hladiny cholesterolu- $\mu$

   *Ověření předpokladů:* bez splnění předpokladu normality nelze příklad řešit

   *Intervalový odhad parametru:*

   $\overline{x} = 6,3$ mmol/l $\quad s = 1,3$ mmol/l $\quad 1-\alpha = 0,95 \quad P(5,8 < \mu < 6,8) = 0,95$

   95% intervalový odhad střední hladiny cholesterolu je $(5,8; 6,8)$ mmol/l.

   [$(5,8; 6,8)$ mmol/l]

4. Předpokládejme, že v náhodném výběru 200 mladých mužů má 120 z nich vyšší než doporučenou hladinu cholesterolu v séru. Určete 95% interval spolehlivosti pro procento mladých mužů s vyšší hladinou cholesterolu v populaci.

   *Odhadovaný parametr:* relativní četnost mladých mužů s vyšší hladinou cholesterolu v populaci - $\pi$

   *Ověření předpokladů:*

   předpoklad 1: $n > 30 \Rightarrow n = 200 \Rightarrow$OK
   předpoklad 2: $n/N < 0,05 \Rightarrow$ výběr menší než 5 % populace považujeme za splněný
   předpoklad 3: $n \cdot p \cdot (1-p) > 9 \Rightarrow 200 \cdot \frac{120}{200} \cdot (1 - \frac{120}{200}) = 40 > 9 \Rightarrow$OK

   *Intervalový odhad parametru (Waldův – dle CLV)*:

   $$p = \frac{120}{200} \qquad 1-\alpha = 0,95$$

   $$P\left(p - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}} < \pi < p + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha$$

   $$P(0,532 < \pi < 0,668) = 0,95$$

   95% Waldův intervalový odhad pravděpodobností, že muž v dané věkové kategorii bude mít vyšší než doporučenou hladinu cholesterolu v séru je $(53,2\ \%; 66,8\ \%)$.

   *Intervalový odhad parametru (Clopperův - Pearsonův – Statgraphics):*

   $P(0,528 < \pi < 0,668) = 0,95$

   95% Cloperův-Pearsonův intervalový odhad pravděpodobností, že muž v dané věkové kategorii bude mít vyšší než doporučenou hladinu cholesterolu v séru je $(52,8\ \%; 66,8\ \%)$.

   [Waldův odhad: $(53,2\ \%; 66,8\ \%)$, C-P odhad: $(52,8\ \%; 66,8\ \%)$]

5. V rámci výzkumné studie pracujeme s náhodným výběrem 70 žen z české populace. U každé z žen byl změřen hemoglobin s přesností 0,1 g/100 ml. Naměřené hodnoty jsou v uvedeny v souboru Hemoglobin.xls. Nalezněte 95% intervalové odhady směrodatné odchylky a střední hodnoty hemoglobinu v populaci českých žen. (Normalitu ověřte na základě exploračních grafů.)

   *Odhadované parametry*: střední hodnota hladiny hemoglobinu - $\mu$,
   směrodatná odchylka hladiny hemoglobinu - $\sigma$

   *Ověření předpokladů*: pomocí exploračních grafů $\Rightarrow$ OK

   *Intervalový odhad parametru 1*:

   $$P(11,65 < \mu < 12,32) = 0,95$$

   95% intervalový odhad střední hodnoty obsahu hemoglobinu v krvi českých žen je $(11,7; 12,3)$ g/100 ml.

   *Intervalový odhad parametru 2*:

   $$P(1,2 < \sigma < 1,7) = 0,95$$

   95% intervalový odhad směrodatné odchylky obsahu hemoglobinu v krvi českých žen je $(1,2; 1,7)$ g/100 ml.

   [$(11,7; 12,3)$ g/100 ml, $(1,2; 1,7)$ g/100 ml]

6. Jaký musí být počet pozorování, jestliže chceme s pravděpodobností 0,95 stanovit průměrnou hodnotu hemoglobinu u novorozenců s chybou nejvýše 1,0 g/l. Populační rozptyl hodnot se odhaduje hodnotou 46,0 $g^2/l^2$.

   *Odhad*: rozsahu výběru

   *Ověření předpokladů*: bez splnění předpokladu normality nelze příklad řešit

   *Minimální počet novorozenců, které musíme vyšetřit, abychom stanovili střední hodnotu hemoglobinu u novorozenců s přesností na 1,0 g/l*:

   $\sigma = \sqrt{46}$ g/l $\quad \Delta = 1,0$ g/l $\quad 1 - \alpha = 0,95 \quad n \geq 180$ (Statgraphics)

   Minimální počet novorozenců, které musíme vyšetřit, abychom stanovili střední hodnotu hemoglobinu u novorozenců s přesností na 1,0 g/l je 180.

   [180]

7. V průběhu experimentu sledujeme vliv chlazení (skupina 1 - žádné, skupina 2 - chlazení vodou) okolních struktur na největší rozměr poškození tkáně slinivky břišní slinivka.xls. Kvantifikujte efekt vlivu chlazení a určete jeho 95% intervalový odhad.

   *Odhadované parametry*: střední hodnota poškození tkáně skupina 1 - $\mu_1$,
   střední hodnota poškození tkáně skupina 2 - $\mu_2$

   *Ověření předpokladů*: normalita rozdělení poškození tkáně - skupina 1 $\Rightarrow$ OK
   normalita rozdělení poškození tkáně - skupina 2 $\Rightarrow$ OK
   shoda rozptylů zamítnuta pro $\alpha = 0,05$

   *Intervalový odhad středního poškození tkáně pro skupinu 1*:

   $$P(24,83 < \mu_1 < 25,1) = 0,95$$

   *Intervalový odhad středního poškození tkáně pro skupinu 2*:

   $$P(21,29 < \mu_2 < 22,39) = 0,95$$

   *Intervalový odhad rozdílu středních poškození tkáně pro skupinu 1 a skupinu 2*:

   $$P(2,59 < \mu_1 - \mu_2 < 3,75) = 0,95$$

   Intervalový odhad rozdílu středních poškození tkáně neobsahuje nulu, tzn. $\mu_1$ je statistick významně vyšší než $\mu_2$ $\Rightarrow$ dochází k výrazně vyššímu poškození tkáně, jestliže není prováděno chlazení okolních struktur.

   $[(24,83; 25,19),\ (21,29; 22,39),\ (2,59; 3,75)]$

# 10. Testování hypotéz - základní terminologie

## 10.1 Test z teorie

Doplňte

a) Statistická hypotéza je výrok o ..... [rozdělení náhodné veličiny]

b) Rozhodovací proces, který používáme k učinění závěru o rozdělení náhodné veličiny na základě výběrového souboru a hypotéz se nazývá ..... [testování hypotéz]

c) Při testování hypotéz se rozhodujeme mezi ..... a ..... hypotézou. [nulovou, alternativní]

d) Obor hodnot testové statistiky (testového kritéria) lze rozdělit na dvě disjunktní množiny nazývané ..... a ..... [kritický obor, obor přijetí]

e) Kritický obor se stanovuje tak, aby pravděpodobnost, že hodnota testové statistiky padne do kritického oboru byla v případě platnosti nulové hypotézy rovna ..... [pravděpodobnosti chyby 1. druhu ($\alpha$)]

f) Pravděpodobnost chyby I. druhu i chyby II. druhu lze snížit, zvýšíme-li ..... [rozsah výběru]

g) Přístup k testování hypotéz, který je založen na rozhodování pomocí kritického oboru bývá nazýván ..... [klasický test]

h) Přístup k testování hypotéz, který je založen na rozhodování pomocí p-hodnoty bývá nazýván ..... [čistý test významnosti]

i) Při testování hypotéz je možno učinit dvě rozhodnutí - ..... nebo ..... [zamítnout, nezamítnout nulovou hypotézu]

j) Je-li p-hodnota=0,03, pak ..... nulovou hypotézu se spolehlivostí 0,95. [zamítáme]

# 11. Jednovýběrové testy hypotéz o parametrech populace

| Typ proměnné | Požadovaný typ analýzy | Předpoklady | Testy, resp. intervalové odhady |
|---|---|---|---|
| Spojitá proměnná | Ověření variability | Normalita | Test o rozptylu (test o směr. odchylce) |
| | | | Intervalový odhad rozptylu (směr. odchylky) |
| | Ověření polohy | Normalita | Studentův $t$-test, (test o střední hodnotě) |
| | | | Intervalový odhad střední hodnoty |
| | | Výběr většího rozsahu | Znaménkový test (test o mediánu) |
| | | Symetrické rozdělení | Wilcoxonův test (test o mediánu) |
| Dichotomická proměnná (0-1) | Ověření shody relativní četnosti s očekávanou pravděpodobností | $n > \frac{9}{p(1-p)}$ | Test o parametru $\pi$ binomického rozdělení |
| | | | Intervalový odhad parametru $\pi$ binomického rozdělení |

## 11.1 Test z teorie

1. Zamítneme-li na základě $t$-testu nulovou hypotézu, pak lze tvrdit, že rozdíl mezi testovanou hodnotou a průměrem výběrového souboru

   a) je na dané hladině významnosti statisticky významný,
   b) není na dané hladině významnosti statisticky významný.

2. Označte všechny parametrické testy, tj. testy vyžadující znalost rozdělení populace

   a) test o střední hodnotě ($t$-test),
   b) test o rozptylu,
   c) mediánový test nebo Wilcoxonův test,
   d) test o parametru $\pi$ binomického rozdělení.

3. Neparametrické testy

   a) nevyžadují splnění žádných předpokladů,
   b) nevyžadují znalost rozdělení populace,
   b) vyžadují znalost rozdělení populace.

4. Neparametrické testy mají

   a) větší sílu testu než jejich parametrické protějšky,
   b) menší sílu testu než jejich parametrické protějšky.

5. Předpokladem pro použití testu o parametru $\pi$ binomického rozdělení je

   a) normalita výběru,
   b) výběr ze spojitého symetrického rozdělení,
   c) dostatečný rozsah výběru ($n > 9/(p(1-p))$, kde $p$ je relativní četnost výskytu sledovaného jevu.

6. Předpokladem pro použití Wilcoxonova testu je

   a) normalita výběru,
   b) výběr ze spojitého symetrického rozdělení,
   c) dostatečný rozsah výběru ($n > 9/(p(1-p))$, kde $p$ je relativní četnost výskytu sledovaného jevu.

7. Chceme-li ověřit, zda lze výrobcem udávanou spotřebu 8,8 l/100km považovat za pravdivou (bylo testováno 11 automobilů, normalita výběru byla zamítnuta), použijeme

   a) test o střední hodnotě ($t$-test),
   b) test o rozptylu,
   c) mediánový test nebo Wilcoxonův test,
   d) test o parametru $\pi$ binomického rozdělení.

8. Chceme-li ověřit, zda lze očekávat, že v prodejně je více než 5 % konzerv s prošlou záruční lhůtou (v kontrolním vzorku 100 konzerv bylo nalezeno 7 konzerv s prošlou záruční lhůtou), použijeme

   a) test o střední hodnotě ($t$-test),
   b) test o rozptylu,
   c) mediánový test nebo Wilcoxonův test,
   d) test o parametru $\pi$ binomického rozdělení.

9. Chceme-li ověřit, zda je průměrná výška dospělé populace v ČR větší než 170 cm (rozsah výběru je 120, byla ověřena normalita výběru), použijeme

   a) test o střední hodnotě ($t$-test),
   b) test o rozptylu,
   c) mediánový test nebo Wilcoxonův test,
   d) test o parametru $\pi$ binomického rozdělení.

10. Pro bavlněnou přízi je předepsaná horní mez variability pevnosti vlákna. Rozptyl pevnosti (která má normální rozdělení) nemá překročit 0,36. Chceme-li ověřit, zda je důvod k podezření na vyšší variabilitu než je stanoveno, použijeme

    a) test o střední hodnotě ($t$-test),
    b) test o rozptylu,
    c) mediánový test nebo Wilcoxonův test,
    d) test o parametru $\pi$ binomického rozdělení.

## 11.2 Příklady

1. Máme výběr 216 pacientů a změřili jsme jejich bílkovinné sérum bilik_serum.xlsx. Ověřte, zda se průměrné bílkovinné sérum (Albumin) všech pacientů tohoto typu (populační průměr $\mu$) statisticky významně liší od hodnoty 35 g/l.

   *Testovaný parametr*: střední hodnota Albuminu - $\mu$ (medián Albuminu - $x_{0,5}$)

   | | | | |
   |---|---|---|---|
   | *Možné testy*: | Studentův $t$-test | Znaménkový test | Wilcoxnův test |
   | | ↓ | ↓ | ↓ |
   | *Předpoklady testů*: | normalita rozdělení ⇒ OK | výběr většího rozsahu | výběr ze spojitého symetrického rozdělení |

   *Zvolený test*: Studentův $t$-test

   *Nulová hypotéza* $H_0$: $\mu = 35$ g/l

   *Alternativní hypotéza* $H_A$: $\mu \neq 35$ g/l

   *Pozorovaná hodnota* $x_{OBS}$: -19,25

   *p-hodnota*: $<< 0,001 < 0,5 \Rightarrow$ zamítáme nulovou hypotézu

   *Rozhodnutí*: Na hladině významnosti 0,05 zamítáme nulovou hypotézu (Studentův $t$-test, $p$-hodnota $<< 0,001$) ⇒ s pravděpodobností 95 % se průměrné bílkovinné sérum statisticky výzmamně liší od hodnoty 35 g/l. [$p$-hodnota $<< 0,001$]

2. V souboru preziti.xls jsou uvedeny doby přežití pro 100 pacientů s rakovinou plic léčených novým lékem. Z předchozích studií je známo, že průměrné přežití takových pacientů bez podávání nového léku je 22,2 měsíce. Lze na základě těchto dat usoudit, že nový lék prodlužuje přežití?

