Pravděpodobnost a statistika pro FEL

Blanka Šedivá, Patrice Marek, Tomáš Ťoupal, Eva Wagnerová

Cíl kurzu: Základní počet pravděpodobnosti, náhodná proměnná, náhodný vektor, limitní věty, statistické soubory, náhodný výběr, odhady parametrů, testování hypotéz, regresní analýza, statistická kontrola kvality.

Poslední aktualizace:17. března 2011

Obsah

1 Cvičení 1 - Náhodný pokus, náhodný jev

Náhodný pokus, náhodný jev. Operace s jevy. Definice pravděpodobnosti jevu, vlastnosti ppsti. Klasická definice pravděpodobnosti a její použití, základní kombinatorické vzorce.

1.1 Teoretická část

1.1.1 Definice základních pojmů

Náhodný pokus je každý proces, jehož výsledek je při jinak stejných počátečních podmínkách nejistý; výsledek nejsme schopni s jistotou předpovědět; množinu všech možných výsledků náhodného pokusu označujeme Ω.

Náhodný jev je jev A je podmnožina množiny Ω ($A \subset \Omega$); náhodné jevy značíme velkými latinskými písmeny z počátku abecedy A,B,C, ...; celá množina Ω je jev jistý; prázdná množina $\emptyset$ je jev nemožný.

Elementární jevy jsou ω_i jsou minimální jevy různé od jevu nemožného

(ω je elementární jev: $\forall A \subset \omega \Rightarrow (A \equiv \omega)$ nebo $(A \equiv \emptyset)$;

elementární jevy jsou párově neslučitelné (ω_1, ω_2 různé elementární jevy, pak $\omega_1 \cap \omega_2 \equiv \emptyset$);

každý jev A lze vyjádřit jako množinu elementárních jevů ($A \equiv \{\omega_1, \omega_2, \dots\}$).

Operace s jevy , protože jevy mají charakter množin, můžeme je graficky znázorňovat pomocí Vénnových diagramů

- $A \equiv B$ rovnocenné jevy
- $\bar{A}$ (nebo A^c nebo A') $\equiv \Omega \backslash A$ jev opačný, doplněk jevu
- $A \subset B$ jev A je podjevem jevu B
- $A \cap B$ průnik jevů ,jev A a zároveň jev B
- $A \cup B$ sjednocení jevů, jev A nebo jev B (nebo oba jevy)
- $A \backslash B$ rozdíl jevů , platí jev A, ale nikoliv jev B
- $A \cap B \equiv \emptyset$ jevy disjunktní, jevy neslučitelné
- $\bigcup A_i \equiv \Omega$ úplný systém jevů
- zákon jedinečnosti: $\forall A, B \;\; \exists!\; A \cap B$ a $\exists!\; A \cup B$
- zákon komutativní: $A \cup B \equiv B \cup A$, resp. $A \cap B \equiv B \cap A$
- zákon asociativní: $(A \cup B) \cup C \equiv A \cup (B \cup C)$ resp. $(A \cap B) \cap C \equiv A \cap (B \cap C)$
- zákon identity: $A \cup \emptyset \equiv A \qquad A \cap \emptyset \equiv \emptyset$, resp. $A \cup \Omega \equiv \Omega \qquad A \cap \Omega \equiv A$
- zákon komplementu: $A \cup \bar{A} \equiv \Omega$, resp. $A \cap \bar{A} \equiv \emptyset$
- zákon distributivní: $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$ resp. $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$
- de Morganovy vzorce: $\overline{A \cup B} \equiv \overline{A} \cap \overline{B}$, resp. $\overline{A \cap B} \equiv \overline{A} \cup \overline{B}$
  obecněji nechť $A_i, i = 1, 2, \dots, A_n$ jsou jevy

$$\overline{\bigcup_{i=1}^{n} A_i} \equiv \bigcap_{i=1}^{n} \overline{A_i}\ , \qquad\qquad \overline{\bigcap_{i=1}^{n} A_i} \equiv \bigcup_{i=1}^{n} \overline{A_i}$$

1.1.2 Definice pravděpodobnosti

Pravděpodobnost jevu - každému jevu A přiřazujeme reálné číslo $\mathcal{P}(A)$; pravděpodobnost (ppst) lze chápat jako předpověď poměrných četností výsledků při mnohonásobném opakování daného pokusu; ppst lze chápat jako kvantitativní ohodnocení stupně jistoty. Existují různé možnosti matematického zavedení pravděpodobnosti - klasická ppst, geometrická ppst, statistická ppst a axiomatická ppst.

Klasická definice pravděpodobnosti - předpoklady:

- $\Omega = \{\omega_1, \omega_2, \omega_3, \ldots, \omega_N\}$ množina možných výsledků pokusu je konečná a neprázdná $(0 < N < \infty)$;
- $p_1, p_2, p_3, \ldots, p_N$ jsou nezáporná čísla splňující $\displaystyle\sum_{i=1}^{N} p_i = 1$;
- $p_1 = p_2 = \cdots = p_N = \dfrac{1}{N}$ všechny výsledky pokusu jsou stejně možné;
- každý jev A lze popsat množinou jevů $\{\omega_{i_1}, \omega_{i_2}, \ldots, \omega_{i_k}\}$ kde ω_i jsou výsledky pokusu příznivé jevu A; pak

$$\mathcal{P}(A) = \frac{N_A}{N}$$

kde

N_A je počet výsledků příznivých jevu A
N je počet všech možných výsledků

Geometrická definice pravděpodobnosti - předpoklady

- Ω jsme schopni vyjádřit jako neprázdnou omezenou oblast v $\mathbb{R}^n$ (například pomocí omezené přímky v $\mathbb{R}^1$, omezené plochy v $\mathbb{R}^2$, omezeného tělesa v $\mathbb{R}^3$)
- jev A jsme schopni vyjádřit jako podoblast oblasti Ω, pak

$$\mathcal{P}(A) = \frac{\lambda(A)}{\lambda(\Omega)}$$

kde $\lambda(A)$ je *míra* oblasti A (délka, obsah plochy, objem tělesa) a $\lambda(\Omega)$ je *míra* oblasti Ω (délka, obsah plochy, objem tělesa).

Vlastnosti pravděpodobnosti - pro všechny jevy A_i, $i = 1, 2, 3, \dots$ platí

- $0 \leq \mathcal{P}(A_i) \leq 1$
- jsou-li A_i a A_j neslučitelné, potom $\mathcal{P}(A_i \cup A_j) = \mathcal{P}(A_i) + \mathcal{P}(A_j)$ resp. obecněji jsou-li $A_i, i = 1, 2, \dots$ neslučitelné, potom $\mathcal{P}(\bigcup_i A_i) = \sum_i \mathcal{P}(A_i)$
- $\mathcal{P}(\Omega) = 1,\ \mathcal{P}(\emptyset) = 0$
- $A_i \subset A_j \Rightarrow \mathcal{P}(A_i) \leq \mathcal{P}(A_j)$
- $\mathcal{P}(\overline{A_i}) = 1 - \mathcal{P}(A_i)$
- $A_i \subset A_j \Rightarrow \mathcal{P}(A_j \backslash A_i) = \mathcal{P}(A_j) - \mathcal{P}(A_i)$
- $\mathcal{P}(A_i \cup A_j) \leq \mathcal{P}(A_i) + \mathcal{P}(A_j)$
- $\mathcal{P}(A_i \cup A_j) = \mathcal{P}(A_i) + \mathcal{P}(A_j) - \mathcal{P}(A_i \cap A_j)$

1.1.3 Základní kombinatorické vzorce

Pro určování počtu možných výsledků používáme vzorce pro permutace, variace a kombinace.