   *Testovaný parametr*: střední hodnota doby přežití - $\mu$ (medián doby přežití - $x_{0,5}$)

   | | | | |
   |---|---|---|---|
   | *Možné testy*: | Studentův $t$-test | Znaménkový test | Wilcoxnův test |
   | | ↓ | ↓ | ↓ |
   | *Předpoklady testů*: | normalita rozdělení ↓ Výrazně zešikmené rozdělení ⇒KO | výběr většího rozsahu ⇓ OK | výběr ze spojitého symetrického rozdělení ↓ Výrazně zešikmené rozdělení ⇒KO |

   *Zvolený test*: Znaménkový test

   *Nulová hypotéza* $H_0$: $x_{0,5} = 22,2$ měsíce

   *Alternativní hypotéza* $H_A$: $x_{0,5} > 22,2$ měsíce

   *Pozorovaná hodnota* $x_{OBS}$: 46 - počet pozorování v náhodném výběru, které překročí hypotetický medián

   *p-hodnota*: $0,758 > 0,5 \Rightarrow$ nezamítáme nulovou hypotézu

   *Rozhodnutí*: Na hladině významnosti 0,05 nezamítáme nulovou hypotézu (Znaménkový test, $p$-hodnota 0,758) ⇒ s pravděpodobností 95 % můžeme soudit, že není statisticky významný důkaz pro tvrzení, že nový lék prodlužuje přežití. [$p$-hodnota $= 0,758$ ]

3. Automat vyrábí pístové kroužky o daném průměru. Výrobce udává, že směrodatná odchylka průměru kroužku je 0,05 mm. K ověření této informace bylo náhodně vybráno 80 kroužků a vypočtena směrodatná odchylka jejich průměru 0,04 mm. Lze tento rozdíl považovat za statisticky významný ve smyslu zlepšení kvality produkce? Ověřte čistým testem významnosti. Předpokládejte, že průměr pístových kroužků má normální rozdělení.

   *Testovaný parametr*: směrodatná odchylka průměru kroužků - $\sigma$

   *Možné testy*: test o rozptylu (test o směrodatné odchylce)

   *Předpoklady testů*: normalita rozdělení $\Rightarrow$ ze zadání splněno

   *Zvolený test*: test o rozptylu (test o směrodatné odchylce)

   *Nulová hypotéza* $H_0$: $\sigma = 0,05$ mm

   *Alternativní hypotéza* $H_A$: $\sigma < 0,05$ mm (zlepšení$\approx$ zmenšení $\sigma$)

   *Pozorovaná hodnota* $x_{OBS}$: 50,56

   *p-hodnota*: $0,006 < 0,05 \Rightarrow$ zamítáme nulovou hypotézu

   *Rozhodnutí*:

   Na hladině významnosti 0,05 zamítáme nulovou hypotézu (test o rozptylu, $p$-hodnota = 0,006) $\Rightarrow$ s pravděpodobností 95 % lze rozdíl považovat za statisticky významný ve smyslu zvýšení kvality produkce.

   [$p$-hodnota $= 0,006$]

4. Automat vyrábí pístové kroužky o daném průměru. Výrobce udává, že směrodatná odchylka průměru kroužku je 0,05 mm. K ověření této informace bylo náhodně vybráno 80 kroužků a byl změřen jejich průměr krouzky.xls. Lze zjištěné výsledky považovat za statisticky významné ve smyslu zlepšení kvality produkce? Ověřte čistým testem významnosti.

   *Testovaný parametr*: směrodatná odchylka průměru kroužků - $\sigma$

   *Možné testy*: test o rozptylu (test o směrodatné odchylce)

   *Předpoklady testů*: normalita rozdělení $\Rightarrow$ podle exploračních grafů OK

   *Zvolený test*: test o rozptylu (test o směrodatné odchylce)

   *Nulová hypotéza* $H_0$: $\sigma = 0,05$ mm

   *Alternativní hypotéza* $H_A$: $\sigma < 0,05$ mm

   *Pozorovaná hodnota* $x_{OBS}$: 22,55

   *p-hodnota*: $<< 0,001 < 0,05 \Rightarrow$ zamítáme nulovou hypotézu

   *Rozhodnutí*:

   Na hladině významnosti 0,05 zamítáme nulovou hypotézu (test orozptylu, $p$-hodnota $<< 0,001$) $\Rightarrow$ s pravděpodobností 95 % lze rozdíl považovat za statisticky významný ve smyslu zvýšení kvality produkce.

   [$p$-hodnota $<< 0,001$]

5. Firma TT udává, že 1% jejich rezistorů nesplňuje požadovaná kritéria. V testované dodávce 1000 ks bylo nalezeno 15 nevyhovujících rezistorů. Potvrzuje tento výsledek tvrzení TT? Ověřte čistým testem významnosti.

   *Testovaný parametr*: relativní četnost televizorů nesplňujících kritéria- $\pi$

   *Možné testy*: test o parametru $\pi$ binomického rozdělení

   *Předpoklady testů*: $np(1-p) > 9 : 1\,000 \cdot \frac{15}{1\,000} \cdot (1 - \frac{15}{1\,000}) = 14,775 > 9 \Rightarrow$ OK

   *Zvolený test*: test o parametru $\pi$ binomického rozdělení

   *Nulová hypotéza* $H_0$: $\pi = 0,01$

   *Alternativní hypotéza* $H_A$: $\pi > 0,01$

   *Pozorovaná hodnota* $x_{OBS}$: —

   *p-hodnota*: $0,08 > 0,05 \Rightarrow$ nezamítáme nulovou hypotézu

   *Rozhodnutí*:

   Na hladině významnosti 0,05 nezamítáme nulovou hypotézu (test o parametru $\pi$ binomického rozdělení, $p$-hodnota = 0,08) $\Rightarrow$ s pravděpodobností 95 % potvrzuje tento výsledek tvrzení firmy TT.

   [$p$-hodnota $= 0,08$]

# 12. Dvouvýběrové testy o shodě parametrů dvou populací

<table>
<tr><th>Typ proměnné</th><th>Požadovaný typ analýzy</th><th colspan="2">Předpoklady</th><th>Testy, resp. intervalové odhady</th></tr>
<tr><td rowspan="8">Dvě nezávislé spojité proměnné</td><td rowspan="2">Ověření shody rozptylů (homoskedasticity)</td><td colspan="2" rowspan="2">Normalita</td><td>$F$-test (test shody rozptylů)</td></tr>
<tr><td>Intervalový odhad poměru rozptylů, resp. směr. odchylek</td></tr>
<tr><td colspan="2">—</td><td>Leveneho test</td></tr>
<tr><td rowspan="5">Ověření shody měr polohy (středních hodnot, resp. mediánů)</td><td rowspan="4">Normalita</td><td rowspan="2">Shoda rozptylů (homoskedasticita)</td><td>Dvouvýběrový Studentův $t$-test (test shody stř. hodnot)</td></tr>
<tr><td>Intervalový odhad rozdílu stř.hodnot</td></tr>
<tr><td rowspan="2">Různé rozptyly (heteroskedasticita)</td><td>Aspinové-Welchův test (test shody stř. hodnot)</td></tr>
<tr><td>Intervalový odhad rozdílu stř.hodnot</td></tr>
<tr><td colspan="2">—</td><td>Mannův-Whitneho test test shody mediánů</td></tr>
<tr><td rowspan="4">Párová (spojitá) data</td><td rowspan="4">Ověření shody úrovně párových dat</td><td colspan="2" rowspan="2">Normalita</td><td>Párový studentův $t$-test</td></tr>
<tr><td>Intervalový odhad střední hodnoty rozdílů</td></tr>
<tr><td colspan="2">Výběry většího rozsahu</td><td>Párový znaménkový test</td></tr>
<tr><td colspan="2">Symetrické rozdělení</td><td>Wilcoxonův párový test</td></tr>
<tr><td rowspan="2">Dvě dichotomické proměnné</td><td rowspan="2">Ověření shody pravděpodobností</td><td colspan="2" rowspan="2">$n_i > \frac{9}{p_i(1-p_i)}, i = 1, 2$</td><td>Test homogenity dvou binomických rozdělení</td></tr>
<tr><td>Intervalový odhad rozdílu parametru binomických rozdělení</td></tr>
</table>

## 12.1 Test z teorie

1. Označte všechny neparametrické (robustní) testy.

   a) dvouvýběrový $t$-test,
   b) párový $t$-test,
   c) Aspinové-Welchův test,
   d) Mannův-Whitneho test,
   e) znaménkový test nebo párový Wilcoxonův test,
   f) dvouvýběrový F-test (test o shodě rozptylů),
   g) test homogenity dvou binomických rozdělení.

2. Předpokladem pro použití Mannova-Whitneyho testu je

   a) normalita obou výběrů,
   b) normalita a homoskedasticita obou výběrů ,
   c) normalita a heteroskedasticita obou výběrů ,
   d) spojitost a stejný tvar rozdělení obou výběrů,
   e) dostatečný rozsah obou výběrů $(n_i > 9/(p_i(1-p_i)), i = 1, 2)$, kde $p_i$ je relativní četnost sledovaného jevu v $i$-tém výběru.

3. Předpokladem pro použití párového $t$-testu je

   a) normalita obou výběrů,
   b) normalita a homoskedasticita obou výběrů ,
   c) normalita a heteroskedasticita obou výběrů ,
   d) spojitost a stejný tvar rozdělení obou výběrů,
   e) dostatečný rozsah obou výběrů $(n_i > 9/(p_i(1-p_i)), i = 1, 2)$, kde $p_i$ je relativní četnost sledovaného jevu v $i$-tém výběru.

4. Předpokladem pro použití Aspinové-Welchova testu je

   a) normalita obou výběrů,
   b) normalita a homoskedasticita obou výběrů,
   c) normalita a heteroskedasticita obou výběrů,
   d) spojitost a stejný tvar rozdělení obou výběrů,
   e) dostatečný rozsah obou výběrů $(n_i > 9/(p_i(1-p_i)), i = 1, 2)$, kde $p_i$ je relativní četnost sledovaného jevu v $i$-tém výběru.

5. Neparametrickým protějškem Aspinové-Welchova testu je

   a) dvouvýběrový $t$-test,
   b) párový $t$-test,
   c) Mannův-Whitneyho test,
   d) znaménkový test nebo párový Wilcoxonův test,
   e) dvouvýběrový $F$-test (test o shodě rozptylů),
   f) test homogenity dvou binomických rozdělení.

6. Neparametrickým protějškem párového $t$-testu je

   a) dvouvýběrový $t$-test,
   b) Aspinové-Welchův test,
   c) Mannův-Whitneyho test,
   d) znaménkový test nebo párový Wilcoxonův test,
   e) dvouvýběrový $F$-test (test o shodě rozptylů),
   f) test homogenity dvou binomických rozdělení.

7. Neparametrickým protějškem dvouvýběrového $t$-testu je

   a) párový $t$-test,
   b) Aspinové-Welchův test,
   c) Mannův-Whitneyho test,
   d) znaménkový test nebo párový Wilcoxonův test,
   e) dvouvýběrový $F$-test (test o shodě rozptylů),
   f) test homogenity dvou binomických rozdělení.

8. Tabáková firma TAB prohlašuje, že jejich cigarety mají nižší obsah nikotinu než cigarety NIK. Obsah nikotinu byl změřen ve 100 cigaretách TAB a 100 cigaretách NIK. Na základě obou výběru byla ověřena homoskedasticita obsahů nikotinu v cigaretách TAB a NIK. Bylo ověřeno, že obsah nikotinu v cigaretách má normální rozdělení. Chceme-li ověřit, zda lze tvrzení firmy TAB prohlásit za nepravdivé, použijeme

   a) dvouvýběrový $t$-test,
   b) párový $t$-test,
   c) Aspinové-Welchův test,
   d) Mannův-Whitneyho test,
   e) znaménkový test nebo párový Wilcoxonův test,
   f) dvouvýběrový $F$-test (test o shodě rozptylů),
   g) test homogenity dvou binomických rozdělení.

9. Při testování ojetí (mm) pneumatik 11 automobilů určité značky byla zamítnuta normalita ojetí pneumatik (mm). Chceme-li ověřit, zda se pravé a levé přední pneumatiky automobilů této značky ojíždějí srovnatelně, použijeme

   a) dvouvýběrový $t$-test,
   b) párový $t$-test,
   c) Aspinové-Welchův test,
   d) Mannův-Whitneyho test,
   e) znaménkový test nebo párový Wilcoxonův test,
   f) dvouvýběrový $F$-test (test o shodě rozptylů),
   g) test homogenity dvou binomických rozdělení.

10. Bylo ověřeno, že hmotnost balení cukru má normální rozdělení. Testujeme-li, zda seřízením výrobní linky došlo ke snížení kolísavosti hmotnosti balení cukru, použijeme

    a) dvouvýběrový $t$-test,
    b) párový $t$-test,
    c) Aspinové-Welchův test,
    d) Mannův-Whitneyho test,
    e) znaménkový test nebo párový Wilcoxonův test,
    (f) dvouvýběrový $F$-test (test o shodě rozptylů),
    g) test homogenity dvou binomických rozdělení.

11. Určete, zda jsou následující tvrzení pravdivá.

    N a) Při neparametrickém testu homogenity dvou binomických rozdělení nemusíme ověřovat žádné předpoklady o výběrech. [požadavek na rozsah výběrů]
    N b) Mannův-Whitneyho test se používá pro ověření shody úrovně ve dvou závislých výběrech.
    N c) Každý test hypotézy $H_0 : \mu_1 = \mu_2$, tj. hypotézy o shodě dvou středních hodnot, je testem párovým.

## 12.2 Příklady

1. Data v souboru cholesterol2.xls udávají hladinu cholesterolu v krvi mužů dvou různých věkových skupin (20-30 letých a 40-50 letých). Ověřte na hladině významnosti 0,05 hypotézu, zda se hladina cholesterolu v krvi starších mužů neliší od hladiny cholesterolu v krvi mladších mužů.

*Testované parametry*: $\mu_M, \mu_S$ $(x_{0,5,M}, x_{0,5,S})$

| *Možné testy*: | Dvouvýběrový Studentův $t$-test | Aspinové-Welshův test | Mannův-Whitneyův test |
|---|---|---|---|
| | ↓ | ↓ | ↓ |
| *Předpoklady testů*: | normalita<br>homoskedasticita<br>nezávislost výběrů | normalita<br>heteroskedasticita<br>nezávislost výběrů | nezávislost výběrů |

Normalita u mladších mužů: $H_0$ : data pocházejí z normálního rozdělení
$H_A$ : data nepocházejí z normálního rozdělení
explorační grafy $\Rightarrow$ OK
Shapiro-Wilkův test:
$x_{OBS} = 0,97$
$p$-hodnota = 0,29 > 0,05
Na hladině významnosti 0,05 nezamítáme nulovou hypotézu o normalitě dat (S-W test, $p$-hodnota = 0,29).

Normalita u starších mužů: $H_0$ : data pocházejí z normálního rozdělení
$H_A$ : data nepocházejí z normálního rozdělení
explorační grafy $\Rightarrow$ OK
Shapiro-Wilkův test:
$x_{OBS} = 0,99$
$p$-hodnota = 0,92 > 0,05
Na hladině významnosti 0,05 nezamítáme nulovou hypotézu o normalitě dat (S-W test,$p$-hodnota = 0,92).

Homoskedasticita: $H_0 : \sigma_M^2 = \sigma_S^2$ ($M$: $n = 99$, $s = 0{,}115$)
$H_A : \sigma_M^2 \neq \sigma_S^2$ ($S$: $n = 85$, $s = 0{,}368$)
$x_{OBS} = 0,10$
$p$-hodnota $F$-testu = 0,001
Na hladině významnosti 0,05 zamítáme nulovou hypotézu o shodě rozptylů (F test, $p$-hodnota = 0,001).

Nezávislost výběrů: Z podstaty zadání splněno.

*Zvolený test*: Aspinové-Welchův test

*Nulová hypotéza* $H_0$: $\mu_M = \mu_S$ ($M$: $\overline{x}_M = 4{,}597$; $S$: $\overline{x}_S = 5{,}101$)

*Alternativní hypotéza* $H_A$: $\mu_M < \mu_S$

*Pozorovaná hodnota* $x_{OBS}$: -12,923

*p-hodnota*: $<< 0,001 \Rightarrow$ zamítáme nulovou hypotézu

*Rozhodnutí*: Na hladině významnosti 0,05 zamítáme nulovou hypotézu (Aspinové-Welshův test, $p$-hodnota $<< 0,001$) $\Rightarrow$ s pravděpodobností 95 % můžeme tvrdit, že je hladina cholesterolu u starších mužů statisticky významně vyšší než u mladších mužů. [$p$-hodnota $<< 0,001$]

2. Údaje v souboru deprese.xls představují délku remise ve dnech z prostého náhodného výběru ze dvou různých skupin pacientů (pacienti s endogenní depresi a pacienti s neurotickou depresí). Ověřte, zda je pozorovaný rozdíl mezi průměrnou délkou remise u těchto dvou skupin pacientů statisticky významný.

*Testované parametry*: $\mu_M, \mu_S$ $(x_{0,5,M}, x_{0,5,S})$

| *Možné testy*: | Dvouvýběrový Studentův $t$-test | Aspinové-Welshův test | Mannův-Whitneyův test |
|---|---|---|---|
| | ↓ | ↓ | ↓ |
| *Předpoklady testů*: | normalita<br>homoskedasticita<br>nezávislost výběrů | normalita<br>heteroskedasticita<br>nezávislost výběrů | nezávislost výběrů |

Normalita u end. skupiny: $H_0$ : data pocházejí z normálního rozdělení
$H_A$ : data nepocházejí z normálního rozdělení
Shapiro-Wilkův test:
$x_{OBS} = 0,68$
$p$-hodnota $<< 0,001$
Na hladině významnosti 0,05 zamítáme nulovou hypotézu o normalitě dat (S-W test, $p$-hodnota $<< 0,001$).
$\Rightarrow$ Není nutné zkoumat další předpoklady.

*Zvolený test*: Mannův-Whitneyův test

*Nulová hypotéza* $H_0$: $x_{0,5,E} = x_{0,5,N}$ ($E$: $x_{0,5,E} = 159{,}5$; $N$: $x_{0,5,N} = 512{,}5$)

*Alternativní hypotéza* $H_A$: $x_{0,5,E} < x_{0,5,N}$

*Pozorovaná hodnota* $x_{OBS}$: 81

*p-hodnota*: $0,62 > 0,05 \Rightarrow$ nezamítáme nulovou hypotézu

*Rozhodnutí*:

Na hladině významnosti 0,05 nezamítáme nulovou hypotézu (Mannův-Whitneyův test, $p$-hodnota = 0,62) $\Rightarrow$ s pravděpodobností 95 % není pozorovaný rozdíl mezi průměrnou délkou remise statisticky významný.

[$p$-hodnota $= 0,62$]

3. Sledujeme osmolalitu moči na lůžkové stanici v 08:00 hodin a v 11:00 hodin u 16 mužů. Na základě výsledků uvedených v souboru osmolalita.xls ověřte, zda se osmolalita statisticky významně zvýšila.

*Testované parametry*: $\mu_{11-8}$

| *Možné testy*: | Párový Studentův $t$-test | Párový znaménkový test | Wilcoxnův párový test |
|---|---|---|---|
| | ↓ | | ↓ |
| *Předpoklady testů*: | normalita počítat přes diference | | spojité rozdělení symetrické kolem mediánu |

Normalita diferencí (11-8): $H_0$ : data pocházejí z normálního rozdělení
$H_A$ : data nepocházejí z normálního rozdělení
Shapiro-Wilkův test:
$x_{OBS} = 0,87$
$p$-hodnota $= 0,06$
Na hladině významnosti 0,05 nezamítáme nulovou hypotézu o normalitě dat (S-W test, $p$-hodnota = 0,06).

*Zvolený test*: Párový Studentův $t$-test

*Nulová hypotéza* $H_0$: $\mu_{11-8} = 0$ $\quad \overline{x}_{11-8} = 8,5; s = 4,8; n = 16$

*Alternativní hypotéza* $H_A$: $\mu_{11-8} > 0$

*Pozorovaná hodnota* $x_{OBS}$: 7,14

*p-hodnota*: $<< 0,001 < 0,05 \Rightarrow$ zamítáme nulovou hypotézu

*Rozhodnutí*:

Na hladině významnosti 0,05 zamítáme nulovou hypotézu (párový $t$-test, $p$-hodnota $<< 0,001$) $\Rightarrow$ s pravděpodobností 95 % můžeme tvrdit, že se hodnota osmolality statisticky významně zvýšila.

[$p$-hodnota $<< 0,001$]

4. Byly testovány polovodičové součástky dvou výrobců - MM a PP. MM prohlašuje, že její výrobky mají nižší procento vadných kusů. Pro ověření tohoto tvrzení bylo z produkce MM náhodně vybráno 200 součástek, z nichž 14 bylo vadných. Podobný experiment byl proveden u firmy PP s výsledkem 10 vadných ze 100 náhodně vybraných součástek.

   a) Otestujte tvrzení firmy MM čistým testem významnosti.

   *Testované parametry*: $\pi_{MM}, \pi_{PP}$

   *Možné testy*: Test homogenity dvou binomických rozdělení

   *Předpoklady testů*: $n_{MM} \cdot p_{MM} \cdot (1 - p_{MM}) > 9 \Rightarrow 200 \cdot \frac{14}{200} \cdot (1 - \frac{14}{200}) = 13,02$
   $n_{PP} \cdot p_{PP} \cdot (1 - p_{PP}) > 9 \Rightarrow 100 \cdot \frac{10}{100} \cdot (1 - \frac{10}{100}) = 9$

   *Zvolený test*: Test homogenity dvou binomických rozdělení

   *Nulová hypotéza* $H_0$: $\pi_{MM} = \pi_{PP}$

   *Alternativní hypotéza* $H_A$: $\pi_{MM} < \pi_{PP}$

   *Pozorovaná hodnota* $x_{OBS}$: $-0,90$

   *p-hodnota*: $0,18 > 0,05 \Rightarrow$ nezamítáme nulovou hypotézu

   *Rozhodnutí*:

   Na hladině významnosti 0,05 nezamítáme nulovou hypotézu (test homogenity dvou binomických rozdělení, $p$-hodnota = 0,18)$\Rightarrow$ s pravděpodobností 95 % nemůžeme tvrdit, že výrobky firmy MM mají nižší procento vadných kusů. [$p$-hodnota $= 0,18$]

   b) Otestujte tvrzení firmy MM prostřednictvím intervalového odhadu na hladině významnosti 0,05.

   *Testované parametry*: $\pi_{MM}, \pi_{PP}$

   *Předpoklady pro použití intervalového odhadu*:

   $n_{MM} \cdot p_{MM} \cdot (1 - p_{MM}) > 9 \Rightarrow 200 \cdot \frac{14}{200} \cdot (1 - \frac{14}{200}) = 13,02$
   $n_{PP} \cdot p_{PP} \cdot (1 - p_{PP}) > 9 \Rightarrow 100 \cdot \frac{10}{100} \cdot (1 - \frac{10}{100}) = 9$

   *Nulová hypotéza* $H_0$: $\pi_{MM} = \pi_{PP}$

   *Alternativní hypotéza* $H_A$: $\pi_{MM} < \pi_{PP}$

   *Rozhodnutí*:

   $P(-0,095 < \pi_{MM} - \pi_{PP} < 0,035) = 0,95$, protože intervalový odhad obsahuje nulu, nelze s 95% pravděpodobnostní tvrdit, že výrobky firmy MM mají nižší procento vadných kusů.

   $[P(-0,095 < \pi_{MM} - \pi_{PP} < 0,035) = 0,95]$

# 13. Vybrané vícevýběrové testy parametrických hypotéz

<table>
<tr><th>Typ proměnné</th><th>Požadovaný typ analýzy</th><th colspan="2">Předpoklady</th><th>Testy</th></tr>
<tr><td rowspan="5">Alespoň tři nezávislé spojité proměnné</td><td rowspan="3">Ověření shody rozptylů (homoskedasticity)</td><td rowspan="2">Normalita</td><td rowspan="2">Vyvážené třídění</td><td>Cochranův test</td></tr>
<tr><td>Hartleyův test</td></tr>
<tr><td>Nevyvážené třídění</td><td>Bartlettův test</td></tr>
<tr><td colspan="2">—</td><td>Leveneův test</td></tr>
<tr><td rowspan="2">Ověření shody měr polohy (středních hodnot, resp. mediánů)</td><td>Normalita</td><td>Shoda rozptylů (homoskedasticita)</td><td>ANOVA (Analýza rozptylu = test shody stř. hodnot (Poznámka: V případě zamítnutí $H_0$ je vhodné provést post hoc analýzu, např. Schéffeho metodou.)</td></tr>
<tr><td colspan="2">Symetrické rozdělení</td><td>Kruskalův-Wallisův test (test shody mediánů, resp. shody distribucí; použití i pro pořadová data. Poznámka: V případě zamítnutí $H_0$ je vhodné provést post hoc analýzu, např. Dunnové metodou.)</td></tr>
<tr><td>Alespoň tři závislé (spojité) proměnné</td><td>Ověření shody úrovně závislých dat</td><td colspan="2">—</td><td>Friedmanův test (Poznámka: V případě zamítnutí $H_0$ je vhodné provést post hoc analýzu Friedmanovou metodou.)</td></tr>
</table>

**Tabulka ANOVA**

| Zdroj variability | Součet čtverců | Počet stupňů volnosti | Rozptyl (prům. součet čtverců) | *F-poměr* | *p-hodnota* |
|---|---|---|---|---|---|
| Skupinový faktor | $SS_B = \sum_{i=1}^{k} n_i(\overline{X}_i - \overline{X})^2$ | $df_B = k - 1$ | $MS_B = \frac{SS_B}{df_B}$ | $\frac{MS_B}{MS_e}$ | $1 - F_0(x_{OBS})$ |
| Reziduální faktor | $SS_e = \sum_{i=1}^{k} (n_i - 1)s_i^2$ | $df_e = n - k$ | $MS_e = \frac{SS_e}{df_e}$ | — | — |
| Celkový faktor | $SS_T = \sum_{i=1}^{k} (X_{ij} - \overline{X})^2$ | $df_T = n - k1$ | — | — | — |

## 13.1 Test z teorie

1. Určete, zda jsou následující tvrzení pravdivá.

N a) Analýza rozptylu (ANOVA) je test shody rozptylů ve více než dvou výběrech.