- permutace n prvků (kolika způsoby lze uspořádat $n$tici prvků); uspořádání prvků skupiny M v daném pořadí
  - počet permutací $P_n = n!$
  - pokud M se skládá z $i_1, i_2, \dots, i_k$ stejných prvků, je počet permutací $P_n = \dfrac{n!}{i_1! i_2! \dots i_k!}$
  - počet permutací s opakováním $P_n = n^n$
- variace n prvků $k$té třídy (kolika způsoby lze z $n$tici prvků vybrat $k$tici, přičemž záleží na pořadí výběru)
  - počet variací $V_n^k = \dfrac{n!}{(n-k)!} = n(n-1)\dots(n-k+1)$
  - počet variací s opakováním $\overline{V_n^k} = n^k$
- kombinace n prvků $k$té třídy (kolika způsoby lze z $n$tici prvků vybrat $k$tici, přičemž nezáleží na pořadí výběru)
  - počet kombinací $C_n^k = \dbinom{n}{k} = \dfrac{n!}{(n-k)!k!}$
  - počet kombinací s opakováním $\overline{C_n^k} = \dbinom{n+k-1}{k}$
- binomické číslo lze přibližně určit za použití Stirlingovy formule pro určení hodnoty $k!$

$$\log k! \approx \log \sqrt{2\pi k} + k(\log k - \log e)$$

- vlastnosti kombinačních čísel
  - $\binom{n}{k} = \binom{n}{n-k}$
  - $\binom{n}{0} = \binom{n}{n} = 1$
  - $\binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1}$
  - $\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n} = 2^n$
  - $\binom{n}{0} - \binom{n}{1} + \cdots + (-1)^n \binom{n}{n} = 0$

1.2 Příklady

1. 120 studentů absolvovalo zkoušku z matematiky a fyziky. 82 studentů udělalo zkoušku z matematiky, 85 studentů zkoušku z fyziky, 77 studentů udělalo obě zkoušky. Určete:

   (a) Kolik studentů udělalo zkoušku z fyziky nebo z matematiky ?

   (b) Kolik studentů neudělalo zkoušku z fyziky ?

   (c) Kolik studentů neudělalo zkoušku z matematiky ?

   (d) Kolik studentů udělalo zkoušku z fyziky a neudělalo zkoušku z matematiky ?

   (e) Kolik studentů udělalo zkoušku z matematiky a neudělalo zkoušku z fyziky ?

2. Jev A spočívá v tom, že náhodně vybrané přirozené číslo je dělitelné pěti a jev B v tom, že toto číslo má na posledním místě nulu. Určete, co znamenají jevy

   (a) $A \cap B$;

   (b) $A \cup B$;

   (c) $\bar{A} \cap B$;

   (d) $A \cup \bar{B}$;

   (e) $\overline{A \cap \bar{B}}$.

3. Výrobek je v rámci výstupní kontroly podroben třem různým zkouškám. Jev A spočívá v tom, že výrobek obstojí při první zkoušce, jev B spočívá v tom, že výrobek obstojí při druhé zkoušce a jev C v tom, že výrobek obstojí při třetí zkoušce. Vyjádřete v množinové symbolice, že výrobek obstojí

   (a) jen v první zkoušce;

   (b) v první a druhé zkoušce, ale neobstojí ve třetí zkoušce;

   (c) ve všech třech zkouškách;

   (d) alespoň v jedné zkoušce;

   (e) alespoň ve dvou zkouškách;

   (f) maximálně ve dvou zkouškách.

4. Charakterizujte množinu elementárních náhodných jevů pro náhodné pokusy

   (a) hod dvěmi mincemi;

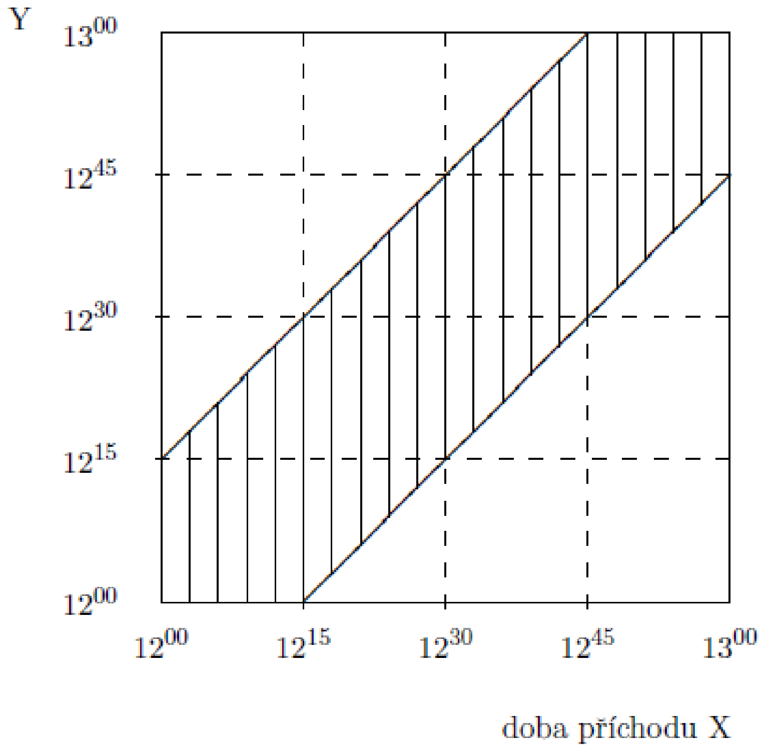
   (b) otočení ruletou.

5. Kolik různých čísel lze vytvořit z číslic 0, 1, 2, 3 a 4.

   (a) smí-li každá z číslic být v čísle obsažena nejvýše jednou, [261]

   (b) je-li počet stejných číslic v čísle neomezený. [nekonečně]

6. V sérii 12 výrobků jsou 3 vadné. Kolika způsoby lze vybrat



Obrázek 1: Příklad 8 - úloha o setkání

(a) šest výrobků, [924]

(b) šest výrobků, všechny bez vady, [84]

(c) šest výrobků, z toho jeden vadný, [378]

(d) šest výrobků, z toho nejvýše dva vadné, [840]

(e) šest výrobků, z toho alespoň dva vadné. [462]

7. V urně je 6 bílých a 3 černé koule. Kolika způsoby lze z urny vytáhnout 4 koule, mají-li mezi nimi být alespoň dvě bílé ? [120]

8. Úloha o setkání: Dva přátelé (X a Y) se domluvili, že přijdou na určité místo v době mezi polednem a jednou hodinou odpoledne. Na místo přijde v tomto časovém intervalu každý z nich zcela náhodně a nezávisle na příchodu toho druhého. Každý bude čekat patnáct minut na příchod druhého, ne déle než do jedné hodiny odpoledne. Úkolem je určit ppst., že se za těchto podmínek sejdou.

   Pravděpodobnost setkání odpovídá podílu obsahu vyšrafované plochy vzhledem k celkové ploše a je $\mathcal{P} = \dfrac{7}{16}$.

1.3 Literatura s dalšími příklady

- Trial KMA – kapitola 30.1 Kombinatorika
- Trial KMA – kapitola 30.2 Náhodné jevy
- Trial KMA – kapitola 30.2 Definice pravděpodobnosti
- Polák, Josef: Středoškolská matematika v úlohách. Strana 74–126
- Reif, Jiří – Kobeda, Zdeněk: Úvod do pravděpodobnosti a spolehlivosti. Strana 9–16

2 Cvičení 2 - Podmíněná ppst, závislost a nezávislost jevů

Pravděpodobnost jevů $A \cap B$, $A \cup B$, $A - B$, $\overline{A}$, ... Závislost a nezávislost jevů, podmíněná pravděpodobnost.

2.1 Teoretická část

2.1.1 Podmíněná pravděpodobnost

Nechť $A, B \in \mathcal{A}$ jsou jevy a nechť $\mathcal{P}(B) > 0$. Podmíněnou pravděpodobnost jevu A za podmínky jevu B definujeme vztahem

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

- je-li $A \cap B = \emptyset$, pak $\mathcal{P}(A|B) = 0$
- $\mathcal{P}(A|B) \neq \mathcal{P}(B|A)$
- $\mathcal{P}(A|B)\, \mathcal{P}(B) = \mathcal{P}(B|A)\, \mathcal{P}(A)$

2.1.2 Závislost a nezávislost jevů

Nezávislost dvou jevů Jevy $A, B \in \mathcal{A}$ se nazývají nezávislé, jestliže platí

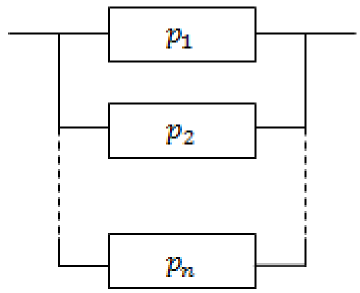
$$\mathcal{P}(A \cap B) = \mathcal{P}(A)\, \mathcal{P}(B)$$

Jevy, které nejsou nezávislé, jsou závislé.
Následující tvrzení jsou ekvivalentní pro $\mathcal{P}(A) \neq 0$ a $\mathcal{P}(B) \neq 0$

- A a B jsou nezávislé
- $\mathcal{P}(A|B) = \mathcal{P}(A)$
- $\mathcal{P}(B|A) = \mathcal{P}(B)$
- $\overline{A}$ a B jsou nezávislé
- A a $\overline{B}$ jsou nezávislé
- $\overline{A}$ a $\overline{B}$ jsou nezávislé

Nezávislost více jevů
Nechť $\{A_i,\ i \in I\}$ je množina jevů. Jevy této množiny se nazývají nezávislé, jestliže pro každé přirozené n a každou podmnožinu $\{i_1, i_2, \dots, i_n\} \subset I$ platí

$$\mathcal{P}(\bigcap_{j=1}^{n} A_{i_j}) = \mathcal{P}(A_{i_1})\, \mathcal{P}(A_{i_2})\, \dots \mathcal{P}(A_{i_n})$$



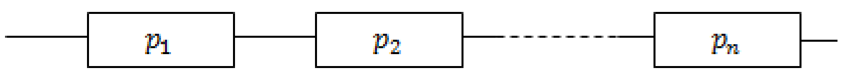
Obrázek 2: Paralelní zapojení prvků

2.1.3 Spolehlivost paralelně a sériově řazených nezávislých prvků

Nechť $p_1, p_2, ...p_n$ jsou pravděpodobnosti poruch prvků $P_1, P_2, ..., P_n$. Předpokládáme, že poruchy jednotlivých prvků jsou na sobě nezávislé.
Pravděpodobnost poruchy celého systému značíme P a spolehlivost celého systému $R = 1 - P$

$$P = p_1 \cdot p_2 \cdot \dots \cdot p_n = \prod_{i=1}^{n} p_i$$

$$R = 1 - p_1 \cdot p_2 \cdot \dots \cdot p_n = 1 - \prod_{i=1}^{n} p_i$$



Obrázek 3: Sériové zapojení prvků

$$P = 1 - (1 - p_1) \cdot (1 - p_2) \cdot \dots \cdot (1 - p_n) = 1 - \prod_{i=1}^{n} (1 - p_i)$$

$$R = (1 - p_1) \cdot (1 - p_2) \cdot \dots \cdot (1 - p_n) = \prod_{i=1}^{n} (1 - p_i)$$

2.2 Příklady

1. Uvažujme následující jevy

   (a) jev A: na kostce padlo číslo „1“ nebo „2“;

   (b) jev B: na kostce padlo číslo sudé („2“,„4“,„6“);

   Spočtěte pravděpodobnosti jevů A, B, $A \cap B$, $A \cup B$, $A|B$ a $B|A$ a rozhodněte, zda jsou jevy A a B nezávislé.

   Řešení:
   Jev $A \cap B$, platí jev A a zároveň jev B odpovídá situaci, kdy na kostce padlo číslo „2“.
   $\mathcal{P}(A) = 1/3$, $\mathcal{P}(B) = 1/2$, $\mathcal{P}(A \cap B) = 1/6$
   $\mathcal{P}(A|B) = \dfrac{\mathcal{P}(A \cap B)}{\mathcal{P}(B)} = \dfrac{1/6}{1/2} = 1/3$,
   $\mathcal{P}(B|A) = 1/2$
   Protože platí $P(A) \cdot P(B) = \dfrac{1}{3} \cdot \dfrac{1}{2} = \dfrac{1}{6} = P(A \cap B)$, jsou jevy A a B nezávislé.

2. Ukázka příkladu, kdy jsou jevy po dvou nezávislé, ale jsou celkově závislé. Uvažujme náhodný pokus „hod dvěmi mincemi“, kdy sledujeme zda na mincích padl líc (L) nebo (R). Množina všech možných výsledků (elementárních jevů) je tedy $\Omega = \{LL, LR, RL, RR\}$ a všechny elementární jevy jsou stejně pravděpodobné, tj. mají pravděpodobnost $\dfrac{1}{4}$.
   Najděte pravděpodobnost a zjistěte zda jsou nezávislé a po dvou nezávislé jevy

   (a) A_1 na první mince padne líc;

   (b) A_2 na druhé minci padne líc;

   (c) A_3 na obou mincích padne totéž.

   Řešení: $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\},\ \mathcal{P}(\omega_i) = 1/4\ A_1 = \{\omega_1, \omega_2\},\ \mathcal{P}(A_1) = 1/2$
   $A_2 = \{\omega_1, \omega_3\},\ \mathcal{P}(A_2) = 1/2$
   $A_3 = \{\omega_1, \omega_4\},\ \mathcal{P}(A_3) = 1/2$
   jevy A_1 a A_2 jsou nezávislé, protože $A_1 \cap A_2 = \{\omega_1\}$ a
   $$\mathcal{P}(A_1 \cap A_2) = 1/4 = 1/2 \cdot 1/2$$
   jevy A_1 a A_3 jsou nezávislé, protože $A_1 \cap A_3 = \{\omega_1\}$ a
   $$\mathcal{P}(A_1 \cap A_3) = 1/4 = 1/2 \cdot 1/2$$
   jevy A_2 a A_3 jsou nezávislé, protože $A_2 \cap A_3 = \{\omega_1\}$ a
   $$\mathcal{P}(A_2 \cap A_3) = 1/4 = 1/2 \cdot 1/2$$
   jevy A_1,A_2 a A_3 jsou závislé, protože $A_1 \cap A_2 \cap A_3 = \{\omega_1\}$ a
   $$\mathcal{P}(A_1 \cap A_2 \cap A_3) = 1/4 \neq 1/2 \cdot 1/2 \cdot 1/2$$

3. Příklad o zapomenutém deštníku
   Roztržitý profesor zapomíná v obchodě deštník s ppstí. $\dfrac{1}{4}$, tedy za podmínky, že s deštníkem do

obchodu dorazí. Postupně navštívil tři obchody a cestou domů zjistí, že deštník nemá. Určete ppsti., že deštník zapomněl v jednotlivých obchodech.