P b) Jedním z předpokladů analýzy rozptylu je alespoň přibližná shoda rozptylů v jednotlivých skupinách.

N c) Reziduální rozptyl (v analýze rozptylu) lze určit jako aritmetický průměr rozptylů v jednotlivých skupinách.

N d) Post hoc analýza znamená, že stanovíme nejprve hypotézy $H_0$, $H_A$, a „následně“ provedeme řešení.

P e) Kruskalův-Wallisův test se nazývá rovněž neparametrická ANOVA.

P f) Hartleyův test homoskedasticity lze použít pouze v případě vyváženého třídění.

P g) Jediným předpokladem Leveneova testu je nezávislost výběrů.

N h) Bartlettův test je neparametrickým protějškem Leveneova testu.

N i) Friedmanův test je neparametrickou obdobou Kruskalova-Wallisova testu.

## 13.2 Příklady

1. Testujeme nulovou hypotézu $\mu_1 = \mu_2 = \mu_3$. Bylo zjištěno, že data, která máme k dispozici jsou výběry z normálního rozdělení splňující předpoklad homoskedasticity (shody rozptylů). Na základě údajů získaných explorační analýzou doplňte tabulku ANOVA a vyplývající závěry.

| **Faktor** | **Rozsah výběru** | **Průměr** | **Výběrová směrodatná odchylka** |
|---|---|---|---|
| **Skupina 1** | 40 | 300 | 33 |
| **Skupina 2** | 40 | 290 | 34 |
| **Skupina 3** | 42 | 310 | 31 |
| **Celkem** | **122** | | — |

| Zdroj variability | Součet čtverců | Počet stupňů volnosti | Rozptyl (prům. součet čtverců) | *F-poměr* | *p-hodnota* |
|---|---|---|---|---|---|
| Skupinový | | | | | |
| Reziduální | | | | | |
| Celkový | | | | | |

Řešení: $\overline{x} = (40 \cdot 30 + 40 \cdot 290 + 42 \cdot 310)/(40 + 40 + 42) = 300,16 \approx 300$

$$SS_B = \sum_{i=1}^{k} n_i(\overline{X}_i - \overline{X})^2 = 40 \cdot (300 - 300)^2 + 40 \cdot (290 - 300)^2 + 42 \cdot (310 - 300)^2 = 8\,200$$

$$SS_e = \sum_{i=1}^{k} (n_i - 1)s_i^2 = (40 - 1) \cdot 33^2 + (40 - 1) \cdot 34^2 + (42 - 1) \cdot 31^2 = 126\,956$$

$$MS_B = \frac{SS_B}{df_B} = \frac{8\,200}{3-1} = 4\,100 \qquad MS_e = \frac{SS_e}{df_e} = \frac{126\,956}{122-3} \cong 1\,066$$

$$SS_T = SS_B + SS_e = 4\,100 + 126\,956 = 131\,056$$

$$x_{OBS} = \frac{MS_B}{MS_e} = \frac{4\,100}{1\,066} = 3,84$$

| Zdroj variability | Součet čtverců | Počet stupňů volnosti | Rozptyl (prům. součet čtverců) | *F-poměr* | *p-hodnota* |
|---|---|---|---|---|---|
| Skupinový | 8 200 | 2 | 4 100 | 3,84 | |
| Reziduální | 126 956 | 119 | 1 066 | — | — |
| Celkový | 131 056 | 121 | — | — | — |

*Nulová hypotéza* $H_0$: $\mu_1 = \mu_2 = \mu_3$

*Alternativní hypotéza* $H_A$: $\neg H_0$

*p-hodnota*: $1 - F_0(x_{OBS}) = 1 - F_{2,119}(3,84) = 0,02 < 0,05$

*Rozhodnutí*: Na hladině významnosti 0,05 zamítáme nulovou hypotézu (ANOVA, *p*-hodnota = 0,02) $\Rightarrow$ s pravděpodobností 95 % nemůžeme tvrdit, že střední hodnoty výběrů nejsou stejné.

*Odhady skupinových efektů (tj. jak se průměry jednotlivých skupin liší od celkového průměru)*:

1: $300 - 300 = 0$
2: $290 - 300 = -10$
3: $310 - 300 = 10$ [*p*-hodnota = 0,02]

2. 122 pacientů, kteří podstoupili operaci srdce, bylo náhodně rozděleno do tří skupin.
**Skupina 1:**
Pacienti dostali 50 % oxidu dusného a 50 % kyslíkové směsi nepřetržitě po dobu 24 hodin.
**Skupina 2:**
Pacienti dostali 50 % oxidu dusného a 50 % kyslíkové směsi pouze během operace.
**Skupina 3:**
Pacienti nedostali žádný oxid dusný, ale dostali 35-50 % kyslíku po dobu 24 hodin.

Data v souboru kyselina listova.xls odpovídají koncentracím soli kyseliny listové v červených krvinkách ve všech třech skupinách po uplynutí 24 hodin ventilace. Ověřte, zda pozorované rozdíly mezi koncentracemi soli kyseliny listové jsou statisticky významné, tj. zda existuje vliv složení směsi na sledovaný parametr.

| | | |
|---|---|---|
| *Možné testy*: | ANOVA<br>↓ | Kruskalův-Wallisův<br>test<br>↓ |
| *Předpoklady testů*: | normalita<br>homoskedasticita<br>nezávislost výběrů | symetrické<br>rozdělení<br>nezávislost výběrů |

*Vizuální posouzení*: Průměry jednotlivých výběrů se jeví jako odlišné, rozptyly podobné, data jsou v každém výběru rozložena symetricky.

*Exaktní posouzení (testování hypotéz)*:

Normalita: $H_0$ : data pochází z normálního rozdělení
$H_A : \neg H_0$

| Shapiro-Wilkův test | $p$-hodnota |
|---|---|
| Skupina 1 | $0{,}97 > 0{,}05 \Rightarrow$ OK |
| Skupina 2 | $0{,}63 > 0{,}05 \Rightarrow$ OK |
| Skupina 3 | $0{,}45 > 0{,}05 \Rightarrow$ OK |

Homoskedasticita: $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$
$H_A : \neg H_0$
Cochranův test: $p$-hodnota $= 0,50 > 0,05 \Rightarrow$ OK

Nezávislost výběrů: Splněna z podstaty zadání.

*Zvolený test*: ANOVA
*Nulová hypotéza* $H_0$: $\mu_1 = \mu_2 = \mu_3$
*Alternativní hypotéza* $H_A$: $\neg H_0$
*p-hodnota*: $<< 0,001 < 0,05 \Rightarrow$ zamítáme nulovou hypotézu

*Rozhodnutí*:Na hladině významnosti 0,05 zamítáme nulovou hypotézu (ANOVA, $p$-hodnota $<< 0,001$) $\Rightarrow$ s pravděpodobností 95 % můžeme tvrdit, že existuje vliv složení směsi na koncentraci kyseliny listové $\Rightarrow$ je vhodné provést post hoc analýzu.

*Odhady skupinových efektů (tj. jak se průměry jednotlivých skupin liší od celkového průměru)*:
Skupina 1: $304,1 - 277,4 = +26,7$
Skupina 2: $250,9 - 277,4 = -26,5$
Skupina 3: $277,2 - 277,4 = -0,2$

*Post-hoc analýza (vícenásobné porovnávání)*:
Všechny tři skupiny jsou navzájem statisticky významně odlišné. [$p$-hodnota $<< 0,001$]

3. Na farmě jsou chována tři plemena králíků. Byl proveden pokus kralici.xls, jehož cílem bylo zjistit, zda i když chováme a vykrmujeme všechny králíky po stejnou dobu a za stejných podmínek, existuje statisticky významný (průkazný) rozdíl mezi plemeny v hmotnostech králíků. Ověřte.

| | | |
|---|---|---|
| *Možné testy*: | ANOVA<br>↓ | Kruskalův-Wallisův<br>test<br>↓ |
| *Předpoklady testů*: | normalita<br>homoskedasticita<br>nezávislost výběrů | symetrické<br>rozdělení<br>nezávislost výběrů |

*Vizuální posouzení*: V datech se nechází jedno odlehlé pozorování, při bližším posouzení dojdeme k závěru, že jde o překlep (není možné, aby králík vážil 38 kg) ⇒ hodnotu opravíme vložením desetinné čárky na 3,81 (nebo odlehlé pozorování odstraníme) a provedeme o tomto úkonu zápis. Poté se průměry jednotlivých výběrů jeví jako odlišné, rozptyly podobné, data jsou v každém výběru rozložena symetricky.

*Exaktní posouzení (testování hypotéz)*:

Normalita: $H_0$ : data pochází z normálního rozdělení
$H_A : \neg H_0$

| Shapiro-Wilkův test | $p$-hodnota |
|---|---|
| český | $0{,}91 > 0{,}05 \Rightarrow$ OK |
| kalifornský | $0{,}17 > 0{,}05 \Rightarrow$ OK |
| vídeňský 3 | $0{,}81 > 0{,}05 \Rightarrow$ OK |

Homoskedasticita: $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$
$H_A : \neg H_0$
Bartlettův test: $p$-hodnota $= 0,25 > 0,05 \Rightarrow$ OK

Nezávislost výběrů: Splněna z podstaty zadání.

*Zvolený test*: ANOVA

*Nulová hypotéza* $H_0$: $\mu_1 = \mu_2 = \mu_3$

*Alternativní hypotéza* $H_A$: $\neg H_0$

*p-hodnota*: $<< 0,001 < 0,05 \Rightarrow$ zamítáme nulovou hypotézu

*Rozhodnutí*:
Na hladině významnosti 0,05 zamítáme nulovou hypotézu (ANOVA, $p$-hodnota $<< 0,001$) ⇒ s pravděpodobností 95 % můžeme tvrdit, že mezi hmotnostmi jednotlivých plemen existuje statisticky významný rozdíl ⇒ je vhodné provést post hoc analýzu.

*Odhady skupinových efektů (tj. jak se průměry jednotlivých skupin liší od celkového průměru)*:

český: $3,49 - 4,11 = -0,62$ kg
kalifornský: $4,99 - 4,11 = 0,88$ kg
vídeňský: $4,04 - 4,11 = -0,07$ kg

*Post-hoc analýza (vícenásobné porovnávání)*:
Všechny tři skupiny jsou navzájem statisticky významně odlišné. [$p$-hodnota $<< 0,001$]

4. Soutěž o nejlepší jakost výrobků obeslali čtyři výrobci A, B, C, D celkem 66 výrobky. Porota sestavila pořadí (uvedeno pouze pořadí výrobku od nejlepšího k nejhoršímu), jež je uvedené v souboru jakost.xls. Na základě uvedených údajů posuďte, zda původ výrobků má vliv na jeho jakost.

   *Možné testy*: Výsledky poroty jsou zadány jako pořadí ⇒ nejedná se o spojitá data ⇒ Kruskalův-Wallisův test

   *Předpoklady testů*:

   —

   *Vizuální posouzení*: Průměry jednotlivých výběrů se nejeví jako odlišné.

   *Exaktní posouzení (testování hypotéz)*:

   —

   *Zvolený test*: Kruskalův-Wallisův test

   *Nulová hypotéza* $H_0$: $x_{0,5;A} = x_{0,5;B} = x_{0,5;C} = x_{0,5;D}$

   *Alternativní hypotéza* $H_A$: $\neg H_0$

   *p-hodnota*: $0,30 > 0,05 \Rightarrow$ nezamítáme nulovou hypotézu

   *Rozhodnutí*:
   Na hladině významnosti 0,05 nezamítáme nulovou hypotézu (Kruskallův-Wallisův test, $p$-hodnota = 0,3) ⇒ s pravděpodobností 95 % můžeme tvrdit, že původ výrobku nemá vliv na jeho jakost. ⇒ nemá smysl provádět post hoc analýzu.

   *Odhady skupinových efektů (tj. jak se mediány jednotlivých skupin liší od celkového mediánu)*:

   A: $x_{0,5;A} - x_{0,5} = 40 - 33,5 = +6,5$
   B: $x_{0,5;B} - x_{0,5} = 24,5 - 33,5 = -9,0$
   C: $x_{0,5;C} - x_{0,5} = 35 - 33,5 = +1,5$
   D: $x_{0,5;D} - x_{0,5} = 35,5 - 33,5 = +2,0$

   POZOR! Pozorované rozdíly byly vyhodnoceny jako statisticky nevýznamné!

   *Post-hoc analýza (vícenásobné porovnávání)*:
   Nemá smysl provádět post hoc analýzu.