Řešení:
jevy A_i: deštník zapomněl v $i$tém obchodě (jevy jsou disjunktní)
jev $A = A_1 \cup A_2 \cup A_3$: děštník v některém z obchodů zapomněl
jevy $A_i|A$: deštník zapomněl v $i$tém obchodě za podmínky, že deštník v některém obchodě zapomněl
$\mathcal{P}(A_1) = \frac{1}{4}$, $\mathcal{P}(A_2) = \frac{3}{4} \cdot \frac{1}{4}$, $\mathcal{P}(A_3) = \frac{3}{4} \cdot \frac{3}{4} \cdot \frac{1}{4}$
$\mathcal{P}(A) = \mathcal{P}(A_1) + \mathcal{P}(A_2) + \mathcal{P}(A_3) = \frac{37}{64}$
$\mathcal{P}(A_i|A) = \frac{\mathcal{P}(A_i \cap A)}{\mathcal{P}(A)} = \frac{\mathcal{P}(A_i)}{\mathcal{P}(A)}$
$\mathcal{P}(A_1|A) = \frac{16}{37}$, $\mathcal{P}(A_2|A) = \frac{12}{37}$, $\mathcal{P}(A_3|A) = \frac{9}{37}$

4. Vláďa a Jarda hrají ruletu. Víme, že 0 nevyhrála. Vláďa vsadil šestici čísel $22-27$, Jarda vsadil na velkou (tzn. $19-36$).
   (a) Jaká je pravděpodobnost, že Vláďa vyhrál, jestliže Jarda vyhrál.
   (b) Jaká je pravděpodobnost, že Jarda vyhrál, jestliže Vláďa vyhrál.
   (c) Jsou jevy „Vláďa vyhrál“ a „Jarda vyhrál“ nezávislé?

Řešení:

Označme A jev „Vláďa vyhrál“ a B jev „Jarda vyhrál“.

Ze zadání víme, že vyhrává jedno číslo z 36.

Pak $P(A) = \frac{6}{36} = 0.167$ a $P(B) = \frac{18}{36} = 0.5$

Vyjádříme elementární jevy jevu $A \cap B$.

$$A \cap B = \{22, 23, 24, 25, 26, 27\}$$

a spočteme $P(A \cap B) = \frac{6}{36} = 0.167$

(a) Počítáme pravděpodobnost $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.167}{0.5} = 0.333$

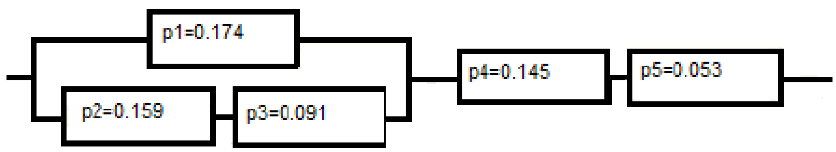
(b) Počítáme pravděpodobnost $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.167}{0.167} = 1$

(c) Pro nezávislé jevy platí $P(A) \cdot P(B) = P(A \cap B)$

$$P(A) \cdot P(B) = \frac{6}{36} \cdot \frac{18}{36} \neq \frac{6}{36} = P(A \cap B)$$

Rovnost neplatí, tedy jevy A a B jsou závislé.

5. Zjistěte pravděpodobnost, že načrtnuté zařízení přestane během doby Δt fungovat, jestliže pravděpodobnost, že $i$tá součástka přestane fungovat během doby Δt je p_i.



Obrázek 4: Schéma zapojení

Řešení:

Nejprve určíme pravděpodobnost selhání bloku (tento blok označme A) složeného z prvků P_2 a P_3:

$p_A = 1 - (1 - p_1) \cdot (1 - p_2) = 1 - (1 - 0.159) \cdot (1 - 0.091) \doteq 0.236.$

Dále spočteme pravděpodobnost selhání bloku (tento blok označme B) složeného z prvků P_1 a A:

$p_B = p_1 \cdot p_A = 0.164 \cdot 0.236 \doteq 0.041$

Nakonec spočteme pravděpodobnost selhání bloku složeného z prvků B, P_4 a P_5:

$p = 1 - (1 - p_B) \cdot (1 - p_4) \cdot (1 - p_5) = 1 - (1 - 0.041) \cdot (1 - 0.145) \cdot (1 - 0.053) \doteq O.224$

Pravděpodobnost selhání načrtnutého zařízení je 22.4%

2.3 Literatura s dalšími příklady

- Trial KMA – kapitola 30.4.1 Podmíněná pravděpodobnost
- Trial KMA – kapitola 30.4.2 Nezávislost jevů
- Reif, Jiří – Kobeda, Zdeněk: Úvod do pravděpodobnosti a spolehlivosti. Strana 17–19.

3 Cvičení 3 - Věta o úplné ppsti, Bayesova věta

3.1 Teoretická část

3.1.1 Věta o úplné pravděpodobnosti

- Nechť $B_1, B_2, \ldots$ tvoří úplný systém disjunktních jevů,
- nechť $P(B_i) > 0$ pro $i = 1, 2, \ldots,$
- nechť jev A je libovolný jev příslušný témuž náhodnému pokusu.

Pak platí

$$P(A) = \sum_i P(A|B_i)\, P(B_i)$$

Důkaz:
Použijeme definice podmíněné pravděpodobnosti $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$ a využijeme vlastnosti pravděpodobnosti pro sjednocení disjunktních jevů
$P(\bigcup_i A_i) = \sum_i P(A_i)$.

3.1.2 Bayesova věta o inverzní pravděpodobnosti

- Nechť $B_1, B_2, \ldots, B_n$ tvoří úplný systém disjunktních jevů,
- nechť $P(B_i) > 0$ pro $i = 1, 2, \ldots, n$ a
- nechť jev A je libovolný jev příslušný témuž náhodnému pokusu takový, že $P(A) > 0$.

Pak platí pro všechna $k = 1, 2, \ldots, n$

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

Důkaz:
Použijeme definici podmíněné pravděpodobnosti a výsledky věty o úplné pravděpodobnosti.

Poznámka: Z hypotéz $B_1, B_2, \ldots, B_n$ nastane při provedení pokusu právě jedna. Jejich pravděpodobnost $P(B_i)$ je známa před provedením pokusu - *a priori* (nezávisle na zkušenosti - na základě rozumu). Víme-li ale, zda při provedení pokusu nastal jev A či nikoli, pak tento fakt mění pravděpodobnosti alternativ na $P(B_i|A)$ - pravděpodobnost *a posteriori* (podle zkušenosti).

3.2 Příklady

1. *Výrobní linky.*
   Na třech výrobních linkách jsou vyráběny identické výrobky.První výrobní linka zajišťuje 60% produkce a ppst. vyrobení vadného výrobku je 1%, druhá výrobní linka zajišťuje 30% produkce a ppst. vyrobení vadného výrobku je 2%, třetí výrobní linka zajišťuje 10% produkce a ppst. vyrobení vadného výrobku je 3%.

(a) Určete ppst., že náhodně vybraný výrobek bude vadný.

(b) Nechť výrobek je vadný. Určete ppst., že náhodně vybraný vadný výrobek pochází z 1., 2. resp. 3. linky.

Řešení: Označme

- jev B_i výrobek je výrobek na i-té lince
- jev A výrobek je vadný

- $P(B_1) = 0.6,\ P(B_2) = 0.3,\ P(B_3) = 0.1$
- $P(A|B_1) = 0.01$
- $P(A|B_2) = 0.02$
- $P(A|B_3) = 0.03$

Ppst., že náhodně vybraný výrobek bude vadný je určena podle věty o úplné ppsti, kde platí $P(A) = 0.6 \cdot 0.01 + 0.3 \cdot 0.02 + 0.1 \cdot 0.03 = 0.015$.

Ppst., že náhodně vybraný vadný výrobek pochází z 1., 2. resp. 3. linky, je určena na základě věty o inverzní pravděpodobnosti.

- $P(B_1|A) = \dfrac{P(A|B_1)\,P(B_1)}{P(A)} = \dfrac{0.6 \cdot 0.01}{0.015} = 0.4$
- $P(B_2|A) = \dfrac{0.3 \cdot 0.02}{0.015} = 0.4$
- $P(B_3|A) = \dfrac{0.1 \cdot 0.03}{0.015} = 0.2$

2. *O výstředním žalářníkovi.*
V žaláři je vězeň odsouzený k smrti. Výstřední žalářník však dá vězni šanci. Přinese 12 černých a 12 bílých kuliček. Pak mu dá dvě prázdné urny a sdělí mu, že zítra přijde kat a náhodně si vybere jednu urnu a z ní náhodně vybere jednu kuličku. Bude-li bílá, dostane vězeň milost. Jak má vězeň rozdělit kuličky, aby maximalizoval pravděpodobnost udělení milosti?