   [$p$-hodnota $= 0,30$]

5. Byl sledován vliv tří preparátů na srážlivost krve. Kromě jiných ukazatelů byl zjišťován tzv. trombinový čas. Údaje o 42 sledovaných osobách jsou zaznamenány v souboru trombin.xls. Závisí velikost trombinového času na tom, jaký byl použit preparát?

| | | |
|---|---|---|
| *Možné testy*: | ANOVA<br>↓ | Kruskalův-Wallisův<br>test<br>↓ |
| *Předpoklady testů*: | normalita<br>homoskedasticita<br>nezávislost výběrů | symetrické<br>rozdělení<br>nezávislost výběrů |

*Vizuální posouzení*: Průměry jednotlivých výběrů se jeví značně odlišně, rozptyly také, data u preparátu A a C možná nesplňují předpoklad normality

*Exaktní posouzení (testování hypotéz)*:

Normalita: $H_0$ : data pochází z normálního rozdělení
$H_A : \neg H_0$
Preparát A: $p$-hodnota S-W testu $= 0,03 < 0,05 \Rightarrow$ zamítáme nulovou hypotézu, data s 95% pravděpodobností nepocházejí z normálního rozdělení

*Zvolený test*: Kruskalův-Wallisův test

*Nulová hypotéza* $H_0$: $x_{0,5;A} = x_{0,5;B} = x_{0,5;C}$

*Alternativní hypotéza* $H_A$: $\neg H_0$

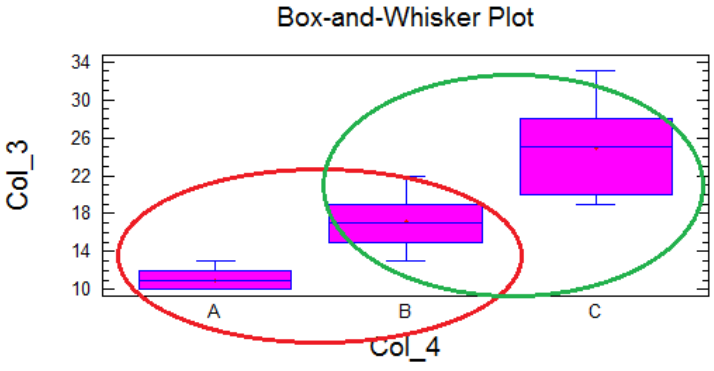
*p-hodnota*: $<< 0,001 < 0,05 \Rightarrow$ zamítáme nulovou hypotézu

*Rozhodnutí*:
Na hladině významnosti 0,05 zamítáme nulovou hypotézu (Kruskalův-Wallisův test, $p$-hodnota $<< 0,001$) $\Rightarrow$ s pravděpodobností 95 % můžeme tvrdit, že velikost trombinového času závisí na použitém preparátu $\Rightarrow$ je vhodné provést post hoc analýzu.

*Odhady skupinových efektů (tj. jak se mediány jednotlivých skupin liší od celkového mediánu)*:

Preparát A: $11,0 - 17,7 = -6,7$
Preparát B: $17,1 - 17,7 = -0,6$
Preparát C: $24,9 - 17,7 = 7,2$

*Post-hoc analýza (vícenásobné porovnávání)*:
Statisticky významný rozdíl byl indikován mezi preparátem A a preparátem C.




[$p$-hodnota $<< 0,001$]

6. Byl sledován vliv tří preparátů na srážlivost krve. Kromě jiných ukazatelů byl zjišťován tzv. trombinový čas. U každé osoby byl stanoven nejprve kontrolní údaj (K), který udává trombinový čas před zahájením pokusu. Pak byly aplikovány preparáty A, B, C, a to každý dostatečně dlouho po odeznění účinku těch předchozích. Údaje o 15 sledovaných osobách jsou uvedeny v souboru trombin2.xls. Závisí velikost trombinového času na tom, jaký byl použit preparát?

   *Možné testy*: Jedná se o závislá data (1 pacient a 4 měření) ⇒ Friedmanův test

   *Předpoklady testů*:

   —

   *Vizuální posouzení*: Průměry jednotlivých výběrů se jeví jako značně odlišné.

   *Exaktní posouzení (testování hypotéz)*:

   —

   *Zvolený test*: Friedmanův test

   *Nulová hypotéza* $H_0$: $x_{0,5;K} = x_{0,5;A} = x_{0,5;B} = x_{0,5;C}$

   *Alternativní hypotéza* $H_A$: $\neg H_0$

   *p-hodnota*: $<< 0,001 < 0,05 \Rightarrow$ zamítáme nulovou hypotézu

   *Rozhodnutí*:
   Na hladině významnosti 0,05 nezamítáme nulovou hypotézu (Friedmanův test, *p*-hodnota $<< 0,001$) ⇒ s pravděpodobností 95 % můžeme tvrdit, že velikost trombinového času závisí na použitém preparátu ⇒ je vhodné provést post hoc analýzu.

   *Odhady skupinových efektů (tj. jak se průměry jednotlivých skupin liší od celkového průměru)*:

   Kontrolní skupina: $10,8 - 16,0 = -5,2$
   Preparát A: $11,0 - 16,0 = -5$
   Preparát B: $16,9 - 16,0 = 0,9$
   Preparát C: $25,4 - 16,0 = 9,4$

   *Post-hoc analýza (vícenásobné porovnávání)*:
   Statisticky významný rozdíl byl indikován mezi kontrolní skupinou a preparátem B, kontrolní skupinou a preparátem C, preparátem A a preparátem B a nakonec mezi preparátem A a preparátem C.

   homogenní skupiny:

| | | |
|---|---|---|
| Kontrolní skupina | X | |
| Preparát A | X | |
| Preparát B | | X |
| Preparát C | | X |

   [*p*-hodnota $<< 0,001$]

# 14. Vybrané testy neparametrických hypotéz

## Testy dobré shody

| Název testu | Předpoklady testu | Testová statistika |
|---|---|---|
| $\chi^2$ test dobré shody | Očekávané četnosti $\geq 2$, alespoň 80% očekávaných četností $> 5$ | $G = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ |

## Analýza závislosti v kontingenční tabulce

| Název testu | Předpoklady testu | Testová statistika |
|---|---|---|
| Analýza závislosti v kontingenční tabulce | Očekávané četnosti $\geq 2$, alespoň 80% očekávaných četností $> 5$ | $K = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ |

- Koeficient korelace $CC = \sqrt{\frac{K}{K+n}}$ (pro čtvercové kontingenční tabulky)
- Korigovaný koeficient kontingence $CC_{cor} = \frac{CC}{CC_{max}}$, kde $CC_{max} = \sqrt{\frac{min(r;s)-1}{min(r;s)}}$ (pro obdélníkové kontingenční tabulky)
- Cramerův koeficient $V = \sqrt{\frac{K}{n(min(r;s)-1)}}$.

Tyto koeficienty se mohou vyskytovat v intervalu (0;1). Čím blíže jsou 1, tím je závislost mezi $X$ a $Y$ těsnější.

## Analýza závislosti v asociační tabulce

- Odhad poměru šancí: $\widehat{OR} = \frac{ad}{bc}$
- Intervalový odhad $OR$: $\left\langle \widehat{OR} \cdot e^{-\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \cdot z_{1-\frac{\alpha}{2}}} ; \widehat{OR} \cdot e^{\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \cdot z_{1-\frac{\alpha}{2}}} \right\rangle$
- Odhad relativního rizika: $\widehat{RR} = \frac{a(c+d)}{c(a+b)}$
- Intervalový odhad $RR$: $\left\langle \widehat{RR} \cdot e^{-\sqrt{\frac{b}{a(a+b)}+\frac{d}{c(c+d)}} \cdot z_{1-\frac{\alpha}{2}}} ; \widehat{RR} \cdot e^{\sqrt{\frac{b}{a(a+b)}+\frac{d}{c(c+d)}} \cdot z_{1-\frac{\alpha}{2}}} \right\rangle$

## 14.1 Test z teorie

1. Lze Kolmogorovův-Smirnovův test použít pro testování normality? (*ANO*, *NE* ) [pouze pro úplně specifikované testy]

2. Použijeme-li $\chi^2$ test dobré shody pro ověření toho, zda je klasická šestistěnná hrací kostka „férová“, pak má v případě platnosti nulové hypotézy testová statistika $\chi^2$ rozdělení se

   a) 4 stupni volnosti,
   b) 5 stupni volnosti,
   c) 6 stupni volnosti.

3. Jak postupujeme v případě, kdy při použití $\chi^2$ testu dobré shody vyšlo po rozdělení dat do dvaceti tříd 7 očekávaných třídních četností rovných 1?

   a) V tomto případě nelze $\chi^2$ testem dobré shody rozhodnout.
   b) Prohlásíme předpoklady testu za splněné.
   c) Sloučíme příslušné sousední třídy.

4. Empirická distribuční funkce je funkce

   a) diskrétní,
   b) spojitá,
   c) zleva spojitá,
   d) zprava spojitá.

5. Čím je mozaikový graf členitější, tím je pozorovaná závislost mezi veličinami v kontingenční tabulce

   a) slabší,
   b) silnější.

6. Analyzujeme-li závislost v kontingenční tabulce, která má 4 řádky a 5 sloupců, pak $\chi^2$ test nezávislosti můžeme použít, pokud alespoň

   a) 4,
   b) 10,
   c) 16, [tj. 80 % ze 45]
   d) 20,

   očekávaných četností je větších než 5 a ostatní jsou rovny alespoň

   a) 0,
   b) 1,
   c) 2.

7. Koeficient kontingence ($K$)

   a) nabývá hodnot z intervalu $(0; 1)$,
   b) nabývá hodnot z intervalu $\langle 0; 1\rangle$,
   c) může nabývat hodnot větších než 1.

8. (*Kontingenční*, *Asociační*) tabulka je speciálním případem (*kontingenční*, *asociační*) tabulky.

9. Je-li odhad relativního rizika $\widehat{RR} = 1{,}2$, pak

   a) mezi znaky v asociační tabulce existuje závislost,
   b) mezi znaky v asociační tabulce neexistuje závislost,
   c) o závislosti znaků v asociační tabulce musí rozhodnout test.

10. Je-li odhad poměru šancí $\widehat{OR} = 10{,}2$, pak je u exponované populace

    a) nižší šance výskytu nemoci než u neexponované populace,
    b) vyšší šance výskytu nemoci než u neexponované populace,
    c) stejná šance výskytu nemoci jako u neexponované populace.

11. Odhadujeme-li se spolehlivostí 0,95, že relativní riziko $RR \in (0,87; 1,45)$, pak na hladině významnosti 0,05 (tj. se spolehlivostí 0,95)

    a) nezamítáme hypotézu o nezávislosti znaků $X$ a $Y$,
    b) zamítáme hypotézu o nezávislosti znaků $X$ a $Y$,
    c) nelze o nezávislosti znaků $X$ a $Y$ rozhodnout.

## 14.2 Příklady

1. Hodilo se 6 000 krát hrací kostkou a zaznamenaly se počty padlých ok.

| $x_i$ (číslo které padlo) | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $n_i$ (četnost jeho výskytu) | 979 | 1 002 | 1 015 | 980 | 1 040 | 984 |

Je možné na základě příslušného testu na hladině významnosti 0,05 spolehlivě tvrdit, že kostka je „falešná“, tj. že pravděpodobnosti všech čísel na kostce nejsou stejné?

*Nulová hypotéza* $H_0$: Výběr pochází z populace, kde jsou relativní četnosti jednotlivých variant rovny číslům $\left\{\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right\}$.

*Alternativní hypotéza* $H_A$: $\neg H_0$

*Zvolený test*: $\chi^2$ test dobré shody

*Předpoklady testu*: Očekávané četnosti $\geq 2$, alespoň 80% očekávaných četností $> 5$ $\Rightarrow$ OK

| $x_i$ (číslo které padlo) | 1 | 2 | 3 | 4 | 5 | 6 | Celkem |
|---|---|---|---|---|---|---|---|
| Pozorované četnosti | 979 | 1 002 | 1 015 | 980 | 1 040 | 984 | 6 000 |
| Očekávané pravděpodobnosti | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | 1 |
| Očekávané četnosti | 1 000 | 1 000 | 1 000 | 1 000 | 1 000 | 1 000 | — |

*Počet stupňů volnosti*: $k - 1 = 6 - 1 = 5$

$$x_{OBS}: \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \frac{(979 - 1\,000)^2}{1\,000} + \frac{(1\,002 - 1\,000)^2}{1\,000} + \cdots + \frac{(979 - 1\,000)^2}{1\,000} = 2{,}93$$

*p-hodnota*: $1 - F_0(2{,}93) = 0{,}71 > 0{,}05 \Rightarrow$ nezamítáme nulovou hypotézu

*Rozhodnutí*:
Na hladině významnosti 0,05 nezamítáme nulovou hypotézu ($\chi^2$ test dobré shody, $p$-hodnota = 0,71) $\Rightarrow$ s pravděpodobností 95 % není možné tvrdit, že je kostka falešná.