Řešení: Označme

- jev B_j kat vybral $j$tou urnu $(j = 1, 2)$
- n počet kuliček v první urně
- i počet bílých kuliček v první urně
- jev A kat vytáhl bílou kuličku, $P(A_{(n,i)})$ ppst. vytažení bílé kuličky.

Pak platí

- $P(B_1) = 1/2,\ P(B_2) = 1/2$
- $P(A|B_1) = \dfrac{i}{n},\ P(A|B_2) = \dfrac{12-i}{24-n}$

a podle věty o úplné ppsti platí $P(A_{(n,i)}) = \dfrac{1}{2}(\dfrac{i}{n} + \dfrac{12-i}{24-n})$

3. *Příklad z lékařské diagnostiky*

Označme jev CH, že náhodně vybraná osoba má sledovanou chorobu. $P(CH) = 0.5\%$. Předpokládejme, že určitý test na odhalení choroby má následující výsledky:

- Má-li osoba sledovanou chorobu, poskytne test pozitivní výsledek v 95% případů (senzitivita testu).
- Nemá-li osoba sledovanou chorobu, poskytne test negativní výsledek v 90% případů (specificita testu).

Jestliže u náhodně vybrané osoby byl výsledek testu pozitivní, jaká je ppst., že skutečně má sledovanou chorobu?

Řešení:

- jev +: výsledek testu byl pozitivní,
- jev - : byl negativní

Známe ppsti

- $P(+|CH) = 0.95$:
  osoba, která má danou chorobu, má pozitivní výsledek testu
- $P(-|CH) = 0.05$:
  osoba, která má danou chorobu, má negativní výsledek testu
- $P(+|\overline{CH}) = 0.10$:
  osoba, která nemá danou chorobu, má pozitivní výsledek testu
- $P(-|\overline{CH}) = 0.90$:
  osoba, která nemá danou chorobu, má negativní výsledek testu

Podle věty o úplné ppsti platí

$$P(+) = 0.95 \cdot 0.005 + 0.10 \cdot 0.995 = 0.10425$$

a podle Bayesovy věty platí

$$P(CH|+) = \frac{P(+|CH)\,P(CH)}{P(+)} = \frac{0.95 \cdot 0.005}{0.10425} \doteq 4.56\%$$

a dále platí

$$P(CH|-) = \frac{P(-|CH)\,P(CH)}{P(-)} = \frac{0.05 \cdot 0.005}{1 - 0.10425} \doteq 0.03\%$$

$$P(\overline{CH}|+) = \frac{P(+|\overline{CH})\,P(\overline{CH})}{P(+)} = \frac{0.10 \cdot 0.995}{0.10425} \doteq 95.44\%$$

$$P(\overline{CH}|-) = \frac{P(-|\overline{CH})\,P(\overline{CH})}{P(-)} = \frac{0.90 \cdot 0.995}{1 - 0.10425} \doteq 99.97$$

4. *Vadné výrobky*

   Ve skladu je 1000 výrobků, přičemž 100 výrobků pochází od 1. dodavatele, 600 od 2. dodavatele a zbytek od 3. dodavatele. Pravděpodobnost, že 1. dodavatel dodal vadný výrobek, je 1%, u 2. je to 0,5% a u 3. dodavatele 2%.

   Ze skladu náhodně vybereme jeden výrobek a tento výrobek je vadný. S jakou pravděpodobností pochází od 1. dodavatele?

5. *Nákup auta*

   Manželka jde do autosalonu Škoda pro nový vůz. Nemůže se však rozhodnout mezi Octavií a Passatem. Rozhodne se tedy, že si z vozů, které jsou na prodejně, jeden vybere zcela náhodně. Z těchto vozů je 60% Octavií a 40% Passatů, dále víme, že dieslový motor má 30% Octavií a 50% Passatů.

   (a) S jakou pravděpodobností manželka přijede vozem s dieslovým motorem?

   (b) Manželka si vybrala vůz s dieslovým motorem. S jakou pravděpodobností se jedná o Passata?

3.3 Literatura s dalšími příklady

- Trial KMA – kapitola 30.4.3. Věta o úplné pravděpodobnosti
- Trial KMA – kapitola 30.4.4. Inverzní Bayesova věta
- Brousek, Jan – Ryjáček, Zdeněk: Sbírka řešených příkladů z počtu pravděpodobnosti. Strana 23, 24, 42–47.

4 Cvičení 4 - Náhodná veličina

4.1 Teoretická část

Náhodná veličina (NV) je libovolná reálná funkce X definovaná na množině elementárních jevů ω pravděpodobnostního prostoru Ω.

Náhodná veličina je funkce $X : \Omega \to \mathbb{R}$ taková, že pro $B \subset \mathbb{R}$ platí $X^{-1}(B) \in \mathbb{A}$

Náhodnou veličinu značíme velkými písmeny: $X, Y, Z, ...$ nebo X_1, X_2 nebo vybranými písmeny řecké abecedy $\xi(ksi)$, η(eta), ζ(dzeta). Konkrétní realizaci náhodné proměnné značíme malými písmeny x, y, z nebo x_1, x_2 apod.

Příklad náhodné veličiny:
- počet zákazníků obsloužených v supermarketu za 1 hodinu prodeje
- počet členů domácnosti v souboru plzeňských domácností
- počet novorozenců v porodnici za 24 hodin

Distribuční funkce každému reálnému číslu přiřazuje pravděpodobnost, že náhodná veličina nabude hodnoty menší nebo rovné než toto číslo.

$F(x) = P\{X \leq x\}$ pro všechna $x \in \mathbb{R}$

Pomocí distribuční funkce je charakterizováno pravděpodobnostní chování náhodné veličiny.

Vlastnosti distribuční funkce:

- $0 \leq F(x) \leq 1$ (hodnoty distribuční funkce leží mezi 0 a 1)
- $F(x_1) \leq F(x_2)$ $\forall x_1 < x_2$ (je neklesající funkcí)
- $\lim\limits_{x \to -\infty} F(x) = 0$
- $\lim\limits_{x \to +\infty} F(x) = 1$
- $F(x)$ je zprava spojitá
- $F(x)$ má nejvýše spočetně bodů nespojitosti

Střední hodnota $E(X)$ je jednou z charakteristik polohy náhodné veličiny X.

Vlastnosti střední hodnoty: pro náhodné veličiny X a Y $(E(X), E(Y) < \infty)$ a $a, b \in \mathbb{R}$ platí

- $E(a) = a$
- $E(a \cdot X) = a \cdot E(X)$
- $E(X + Y) = E(X) + E(Y)$
- pro X, Y nezávislé: $E(X \cdot Y) = E(X) \cdot E(Y)$
- $P(X \geq 0) = 1 \Rightarrow E(X) \geq 0$

Rozptyl $D(X)$ je jednou z charakteristik variability náhodné veličiny X.

Rozptyl $D(X)$ (další označení $var(X)$ nebo $\sigma^2(X)$) je definován jako:

$$D(X) = E((X - E(X))^2)$$

Pro výpočet se používá výpočetní tvar rozptylu:

$$D(X) = E(X^2) - E^2(X)$$

Vlastnosti rozptylu: pro náhodné veličiny X a Y $(D(X),\, D(Y) < \infty)$ a $a,\, b \in \mathbb{R}$ platí

- $D(X) \geq 0$
- $D(a) = 0$
- $D(a \cdot X) = a^2 \cdot D(X)$
- $D(X + Y) = D(X) + D(Y)$
- $D(X - Y) = D(X) + D(Y)$

Směrodatná odchylka $\sigma(X) = \sqrt{D(X)}$

Nejčastěji se vyskytují následující dva případy náhodných veličin

- diskrétní případ (diskrétní náhodná veličina)
- spojitý případ (absolutně spojitá náhodná veličina)

Diskrétní náhodná veličina může nabývat konečně nebo spočetně hodnot $\{x_1, x_2, \ldots, x_n, \ldots\}$

Např.: Počet bodů získaných z jedné zápočtové práce z PSE je diskrétní náhodná veličina, která může nabývat hodnot 0, 1, 2, ... 20.

Pravděpodobnost toho, že náhodná veličina X bude mít hodnotu x_i s pravděpodobností p_i budeme zapisovat $P(X = x_i) = p_i$.

Musí platit $\sum\limits_i p_i = 1$.

Pravděpodobnostní funkce P je pro diskrétní náhodnou veličinu definována:

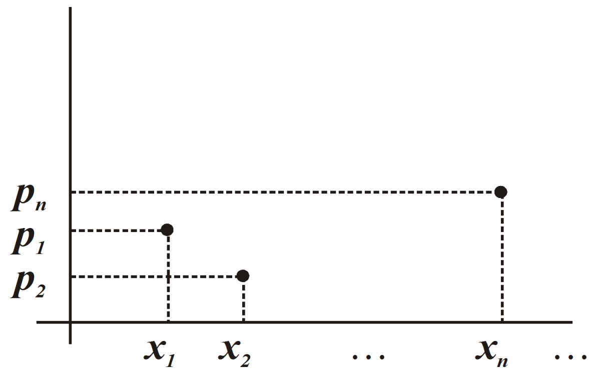
$$P(x) = \begin{cases} p_i & \text{pro } \; x = x_i \;\; i = 1, 2, \ldots \\ 0 & \text{pro } \; x \neq x_i \end{cases}$$

Distribuční funkce diskrétní náhodné veličiny je po částech konstantní:

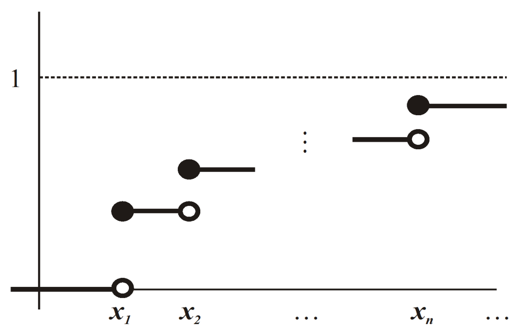
$$F(x) = \sum_{x_i \leq x} p_i$$

Vztahy pro diskrétní distribuční funkci (realizace náhodné veličiny $\{x_1, x_2, \ldots, x_n, \ldots\}$)

- definice $P(X \leq x_i) = F(x_i)$
- $P(X < x_i) = P(X \leq x_{i-1}) = F(x_{i-1})$
- $P(X \geq x_i) = 1 - P(X < x_i) = 1 - F(x_{i-1})$



Obrázek 5: Pravděpodobnostní funkce



Obrázek 6: Distribuční funkce

- $P(X > x_i) = 1 - P(X \leq x_i) = 1 - F(x_i)$
- $P(X = x_i) = P(X \leq x_i) - P(X < x_i) = F(x_i) - F(x_{i-1})$
- $P(x_a < X < x_b) = F(x_{b-1}) - F(x_a)$

Střední hodnota $E(X)$ se pro diskrétní náhodnou veličinu vypočte:

$$E(X) = \sum_i x_i \cdot p_i$$

Rozptyl $D(X)$ se pro diskrétní náhodnou veličinu vypočte:

$$D(X) = \sum_i x_i^2 \cdot p_i - E^2(X)$$

4.2 Příklady

1. Nakreslete pravděpodobnostní a distribuční funkci pro hod kostkou.

2. V dílně pracují 2 stroje (nezávisle na sobě). Pravděpodobnost, že se porouchá první stroj je 0.2. Pravděpodobnost, že se porouchá druhý stroj je 0.3. Náhodná veličina X bude označovat počet porouchaných strojů.

   (a) Určete pravděpodobnostní a distribuční funkci této náhodné veličiny.

   (b) Vypočtěte střední hodnotu a rozptyl této náhodné veličiny.

   Řešení: $\Omega = \{\omega_1, \omega_2, \omega_3\} = \{0, 1, 2\}$
   A...porouchá se první stroj ... $P(A) = 0.2$
   $\overline{A}$...neporouchá se první stroj ... $P(\overline{A}) = 1 - 0.2 = 0.8$
   B...porouchá se druhý stroj ... $P(B) = 0,3$
   $\overline{B}$...neporouchá se druhý stroj ... $P(\overline{B}) = 1 - 0.3 = 0.7$

   a) $P(0) = P(\overline{A}) \cdot P(\overline{B}) = 0.8 \cdot 0.7 = 0.56$
   $P(1) = P(A) \cdot P(\overline{B}) + P(\overline{A}) \cdot P(B) = 0.2 \cdot 0.7 + 0.8 \cdot 0.3 = 0.38$
   $P(2) = P(A \cap B) = P(A) \cdot P(B) = 0.06$

   $F(0) = P(x \leq 0) = 0.56$
   $F(1) = P(x \leq 1) = 0.94$
   $F(2) = P(x \leq 2) = 1$

   b) $E(X) = \omega_1 \cdot P(\omega_1) + \omega_2 \cdot P(\omega_2) + \omega_3 \cdot P(\omega_3) = 0 \cdot 0.56 + 1 \cdot 0.38 + 2 \cdot 0.06 = 0.5$
   $E(X^2) = \omega_1^2 \cdot P(\omega_1) + \omega_2^2 \cdot P(\omega_2) + \omega_3^2 \cdot P(\omega_3) = 0^2 \cdot 0.56 + 1^2 \cdot 0.38 + 2^2 \cdot 0.06 = 0.62$
   $D(X) = E(X^2) - E(X)^2 = 0.62 - 0.5^2 = 0.37$

3. V osudí je 5 kuliček - 2 bílé a 3 černé. Postupně jsou vytahovány kuličky (bez vracení zpět) dokud není vytáhnuta černá kulička.

   (a) Vypočtěte pravděpodobnostní funkci a distribuční funkci a nakreslete.

   (b) Vypočtěte střední hodnotu a rozptyl počtu tahů potřebných pro vytažení černé kuličky.

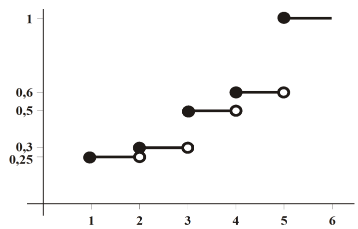
   Řešení: Náhodná veličina X je tah, ve kterém bude vytažena černá kulička, náhodná veličina nabývá hodnot $1, 2, 3$
   a) $P(1) = 3/5 = 0.6$
   $P(2) = (1 - 3/5) \cdot 3/4 = 0.3$
   $P(3) = (2/5) \cdot (1/4) \cdot 3/3 = 0.1$

   $F(0) = P(x \leq 0) = 0.6$
   $F(1) = P(x \leq 1) = 0.9$

$F(2) = P(x \leq 2) = 1$

b) $E(X) = \omega_1 \cdot P(\omega_1) + \omega_2 \cdot P(\omega_2) + \omega_3 \cdot P(\omega_3) = 1 \cdot 0.6 + 2 \cdot 0.3 + 3 \cdot 0.1 = 1,5$
$E(X^2) = \omega_1^2 \cdot P(\omega_1) + \omega_2^2 \cdot P(\omega_2) + \omega_3^2 \cdot P(\omega_3) = 1^2 \cdot 0.6 + 2^2 \cdot 0.3 + 3^2 \cdot 0.1 = 2.7$
$D(X) = E(X^2) - E(X)^2 = 2.7 - 1.5^2 = 0.45$

4. Z daného grafu distribuční funkce nakreslete graf pravděpodobnostní funkce.



Obrázek 7: Distribuční funkce (příklad 4)

5. Necht' náhodná veličina X nabývá hodnot 6,7,8,9 a 10 s pravděpodobností $P(6) = 0.67$, $P(7) = 0.05$, $P(8) = 0.06$, $P(9) = 0.16$ a $P(10) =???$

   (a) Spočtěte $P(10)$, $E(X)$ a $D(X)$.