[$p$-hodnota $= 0{,}71$]

2. Výrobní firma odhaduje počet poruch určitého zařízení během 100 hodin pomocí Poissonova rozdělení s parametrem 1,2. Zaměstnanci zaznamenali pro kontrolu skutečné počty poruch celkem ve 150 100hodinových intervalech (výsledky jsou uvedeny v tabulce). Ověřte čistým testem významnosti, zda má počet poruch daného zařízení během 100 hodin skutečně Poissonovo rozdělení s parametrem $\lambda t = 1,2$.

| $x_i$ (počet poruch během 100 hodin provozu) | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $n_i$ (počet pozorování) | 52 | 48 | 36 | 10 | 4 |

*Nulová hypotéza* $H_0$: Naměřené hodnoty jsou výběrem z Poissonova rozdělení s parametrem $\lambda t = 1,2$.

*Alternativní hypotéza* $H_A$: $\neg H_0$

*Zvolený test*: $\chi^2$ test dobré shody

*Předpoklady testu*: Očekávané četnosti $\geq 2$, alespoň 80% očekávaných četností $> 5$ $\Rightarrow$ OK

| $x_i$ (počet poruch během 100 hodin provozu) | 0 | 1 | 2 | 3 | 4 | Celkem |
|---|---|---|---|---|---|---|
| $n_i$ (počet pozorování) | 52 | 48 | 36 | 10 | 4 | 150 |
| Očekávané pravděpodobnosti | 0,30 | 0,36 | 0,22 | 0,09 | 0,03 | 1 |
| Očekávané četnosti | 45,18 | 54,21 | 32,53 | 13,01 | 3,90 | — |

*Počet stupňů volnosti*: $k - 1 = 5 - 1 = 4$

$x_{OBS}$: $\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \frac{(52 - 45,18)^2}{45,18} + \frac{(48 - 54,21)^2}{54,21} + \cdots + \frac{(4 - 3,9)^2}{3,9} = 2,81$

*p-hodnota*: $1 - F_0(2,81) = 0,59 > 0,05 \Rightarrow$ nezamítáme nulovou hypotézu

*Rozhodnutí*:
Na hladině významnosti 0,05 nezamítáme nulovou hypotézu ($\chi^2$ test dobré shody, $p$-hodnota = 0,59) $\Rightarrow$ s pravděpodobností 95 % můžeme tvrdit, že počet poruch za 100 hodin má Poissonovo rozdělení pravděpodobnosti s parametrem $\lambda t = 1,2$.

[$p$-hodnota $= 0,59$]

3. Zaměstnanci zaznamenali pro kontrolu počty poruch celkem ve 150 100hodinových intervalech (výsledky jsou uvedeny v tabulce). Ověřte čistým testem významnosti, zda má počet poruch daného zařízení během 100 hodin skutečně Poissonovo rozdělení.

| $x_i$ (počet poruch během 100 hodin provozu) | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $n_i$ (počet pozorování) | 52 | 48 | 36 | 10 | 4 |

*Nulová hypotéza* $H_0$: Naměřené hodnoty jsou výběrem z Poissonova rozdělení.

*Alternativní hypotéza* $H_A$: $\neg H_0$

*Odhad parametru Poissonova rozdělení*: $\hat{\lambda t} = \frac{\sum_{i=1}^{n} x_i n_i}{n} = \frac{1}{150}(50 \cdot 0 + 48 \cdot 1 + 36 \cdot 2 + 10 \cdot 3 + 4 \cdot 4) = 1,1$

*Zvolený test*: $\chi^2$ test dobré shody

*Předpoklady testu*: Očekávané četnosti $\geq 2$, alespoň 80% očekávaných četností $> 5$ $\Rightarrow$ OK

| $x_i$ (počet poruch během 100 hodin provozu) | 0 | 1 | 2 | 3 | 4 | Celkem |
|---|---|---|---|---|---|---|
| $n_i$ (počet pozorování) | 52 | 48 | 36 | 10 | 4 | 150 |
| Očekávané pravděpodobnosti | 0,33 | 0,37 | 0,20 | 0,07 | 0,03 | 1 |
| Očekávané četnosti | 49,93 | 54,92 | 30,21 | 11,08 | 3,05 | — |

*Počet stupňů volnosti*: $k - 1 - h = 5 - 1 - 1 = 3$

$x_{OBS}$: $\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \frac{(52 - 49,93)^2}{49,93} + \frac{(48 - 54,92)^2}{54,92} + \cdots + \frac{(4 - 3,05)^2}{3,05} = 2,47$

*p-hodnota*: $1 - F_0(2,47) = 0,65 > 0,05 \Rightarrow$ nezamítáme nulovou hypotézu

*Rozhodnutí*:
Na hladině významnosti 0,05 nezamítáme nulovou hypotézu ($\chi^2$ test dobré shody, $p$-hodnota = 0,65) $\Rightarrow$ s pravděpodobností 95 % můžeme tvrdit, že počet poruch za 100 hodin má Poissonovo rozdělení pravděpodobnosti.

[$p$-hodnota $= 0,65$]

4. Na dálnici byly v průběhu několika minut měřeny časové odstupy (s) mezi průjezdy jednotlivých vozidel. Zjištěné hodnoty těchto odstupů jsou zaznamenány v souboru dalnice.xlsx.

   (viz prezentace přednášky)

   *Nulová hypotéza* $H_0$: Časové odstupy lze považovat za veličinu s normálním rozdělením.

   *Alternativní hypotéza* $H_A$: $\neg H_0$

   *Odhad parametrů normálního rozdělení*: $\hat{\mu} = \overline{x} = 4,6$, $\hat{\sigma} = s = 3,3$

   *Zvolený test*: $\chi^2$ test dobré shody

   *Počet stupňů volnosti*: $k - 1 - h = 12 - 1 - 2 = 9$

   *p-hodnota*: $<< 0,001 < 0,05 \Rightarrow$ zamítáme nulovou hypotézu

   *Rozhodnutí*:
   Na hladině významnosti 0,05 zamítáme nulovou hypotézu ($\chi^2$ test dobré shody, $p$-hodnota $<< 0,001$) $\Rightarrow$ s pravděpodobností 95 % můžeme tvrdit, že časové odstupy nelze považovat za veličinu s normálním rozdělením pravděpodobnosti.

   [$p$-hodnota $<< 0,001$]

5. Rozhodněte na základě datového souboru experimentovani-s-telem.xls (Dudová, J. - Experimentování s tělem (výsledky průzkumu), 2013. Dostupné online na http://experimentovani-s-telem.vyplnto.cz), zda existuje souvislost mezi pohlavím respondentů a tím, zda mají tetování. Pro posouzení míry kontingence použijte Cramerovo $V$.

   *Populace*: Všichni, kteří mohou vyplnit dotazník na vyplnto.cz.

   *Předpoklad pro použití metod statistické indukce*: K dispozici je reprezentativní výběr z populace!!!

   *Vizuální posouzení*: Mozaikový graf je členitý, mezi pohlavím respondentů a přítomností tetování by mohla být závislost.

   *Míra kontingence*: $V = \sqrt{\frac{K}{n(min(r;s)-1)}} = \sqrt{\frac{9,59}{596 \cdot (2-1)}} = 0,13$

   $\Rightarrow$ slabá závislost

   *Zvolený test*: $\chi^2$ test nezávislosti v kontingenční tabulce

   *Předpoklady testu*: Očekávané četnosti $\geq 2$, alespoň 80% očekávaných četností $> 5$ $\Rightarrow$ OK

   *Nulová hypotéza* $H_0$: Mezi pohlavím respondentů a přítomností tetování neexistuje závislost.

   *Alternativní hypotéza* $H_A$: $\neg H_0$

   *p-hodnota*: $0,002 < 0,05 \Rightarrow$ zamítáme nulovou hypotézu

   *Rozhodnutí*:
   Na hladině významnosti 0,05 zamítáme nulovou hypotézu ($\chi^2$ test nezávislosti v kontingenční tabulce, $p$-hodnota = 0,002) $\Rightarrow$ s pravděpodobností 95 % můžeme tvrdit, že mezi pohlavím respondentů a tím, zda mají tetování existuje statisticky významná závislost, která není příliš těsná. (Cramerovo $V$ = 0,13)

   [$p$-hodnota $= 0,002$]

6. Pro diferencovaný přístup v personální politice potřebuje vedení podniku vědět, zda spokojenost v práci závisí na tom, jedná-li se o pražský závod či závody mimopražské. Výsledky šetření jsou v následující tabulce. Zobrazte data pomocí mozaikového grafu a na základě testu nezávislosti v kombinační tabulce rozhodněte o závislosti spokojenosti v zaměstnání na umístnění podniku. Pro posouzení míry kontingence použijte Cramerovo V.

| Stupeň spokojenosti | Místo | |
|---|---|---|
| | Praha | Venkov |
| Velmi spokojen | 15 | 40 |
| Spíše spokojen | 50 | 130 |
| Spíše nespokojen | 25 | 10 |
| Velmi nespokojen | 10 | 20 |

   *Vizuální posouzení*: Mozaikový graf je členitý, mezi spokojeností v zaměstnání a umístěním podniku by mohla existovat závislost.
   (Pozor na ordinárnost proměnné)

   *Míra kontingence*: $V = \sqrt{\frac{K}{n(min(r;s)-1)}} = \sqrt{\frac{9,59}{300(4-1)}} = 0,296$

   $\Rightarrow$ středně silná závislost

| | |
|---|---|
| *Zvolený test*: | $\chi^2$ test nezávislosti v kontingenční tabulce |
| *Předpoklady testu*: | Očekávané četnosti $\geq 2$, alespoň 80% očekávaných četností $> 5$ $\Rightarrow$ OK |
| *Nulová hypotéza* $H_0$: | Spokojenost v zaměstnání nezáleží na umístění podniku |
| *Alternativní hypotéza* $H_A$: | $\neg H_0$ |
| *p-hodnota*: | $<< 0,001 < 0,05 \Rightarrow$ zamítáme nulovou hypotézu |

*Rozhodnutí*:
Na hladině významnosti 0,05 zamítáme nulovou hypotézu ($\chi^2$ test nezávislosti v kontingenční tabulce, phodnota $<< 0,001$) $\Rightarrow$ s pravděpodobností 95 % můžeme tvrdit, že mezi spokojeností v zaměstnání a umístěním podniku existuje závislost.

[$p$-hodnota $= 0,002$]

7. V letech 1965 až 1968 bylo v kohortové studii kardiovaskulárních onemocnění v rámci „Honolulu Heart Program“ zahájeno sledování 8 006 mužů, z nichž 7 872 nemělo při zahájení studie v anamnéze mrtvici (apoplexii). Z tohoto počtu bylo 3 435 kuřáků a 4 437 nekuřáků. Při jejich sledování po dobu 12 let dostalo mrtvici 171 mužů ve skupině kuřáků a 117 mužů ve skupině nekuřáků. (Zdroj: Malý, M., Zvárová, M., *Statistické metody v epidemiologii*, Praha, 2003, ISBN: 8 024 607 654)

   a) Zapište zjištěné výsledky do asociační tabulky.

      Řešení:

| kuřáci\apoplexie | ANO | NE | Celkem |
|---|---|---|---|
| ANO | 171 | 3 264 | 3 435 |
| NE | 117 | 4 320 | 4 437 |
| Celkem | 288 | 7 584 | 7 872 |

   b) Na základě vizuálního posouzení odhadněte vliv kouření na výskyt kardiovaskulárních onemocnění.

      Řešení:

      Mozaikový graf je členitý, odhadujeme, že kouření má vliv na výskyt kardiovaskulárních onemocnění.

   c) Určete absolutní riziko vzniku kardiovaskulárních onemocnění u kuřáků a nekuřáků.

      Řešení:

$$\text{exponovaní} = \text{kuřáci} = \frac{a}{a+b} = \frac{171}{3\,435} = 0,050$$

$\Rightarrow$ na 1 000 kuřáků připadá skoro 50 s onemocněním

$$\text{neexponovaní} = \text{nekuřáci} = \frac{c}{c+d} = \frac{117}{4\,437} = 0,026$$

$\Rightarrow$ na 1 000 nekuřáků připadá 26 s onemocněním

[$0,050; 0,026$]

d) Určete relativní riziko (včetně 95% intervalového odhadu) vzniku kardiovaskulárních onemocnění u kuřáků a nekuřáků. Vysvětlete praktický význam zjištěných výsledků.

Řešení:

$$\widehat{RR} = \frac{a(c+d)}{c(a+b)} = \frac{171 \cdot 4\,437}{117 \cdot 3\,435} = 1,89$$

$\Rightarrow$ Riziko vzniku onemocnění je u kuřáků cca 1,89 × vyšší než u nekuřáků.

$$\sqrt{\frac{b}{a(a+b)} + \frac{d}{c(c+d)}} \cdot z_{1-\frac{\alpha}{2}} = sqrt\frac{3\,264}{171 \cdot 3\,435} + \frac{4\,320}{117 \cdot 4\,342} \cdot 1,96 = 0,23$$

$$P(1,50 < \widehat{RR} < 2,38) = 0,95$$

Intervalový odhad nezahrnuje 1 $\Rightarrow$ s 95% pravděpodobností kouření zvyšuje riziko kardiovaskulárních onemocnění.

$$[\widehat{RR} = 1,89; P(\widehat{RR} \in \langle 1,50; 2,38\rangle) = 0,95]$$

e) Určete absolutní šance vzniku kardiovaskulárních onemocnění u kuřáků a nekuřáků.