   (b) Načrtněte graf pravděpodobnostní a distribuční náhodné veličiny X.

   Řešení: $P(10) = 0.06; E(X) = 6.89; D(X) = 1.898$

6. Náhodná veličina X nabývá hodnot $\{1, 2, 3\}$. Je dána funkce $P(X) = \dfrac{C}{X!}$. Určete hodnotu C tak, aby se jednalo o pravděpodobnostní funkci. Spočtěte $E(X), D(X)$.

   Řešení. $P(1) = \dfrac{C}{1!}$, $P(2) = \dfrac{C}{2!}$, $P(3) = \dfrac{C}{3!}$

   $C = 6/10$; $E(X) = 1.5$; $D(X) = 0.45$

4.3 Literatura s dalšími příklady

- Trial KMA – kapitola 30.5.4. Obecná diskrétní rozdělení
- Brousek, Jan – Ryjáček, Zdeněk: Sbírka řešených příkladů z počtu pravděpodobnosti. Strana 22–32.

5 Cvičení 5 - Alternativní, hypergeometrické a binomické rozdělení pravděpodobnosti.

5.1 Teoretická část

Alternativní rozdělení $A(p)$**:** $0 < p < 1$

Náhodná veličina X nabývá pouze dvou hodnot $X = 0$, pokud jev nenastal a $X = 1$ pokud jev nastal. p tedy označuje pravděpodobnost toho, že jev nastane.

Pravděpodobnostní funkce: $P(x) = \begin{cases} 1-p & \text{pro } \; x = 0 \\ p & \text{pro } \; x = 1 \\ 0 & \text{jinde} \end{cases}$

Distribuční funkce: $F(x) = \begin{cases} 0 & \text{pro } \; x < 0 \\ 1-p & \text{pro } \; 0 \leq x < 1 \\ 1 & \text{pro } \; x \geq 1 \end{cases}$

Střední hodnota: $E(X) = 0 \cdot (1-p) + 1 \cdot p = p$

Rozptyl: $D(X) = 0^2 \cdot (1-p) + 1^2 \cdot p - p^2 = p\,(1-p)$

Použití: teoretický základ pro další typy rozdělení

Binomické rozdělení $Bi(p; n)$**:** $0 < p < 1;\, n \in \mathbf{N}$

Binomické rozdělení popisuje četnost výskytu náhodného jevu v n nezávislých pokusech, v nichž má jev stále stejnou pravděpodobnost. Náhodná veličina X nabývá hodnot $\{0\,;\,1\,;\,2\,;\ldots,\,n\}$.

Pravděpodobnostní funkce: $P(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad$ pro $x = 0, 1, \ldots, n$

Distribuční funkce: $F(x) = \begin{cases} 0 & \text{pro } \; x < 0 \\ \displaystyle\sum_{t=0}^{x} \binom{n}{t} p^t (1-p)^{n-t} & \text{pro } \; 0 \leq x < n \\ 1 & \text{pro } \; x \geq n \end{cases}$

Střední hodnota: $E(X) = n \cdot p$

Rozptyl: $D(X) = n \cdot p\,(1-p)$

Použití:

- pravděpodobnost úspěchu v jednom pokusu je p, náhodná veličina $X \sim Bi(p, n)$ charakterizuje počet úspěšných pokusů při n nezávislých opakováních
- podíl výrobků z danou vlastností v základním souboru je p, náhodná veličina $X \sim Bi(p, n)$ charakterizuje počet výrobků s danou vlastností ve výběru rozsahu n, pokud prvky po výběru vracíme zpět

Poznámky:

- alternativní rozdělení je $Bi(p; n = 1)$

- $A_i \sim A(p)$, $i = 1, 2, \ldots, n$ nezávislé jevy $X = \sum_{i=1}^{n} A_i \sim Bi(p; n)$. Binomické rozdělení lze chápat jako rozdělení součtu n vzájemně nezávislých náhodných veličin, řídících se týmž alternativním rozdělením.

Hypergeometrické rozdělení $H(N; M; n)$**:** $0 < M < N\,;\ 0 < n \leq N\,;\ n, N, M \in \mathbf{N}$

Hypergeometrické rozdělení je rozdělení náhodné veličiny, kdy při opakování náhodného pokusu je výskyt sledovaného jevu závislý na výsledcích předcházejících pokusů. Jde tedy o pokusy, které jsou na sobě závislé. Typickým představitelem je výběr prvků bez vracení. V takovém případě můžeme N považovat za celkový počet prvků souboru a M za počet prvků souboru, které mají sledovanou vlastnost. Počet prvků vybraných z tohoto souboru bez vracení je pak n.

Náhodná veličina X nabývá hodnot: $\{0\,;\,1\,;\,2\,;\ldots\min(n, M)\}$

Pravděpodobnostní funkce: $P(x) = \dfrac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$ pro $x = 0, 1, \ldots, \min(n, M)$

Střední hodnota: $E(X) = n\,\dfrac{M}{N}$

Rozptyl: $D(X) = n\,\dfrac{M}{N}\left(1 - \dfrac{M}{N}\right)\dfrac{N-n}{N-1}$

Poznámka:

- Binomické rozdělení je limitním případem hypergeometrického rozdělení pro $n \to \infty$ a $\dfrac{M}{N} \to 0$ je $HG(N; M; n) \approx Bi(p = \dfrac{M}{N}; n)$

Použití:

- v souboru N prvků má M prvků sledovanou vlastnost, provedeme výběr n prvků, přičemž vybraný prvek do souboru nevracíme, náhodná veličina $X \sim HG(N; M; n)$ charakterizuje počet prvků se sledovanou vlastností ve výběru n
- pokud značíme $p = \dfrac{M}{N}$ procento prvků se sledovanou vlastností v souboru a dále uvažujme, že vybíraných prvků je „tak moc“, že nezáleží na tom, zda po výběru prvek vracíme nebo nevracíme, pak $HG(N; M; n) \approx Bi(p = \dfrac{M}{N}; n)$, stačí $n \geq 30$ a $p \leq 0.1$

5.2 Příklady

1. U hodu jednou kostkou sledujeme, zda padla 6, tj. X=1 v případě, že na kostce padla 6 a X=0 v případě, že na kostce nepadla 6.
   (a) Nakreslete pravděpodobnostní a distribuční funkci náhodné veličiny X.
   (b) Vypočtěte střední hodnotu a rozptyl náhodné veličiny X.

   Řešení: $E(X) = \frac{1}{6}$; $D(X) = \frac{1}{6} \cdot \frac{5}{6} = \frac{5}{36}$

2. U hodu třemi kostkami sledujeme, kolikrát padla 6.
   (a) Nakreslete pravděpodobnostní a distribuční funkci náhodné veličiny.
   (b) Vypočtěte střední hodnotu a rozptyl náhodné veličiny.

   Řešení: $n$krát opakujeme pokus, při každém pokusu je pravděpodobnost sledovaného jevu stejná. Použijeme tedy binomické rozdělení. Náhodná veličina $X \sim Bi(n,p), n = 3, p = \frac{1}{6}$ udává počet hodů, při kterých padla na kostce 6.