Řešení:

$$\text{kuřáci} = \frac{a}{b} = \frac{171}{3\,264} = 0,052$$

$\Rightarrow$ Šance vzniku onemocnění u kuřáků je 52:1 000.

$$\text{nekuřáci} = \frac{c}{d} = \frac{117}{4\,320} = 0,027$$

$\Rightarrow$ Šance vzniku onemocnění u nekuřáků je 27:1 000.

[kuřáci: 0,052; nekuřáci: 0,027]

f) Určete relativní šanci (včetně 95% intervalového odhadu) vzniku kardiovaskulárních onemocnění u kuřáků a nekuřáků. Vysvětlete praktický význam zjištěných výsledků.

Řešení:

$$\widehat{OR} = \frac{ad}{bc} = \frac{171 \cdot 4320}{117 \cdot 3264} = 1,93$$

$$\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \cdot z_{1-\frac{\alpha}{2}} = \sqrt{\frac{1}{171} + \frac{1}{3\,264} + \frac{1}{117} + \frac{1}{4\,320}} \cdot 1,96 = 0,24$$

$$P(1,52 < \widehat{OR} < 2,46) = 0,95$$

Intervalový odhad nezahrnuje 1 $\Rightarrow$ s 95% pravděpodobností kouření zvyšuje šanci na výskyt kardiovaskulárních onemocnění.

$$[\widehat{OR} = 1,93;\ P(\widehat{OR} \in \langle 1,52; 2,46\rangle) = 0,95]$$

g) Rozhodněte na hladině významnosti 0,05 o závislosti výskytu kardiovaskulárních chorob na kouření.

Řešení:

Protože intervalové odhady $OR$ a $RR$ nezahrnují 1, zamítáme hypotézu o nezávislosti výskytu kardiovaskulárních chorob na kouření. (Nebo lze použít test nezávislosti v asociační tabulce.)

# 15. Korelační a regresní analýza

## Analýza závislosti v normálním rozdělení

Pearsonův (výběrový) korelační koeficient: $r = \frac{s_{XY}}{s_X \cdot s_Y}$, kde $s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$, $s_X (s_Y)$ je výběrová směrodatná odchylka proměnné $X$ $(Y)$.

| Název testu | Testované parametry | Předpoklady testu | Testová statistika $T(X,Y)$ | Nulové rozdělení |
|---|---|---|---|---|
| Test nulovosti korelačního koeficientu | $\rho$ | normalita | $T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ | $t_{n-2}$ |

## Analýza závislosti ordinálních veličin

Spearmanův korelační koeficient: $r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_{X_i} - R_{Y_i})^2$

| Název testu | Testované parametry | Předpoklady testu | Testová statistika $T(X,Y)$ | Kritický obor |
|---|---|---|---|---|
| Test nulovosti korelačního koeficientu | $\rho$ | — | $T = \lvert r_s \rvert$ | $W = \{T : T \geq r_S^*(\alpha)\}$ (Tabuka 15) |

## Doporučený postup při korelační a regresní analýze

1. Explorační analýza korelačního pole (případný odhad typu regresní funkce, identifikace vlivných bodů)
2. Odhad koeficientů regresní funkce (aplikace vyrovnávacího kritéria - např. metody nejmenších čtverců)
3. Verifikace modelu, tj. ověření předpokladů lineárního modelu
   a) *Celkový F-test* - testujeme, zda hodnota vysvětlované proměnné závisí na lineární kombinaci vysvětlujících proměnných, tj. testujeme nulovou hypotézu $H_0 : \beta_1 = \cdots = \beta_k$ vůči alternativě $H_A : \overline{H_0}$. Pokud bychom nulovou hypotézu nezamítli, znamenalo by to, že model je chybně specifikován.
   b) *Dílčí t-testy* - umožňují testovat oprávněnost setrvání vysvětlující proměnné v regresním modelu. Testujeme (postupně pro jednotlivá i)nulovou hypotézu ve tvaru $H_0 : \beta_i = 0$ vůči alternativě $H_A : \beta_i \neq 0$ pro $i = 0, 1, \ldots, k$. Pokud pro konkrétní $i$ nelze zamítnout nulovou hypotézu, je třeba zvážit setrvání příslušné vysvětlující proměnné v modelu.
   c) *Analýza reziduí* - ověřujeme předpoklady pro použití lineárního regresního modelu.
      - ověření normality reziduí - testy dobré shody,
      - ověření nulovosti střední hodnoty - vizuálně na základě grafu reziduí a odhadovaných hodnot závisle proměnné (rezidua musí kolísat kolem nuly) + dvouvýběrový $t$ test,
      - ověření homoskedasticity - vizuálně na základě grafu reziduí a odhadovaných hodnot závisle proměnné (rezidua se systematicky nezvyšují ani se systematicky nesnižují spolu s rostoucími odhadovanými hodnotami),

- ověření autokorelace reziduí - vizuálně na základě grafu reziduí a odhadovaných hodnot závisle proměnné (autokorelace projeví tak, že se rezidua systematicky snižují nebo zvyšují, resp. můžeme mezi rezidui a předpovídanými hodnotami pozorovat nelineární závislost) + Durbinova-Watsonova statistika.

d) *Multikolinearita* - v případě vícenásobné regrese musíme ověřit, zda neexistuje multikolinearita mezi regresory.

e) *Ověření kvality modelu* - *index determinace* $R^2$ (udává kolik procent vysvětlované proměnné bylo vysvětleno modelem), *koeficient korelace* $r$ (míra korelace mezi závisle proměnnou a regresorem v případě přímkové regrese), *koeficient vícenásobné korelace* $r_{Y\cdot x_1,x_2,\ldots,x_k}$ (míra korelace mezi závisle proměnnou na lineární kombinaci regresorů $x_1, x_2, \ldots, x_k$), *koeficienty parciální korelace*, např. $r_{Y,x_1\cdot x_2,\ldots,x_k}$ (míra korelace mezi závisle proměnnou a jedním z regresorů při vyloučení vlivu ostatních regresorů).

4. Využití verifikovaného modelu k predikci - odhad střední hodnoty závisle proměnné při daných hodnotách regresorů (pás spolehlivosti), odhad individuální hodnoty závisle proměnné při daných hodnotách regresorů (pás predikce). Pozor na extrapolaci!

## 15.1 Test z teorie

1. Kvalita 50 různých výukových materiálů byla dvěma odborníky hodnocena na stupnici od 1 do 5. Vhodnou mírou závislosti mezi hodnocením jednotlivých odborníků je

   a) Spearmanův korelační koeficient,
   b) Pearsonův korelační koeficient,
   c) korigovaný koeficient kontingence,
   d) Cramerovo $V$.

2. Hodnoty Pearsonova korelačního koeficientu blízké nule vypovídají o tom, že

   a) sledované veličiny $X$ resp. $Y$ nenesou prakticky žádnou informaci o $Y$ resp. $X$,
   b) mezi sledovanými veličinami $X$ a $Y$ existuje silná lineární závislost,
   c) mezi sledovanými veličinami $X$ a $Y$ neexistuje silná lineární závislost,
   d) sledované veličiny $X$ a $Y$ jsou nezávislé.

3. Hodnoty Pearsonova korelačního koeficientu blízké -1 vypovídají o tom, že

   a) sledované veličiny $X$ resp. $Y$ nenesou prakticky žádnou informaci o $Y$ resp. $X$,
   b) na měřených objektech jsou nízké hodnoty veličiny $X$ doprovázeny spíše vysokými hodnotami veličiny $Y$,
   c) na měřených objektech jsou nízké hodnoty veličiny $X$ doprovázeny spíše nízkými hodnotami veličiny $Y$.

4. Regresní a korelační analýza umožňuje získat informace o

   a) tvaru a síle závislosti mezi kvalitativními proměnnými,
   b) tvaru a síle závislosti mezi kvantitativními proměnnými,
   c) tvaru a síle závislosti mezi kvantitativními proměnnými, mezi nimiž je lineární vztah.

5. V případě, že jsou splněny předpoklady lineárního regresního modelu, pak metoda nejmenších čtverců umožňuje nalézt

   a) přesný funkční předpis hledané regresní funkce,
   b) index determinace,
   c) nejlepší odhad koeficientů hledané regresní funkce.

6. Lze metodami lineární regrese nalézt regresní funkci ve tvaru mocninné funkce $f : y = ax^b$, kde $a, b \in \mathbb{R} \setminus \{0\}$?

   a) Ano, tato funkce je lineární vzhledem k parametrům.
   b) Ano, tuto funkci můžeme linearizovat logaritmováním funkčního předpisu.
   $[ln(y) = ln(a) + b \cdot ln(x); y^* = a^* + bx^*]$
   c) Ne, tuto funkci nelze použít k vyjádření regresní funkce.
   d) Ne, toto lze řešit metodami nelineární regrese.

7. Lze metodami lineární regrese nalézt regresní funkci ve tvaru mocninné funkcef: $y = a_0 + a_1x_1 + \cdots + a_kx_k$, kde $a_i \in \mathbb{R}, i = 0, 1, \ldots, k$?

   a) Ano, tato funkce je lineární vzhledem k parametrům.
   b) Ano, tuto funkci můžeme linearizovat logaritmováním funkčního předpisu.
   c) Ne, tuto funkci nelze použít k vyjádření regresní funkce.
   d) Ne, toto lze řešit metodami nelineární regrese.

8. Lze metodami lineární regrese nalézt regresní funkci ve tvaru mocninné funkce $f : y = a_0 + a_1 e^{a_2 x_2}$,kde $a_i \in \mathbb{R}, i = 0, 1, 2$?

   a) Ano, tato funkce je lineární vzhledem k parametrům.
   b) Ano, tuto funkci můžeme linearizovat logaritmováním funkčního předpisu.
   c) Ne, tuto funkci nelze použít k vyjádření regresní funkce.
   (d) Ne, toto lze řešit metodami nelineární regrese.

9. Koeficienty regresní funkce jsou

   a) konstanty,
   (b) náhodné veličiny.

10. Index determinace může nabývat hodnot z intervalu

   a) $\langle -1; 1 \rangle$,
   (b) $\langle 0; 1 \rangle$,
   c) $\langle 0; \infty \rangle$.

11. Rezidua jsou odchylky

   (a) pozorovaných a odhadovaných hodnot závislé proměnné,
   b) pozorovaných a odhadovaných hodnot nezávislé proměnné,
   c) pozorovaných a odhadovaných regresních funkcí.

12. S rostoucím rozptylem reziduí se odhad rozptylu odhadů regresních koeficientů

   (a) zvyšuje,
   b) snižuje.

13. S rostoucím rozptylem jednotlivých regresorů se rozptyl odhadů regresních koeficientů

   a) zvyšuje,
   (b) snižuje.

14. K ověření, zda hodnota vysvětlované proměnné závisí na lineární kombinaci všech vysvětlujících proměnných, používáme

   (a) celkový $F$-test,
   b) dílčí $t$-testy,
   c) analýzu reziduí,
   d) index determinace.

15. K testování oprávněnosti setrvání jednotlivých vysvětlujících proměnných v regresním modelu používáme

   a) celkový $F$-test,
   (b) dílčí $t$-testy,
   c) analýzu reziduí,
   d) index determinace.

16. Rezidua považujeme za nekorelované, pokud Durbin-Watsonova statistika leží vintervalu

    a) $\langle -1; 1\rangle$,
    b) $\langle 0; 1\rangle$,
    c) $\langle 0; \infty\rangle$,
    d) $\langle 1,4; 2,6\rangle$,
    e) $\langle 1,4; 2,6)$ nebo $(2,6; \infty\rangle$.

17. Pojmem multikolinearita označujeme

    a) lineární závislost mezi vysvětlovanou proměnnou a jednotlivými vysvětlujícími proměnnými,
    b) lineární závislost mezi vysvětlujícími proměnnými,
    c) lineární závislost mezi vysvětlovanými proměnnými,
    d) lineární závislost mezi jednotlivými regresními funkcemi.

18. Pás spolehlivosti (odhad střední hodnoty závisle proměnné při daných hodnotách regresorů) je

    a) stejně široký jako,
    b) širší než,
    c) užší než,
    c) pás predikce.

19. Odhad závislé proměnné pro hodnoty regresorů ležící mimo interval pozorovaných hodnot označujeme jako

    a) interpolaci,
    b) extrapolaci,
    c) korelaci.

    pás predikce.

20. Doplňte:

    a) K ověření, zda hodnota vysvětlované proměnné závisí na lineární kombinaci vysvětlujících proměnných, používáme .....

    [celkový $F$-test]

    b) K testování oprávněnosti setrvání jednotlivých vysvětlujících proměnných v regresním modelu používáme .....

    [dílčí $t$ testy]

    c) Při analýze reziduí ověřujeme ....., ....., ....., .....

    [normalitu, nulovou stř. hodnotu, homoskedasticitu, neautokorelovanost reziduí]

    d) Pojmem multikolinearita označujeme .....

    [lineární závislost mezi regresory]

## 15.2 Příklady

1. Máme k dispozici výsledky prvního a druhého zápočtového testu deseti studentů. Na hladině významnosti 0,05 testujte hypotézu, že výsledky zápočtových testů jsou kladně korelované.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $X_i$ (1.test) | 7 | 8 | 10 | 4 | 14 | 9 | 6 | 2 | 13 | 5 |
| $Y_i$ (2.test) | 9 | 7 | 12 | 6 | 15 | 6 | 8 | 4 | 11 | 8 |

*Vizuální posouzení závislosti:*

Graf poukazuje na středně silnou korelaci, neobsahuje odlehlá pozorování.

*Korelační koeficienty pro spojitá data:*

- Pearsonův - pouze pro data pocházející z normálního rozdělení
- Spearmanův - pouze v případě, kdy považujeme hodnotící škálu za diskrétní, resp. když není splněn předpoklad normality

*Ověření normality*:

$H_0$ : data pochází z normálního rozdělení
$H_A : \neg H_0$

| Shapiro-Wilkův test | $p$-hodnota |
|---|---|
| Test 1 | 0,94 > 0,05 ⇒ OK |
| Test 2 | 0,73 > 0,05 ⇒ OK |

Na hladině významnosti 0,05 nezamítáme předpoklad normality.
(Shapiro-Wilkův test, $p$-hodnoty: 0,94 a 0,73)

*Zvolený korelační koeficient*: Pearsonův

*Bodový odhad korelačního koeficientu*: $r = 0,845$

*Test, zda lze korelační koeficient považovat za kladný*:

$H_0 : \rho = 0$ data pochází z normálního rozdělení
$H_A : \rho > 0$
$p$-hodnota $= 0,001 < 0,05 \Rightarrow$ zamítáme nulovou hypotézu

Na hladině významnosti 0,05 zamítáme nulovou hypotézu ⇒ s pravděpodobností 95 % můžeme tvrdit, že mezi výsledky prvního a druhého testu existuje statisticky významná pozitivní korelace ($r \cong 0{,}845$).

[$p$-hodnota $= 0,021$]

2. V níže uvedené tabulce je zaznamenána spotřeba alkoholu a úmrtnost na cirhózu jater a alkoholismus ve vybraných zemích. Určete, zda úmrtnost na cirhózu jater a alkoholismus závisí na spotřebě alkoholu. (Zadání příkladu bylo převzato z [1]).

| země | spotřeba alkoholu [l/osoba] | úmrtnost na cirhózu jater a alkoholismus [počet zemřelých na 100 000 obyvatel] |
|---|---|---|
| Finsko | 3,9 | 3,6 |
| Norsko | 4,2 | 4,3 |
| Irsko | 5,6 | 3,4 |
| Holandsko | 5,7 | 3,7 |
| Švédsko | 6,0 | 7,2 |
| Anglie | 7,2 | 3,0 |
| Belgie | 10,8 | 12,3 |
| Rakousko | 10,9 | 7,0 |
| SRN | 12,3 | 23,7 |
| Itálie | 15,7 | 23,6 |
| Francie | 24,7 | 46,1 |

*Vizuální posouzení závislosti:*

Graf poukazuje na silnou korelaci, neobsahuje odlehlá pozorování.

*Korelační koeficienty pro spojitá data:*

- Pearsonův
- Spearmanův

*Ověření normality*:

$H_0$ : data pochází z normálního rozdělení
$H_A : \neg H_0$

| Shapiro-Wilkův test | $p$-hodnota |
|---|---|
| Spotřeba | $0{,}034 < 0{,}05$ |
| Úmrtnost | $0{,}002 < 0{,}05$ |

Na hladině významnosti 0,05 zamítáme předpoklad normality pro oba výběry. (Shapiro-Wilkův test, $p$-hodnoty: 0,034 a 0,002)

*Zvolený korelační koeficient*: Spearmanův

*Bodový odhad korelačního koeficientu*: $r_S \cong 0,773$

*Test, zda lze korelační koeficient považovat za kladný*:

$H_0 : \rho = 0$
$H_A : \rho > 0$
$p$-hodnota $= 0,008 < 0,05 \Rightarrow$ zamítáme nulovou hypotézu

Na hladině významnosti 0,05 zamítáme nulovou hypotézu $\Rightarrow$ s pravděpodobností 95 % můžeme tvrdit, že pozitivní korelace mezi spotřebou alkoholu a cirhózou jater je statisticky významná ($r_s \cong 0{,}773$).

[$p$-hodnota $= 0,015$]

3. Procentuální obsah křemíku v surovém železe $K$ závisí na teplotě strusky $T$ (kremik.xls). Navrhněte pro tuto závislost regresní model, verifikujte jej a posuďte jeho kvalitu. Na základě nalezeného modelu odhadněte (včetně příslušného intervalu spolehlivosti) průměrný procentuální obsah křemíku v surovém železe, bude-li při jeho výrobě používaná struska o teplotě 1350 °C. (*Součásti řešení příkladu musí být závěr obsahující: nalezený model, hodnocení jeho věrohodnosti a kvality (slovně) + konkrétní vysvětlení výsledku predikce, včetně komentáře k její věrohodnosti!!!.*)

a) *Vizuální posouzení závislosti:*

Neobsahuje odlehlá pozorování, lineární model

b) *Primárně zvolený regresní model:*

(*na základě vizuálního posouzení a orientačního srovnání alternativních modelů*)

$Y = a + bX$

$K = -2,773 + 0,002 \cdot T \qquad (°C)$

c) *Ověření normality závislé proměnné i regresoru*:

$H_0$ : data pochází z normálního rozdělení
$H_A : \neg H_0$

| Shapiro-Wilkův test | $p$-hodnota |
|---|---|
| $T$ | 0,750 > 0,05 ⇒ OK |
| $K$ | 0,189 > 0,05 ⇒ OK |

Na hladině významnosti 0,05 nezamítáme předpoklad normality.
(Shapiro-Wilkův test, $p$-hodnoty: 0,750 a 0,189)

d) *Dílčí t-testy*:

$H_0 : \alpha = 0$ $\qquad$ $H_0 : \beta = 0$
$H_A : \alpha \neq 0$ $\qquad$ $H_A : \beta \neq 0$
$p$-hodnota $< 0,001$ $\qquad$ $p$-hodnota $< 0,001$

⇒ Konstantu ani teplotu nelze z modelu vypustit.

e) *Optimalizovaný regresní model*:

$K = -2,773 + 0,002 \cdot T \qquad (°C)$

f) *Analýza reziduí*:

*Ověření normality*:

$H_0$ : rezidua lze modelovat normálním rozdělením
$H_A : \neg H_0$
$p$-hodnota S-W testu $= 0,677 > 0,05 \Rightarrow$

Rezidua lze modelovat normálním rozdělením.

*Ověření nulové střední hodnoty*:

$H_0 : \mu_r = 0$
$H_A : \mu_r \neq 0$
$p$-hodnota $t$-testu $>> 0,999$

Na hladině významnosti 0,05 nezamítáme nulovou hypotézu o nulovosti střední hodnoty reziduí.

*Ověření nulové autokorelace*:

$H_0$ : rezidua nejsou autokorelované
$H_A : \neg H_0$
$p$-hodnota Durbin-Wats. testu$= 0,065 > 0,05 \Rightarrow$

Na hladině významnosti 0,05 nezamítáme nulovou hypotézu o nulové autokorelaci reziduí.

*Ověření homoskedasticity*:

Dle vizuálního posouzení grafu závislosti reziduí na predikovaných hodnotách nelze usuzovat na porušení předpokladu homoskedasticity.

g) *Index determinace*: 91,6 %

*Hodnocení kvality modelu*:

Model je velmi kvalitní, teplota vysvětluje obsah křemíku v železe z 91,6 %.

h) *Závěr*:

$$K = -2,773 + 0,002 \cdot T \qquad (°C)$$

i) *Predikce*: $T = 1350$ °C $\Rightarrow K \cong 33,8$ %

*Bodový odhad*: 0,338

*Intervalový odhad*: $P(0,236 < K < 0,439) = 0,95$

*Komentář k věrohodnosti predikce*:

Vzhledem k tomu, že se jedná o interpolaci, považujeme predikci za důvěryhodnou. Se spolehlivostí 95 % lze při teplotě 1305 °C očekávat obsah křemíku v rozmezí 23,6 % - 43,9 %.

4. Proč nestačí při regresní analýze pouze najít odhad regresní funkce a index determinace? Srovnejte výsledky regresní analýzy pro tzv. Anscombův kvartet - Anscombe.xls. Které předpoklady pro použití lineárního regresního modelu jsou porušeny u jednotlivých sad dat?

*Sada 1*:

OK

*Sada 2*:

autokorelace reziduí

*Sada 3*:

přítomnost odlehlé hodnoty

*Sada 4*:

normalita regresoru

5. Byla vyšetřována výška 20 18letých mladíků $y$ a výška jejich rodičů a prarodičů, žijících izolovaně v horské vesnici po několik generací a hledaná lineární závislost mezi závisle proměnnou $y$ a nezávisle proměnnými $x_1$ až $x_7$ (podrobnější popis regresorů - viz Mladici.xls). Navrhněte pro tuto závislost regresní model, verifikujte jej a posuďte jeho kvalitu. Na základě nalezeného modelu predikujte výšku 18letého mladíka z dat jeho rodičů a prarodičů: $x_1 = 51$ cm, $x_2 = 152$ cm, $x_3 = 183$ cm, $x_4 = 155$ cm, $x_5 = 180$ cm, $x_6 = 157$ cm, $x_7 = 178$ cm. (*Součástí řešení příkladu musí být závěr obsahující: nalezený model, hodnocení jeho věrohodnosti a kvality* (*slovně*) + *konkrétní vysvětlení výsledku predikce, včetně komentáře k její věrohodnosti*!!!)

a) *Vizuální posouzení*:

Nelze identifikovat vlivné body.

b) *Identifikace vlivných bodů*:

—

c) *Primárně zvolený regresní model*:

$$Y = -193,7 + 1,4X_1 + 0,8X_2 + 1,0X_3 - 0,1X_4 + 0,1X_5 + 0,1X_6 - 0,1X_7$$

d) *Ověření normality závislé proměnné i regresorů*:

Dle vizuálního posouzení nezamítáme předpoklad normality.

e) *Multikolinearita*:

Maximální korelační koeficient je mezi $X_1$ a $X_2$ (0,431) $\Rightarrow$ předpoklad neexistence multikolinearity nelze zamítmout.

f) *Celkový F-test*:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_7 = 0$
$H_A : \neg H_0$
$p$-hodnota $<< 0,001 < 0,05$

Na hladině významnosti 0,05 zamítáme nulovou hypotézu ($F$-test, $p$-hodnota $<< 0,001$), model je statisticky významný.

g) *Dílčí t-testy*:

$H_0 : \beta_i = 0 \; i = 1, \ldots, 7$
$H_A : \beta_i \neq 0$

| $i$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| $p$-hodnota | 0,015 | 0,021 | 0,003 | $<< 0,001$ | 0,486 | 0,592 | 0,583 | 0,508 |

$\Rightarrow$ regresory $X_4 - X_7$ lze z modelu vypustit.

Na hladině významnosti 0,05 zamítáme nulovou hypotézu (Studentův $t$-test, $p$-hodnoty dány tabulkou), model je statisticky významný.

h) *Optimalizovaný regresní model*:

$Y = -199,7 + 1,4X_1 + 0,7X_2 + 1,1X_3$ (Pozor - nutné provést přepočet)

i) *Analýza reziduí*:

$H_0$ : rezidua lze modelovat normálním rozdělením
$H_A : \neg H_0$
$p$-hodnota Shapirova-Wilkova testu $= 0,302 > 0,05 \Rightarrow$

Rezidua lze modelovat normálním rozdělením.

*Ověření nulové střední hodnoty*:

$H_0 : \mu_r = 0$
$H_A : \mu_r \neq 0$
$p$-hodnota $t$-testu $>> 0,999$

Na hladině významnosti 0,05 nezamítáme nulovou hypotézu o nulovosti střední hodnoty reziduí (Studentův $t$-test, $p$-hodnota $>> 0,999$ ).

*Ověření nulové autokorelace*:

$H_0$ : rezidua nejsou autokorelované
$H_A : \neg H_0$
$p$-hodnota Durbin-Wats. testu$= 0,181 > 0,05 \Rightarrow$

Na hladině významnosti 0,05 nezamítáme nulovou hypotézu o nulové autokorelaci reziduí (Durbin-Wats. test, $p$-hodnota = 0,181).

*Ověření homoskedasticity*:

Dle vizuálního posouzení grafu závislosti reziduí na predikovaných hodnotách nelze usuzovat na porušení předpokladu homoskedasticity.

j) *Index determinace*: 88,7 %

*Hodnocení kvality modelu*:

Model je velmi kvalitní.

k) *Závěr*:

Výšku 18letého chlapce lze predikovat dle jeho porodní délky a výšky rodičů v 18 letech. Dané regresory vysvětlují výšku chlapce z 88,7 %.

l) *Predikce*:

*Bodový odhad*: 177,2 cm

*Intervalový odhad*: $P(170,4 < Y < 184,0) = 0,95$

*Komentář k věrohodnosti predikce*:

Jedná se o interpolaci, tj. nalezená predikce je důvěryhodná. Se spolehlivostí 95 % lze výšku chlapce očekávat v rozmezí 170,5 - 184,0 cm, tj. (177,4 $\pm$ 6,8 cm).