   (a) $P(0) = \binom{3}{0}\frac{1}{6}^0(1-\frac{1}{6})^3 \doteq 0.578$; $P(1) = \binom{3}{1}\frac{1}{6}^1(1-\frac{1}{6})^2 \doteq 0.347$
   $P(2) = \binom{3}{2}\frac{1}{6}^2(1-\frac{1}{6})^1 \doteq 0.069$; $P(3) = \binom{3}{3}\frac{1}{6}^3(1-\frac{1}{6})^0 \doteq 0.0046$

   $F(0) = P(0) \doteq 0.578$; $F(1) = P(0) + P(1) \doteq 0.925$;
   $F(2) \doteq 0.9954$; $F(3) = 1$

   (b) $E(X) = n \cdot p = 3 \cdot \frac{3}{5} = 0.6$; $D(X) = n \cdot p \cdot (1-p) = 3 \cdot \frac{1}{6} \cdot \frac{5}{6} \doteq 0.4167$

3. Mezi 10 výrobky jsou 3 vadné. Postupně vybereme dva výrobky (s vracením zpět). Náhodná veličina A označuje počet vadných výrobků mezi vybranými výrobky.
   (a) Nakreslete pravděpodobnostní a distribuční funkci náhodné veličiny.
   (b) Vypočtěte střední hodnotu a rozptyl náhodné veličiny.

   Řešení: $n$krát opakujeme pokus, při každém pokusu je pravděpodobnost sledovaného jevu stejná. Použijeme tedy binomické rozdělení. Náhodná veličina $X \sim Bi(n,p), n = 2, p = \frac{3}{10}$ udává počet hodů, při kterých padla na kostce 6.

   (a) $P(0) = \binom{2}{0}\frac{3}{10}^0(1-\frac{3}{10})^2 \doteq 0.49$;
   $P(1) = \binom{2}{1}\frac{3}{10}^1(1-\frac{3}{10})^1 \doteq 0.42$;
   $P(2) = \binom{2}{2}\frac{3}{10}^2(1-\frac{3}{10})^0 \doteq 0.09$;

$F(0) = P(0) \doteq 0.49$; $F(1) = 0.91$; $F(2) = 1$

(b) $E(X) = n \cdot p = 3 \cdot \dfrac{6}{10} = 0.6$; $D(X) = n \cdot p \cdot (1-p) = 2 \cdot \dfrac{3}{10} \cdot \dfrac{7}{10} \doteq 0.42$

4. Při akci "Kryštof"policisté na jednom stanovišti zkontrolovali 500 vozidel. Víme, že 1 vozidlo z 200 je kradené. Jaká je pravděpodobnost, že policisté:
   (a) našli 2 kradená vozidla;
   (b) našli méně než 5 kradených vozidel;
   (c) našli alespoň 3 a méně než 5 kradených vozidel.

   Řešení: $n$krát opakujeme pokus, při každém pokusu je pravděpodobnost sledovaného jevu stejná. Použijeme tedy binomické rozdělení. Náhodná veličina $X \sim Bi(n,p), n = 500, p = \dfrac{1}{200} = 0.005$ udává počet kradených vozidel, která policisté při kontrolách naleznou.

   (a) $P(X = 2) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{500}{2} 0.005^2 (1-0.005)^{500-2} = 0.257$
   (b) $P(X < 5) = P(0) + P(1) + ... + P(4) = 0.082 + 0.205 + 0.257 + 0.214 + 0.134 = 0.892$
   (c) $P(3 \leq X < 5) = P(3) + P(4) = 0.214 + 0.134 = 0.348$

5. Mezi 10 výrobky jsou 3 vadné. Postupně vybereme dva výrobky (bez vracení zpět). Náhodná veličina A označuje počet vadných výrobků mezi vybranými výrobky.
   (a) Nakreslete pravděpodobnostní a distribuční funkci náhodné veličiny.
   (b) Vypočtěte střední hodnotu a rozptyl náhodné veličiny.

   Řešení: výrobky zpět do výběru nevracíme, jednotlivé tahy jsou závislé. Použijeme hypergeometrické rozdělení. Náhodná veličina $X \sim H(N; M; n)$; $N = 10, M = 3, n = 2$.

   (a) $P(0) = \dfrac{\binom{3}{0}\binom{7}{2}}{\binom{10}{2}} = \dfrac{21}{45} \doteq 0.467$;
   $P(1) = \dfrac{\binom{3}{1}\binom{7}{1}}{\binom{10}{2}} = \dfrac{21}{45} \doteq 0.467$;
   $P(2) = \dfrac{\binom{3}{2}\binom{7}{0}}{\binom{10}{2}} = \dfrac{3}{45} \doteq 0.067$
   $F(0) \doteq 0.467$; $F(1) = 0.933$; $F(2) = 1$

   (b) $E(X) = n\dfrac{M}{N} = 2 \cdot \dfrac{3}{10} = 0.6$;
   $D(X) = n\,\dfrac{M}{N}\left(1 - \dfrac{M}{N}\right)\dfrac{N-n}{N-1} = 2\,\dfrac{3}{10}\left(1 - \dfrac{3}{10}\right)\dfrac{10-2}{10-1} = \dfrac{56}{150} = 0.3733$

6. Nechť náhodná veličina X pochází z hypergeometrického rozdělení, $X \sim H(N, K, n) = H(10, 4, 4)$. Spočtěte pravděpodobnosti:

(a) $P(X = 3)$
(b) $P(X > 0)$

Řešení: Protože $k \geq 0, k \leq K, k \leq n, n - k \leq N - K$, je $k \in \{0, ..., 4\}$

(a) $P(X = 3) = \frac{\binom{4}{3}\binom{6}{1}}{\binom{10}{4}} \doteq 0.114$
(b) $P(X > 0) = 1 - P(X = 0) \doteq 0.929$

7. Spočtěte pravděpodobnosti výhry ve Sportce (celkový počet čísel v osudí je 49, je vytaženo 6 čísel a následně jedno dodatkové):
   (a) Pro 1. pořadí, tj. ze 6 vybraných jsme tipnuli 6.
   (b) Pro 2. pořadí, tj. ze 6 vybraných jsme tipnuli 5 a zároveň dodatkové číslo.
   (c) Pro 3. pořadí, tj. ze 6 vybraných jsme tipnuli 5.
   (d) Pro 4. pořadí, tj. ze 6 vybraných jsme tipnuli 4.
   (e) Pro 5. pořadí, tj. ze 6 vybraných jsme tipnuli 3.
   (f) Ze 6 vybraných jsme tipnuli 2.
   (g) Ze 6 vybraných jsme tipnuli 1.
   (h) Ze 6 vybraných jsme netipnuli ani jedno číslo.

   Řešení: pro losování sportky je typické, že se losovaná čísla nevrací zpět, proto použijeme hypergeometrické rozdělení. $X \sim H(N; M; n); N = 49, M = 6, n = 6$

   (a) $P(1.\text{pořadí}) = P(X = 6) = \frac{\binom{6}{6}\binom{43}{0}}{\binom{49}{6}} = 7.15 \cdot 10^{-8}$
   (b) $P(X = 5) = \frac{\binom{6}{5}\binom{43}{1}}{\binom{49}{6}}$
   $P(2.\text{pořadí}) = P(X = 5) \cdot \frac{1}{43} = 4.29 \cdot 10^{-7}$ (uhádli jsme dodatkové)
   (c) $P(3.\text{pořadí}) = P(X = 5) \cdot \frac{42}{43} = 1.8 \cdot 10^{-5}$ (neuhádli jsme dodatkové)
   (d) $P(4.\text{pořadí}) = P(X = 4) = 9.69 \cdot 10^{-4}$
   (e) $P(5.\text{pořadí}) = P(X = 3) = 1.77 \cdot 10^{-2}$
   (f) $P(X = 2) = 0.132$
   (g) $P(X = 1) = 0.413$
   (h) $P(X = 1) = 0.436$

8. Nechť náhodná veličina X pochází z hypergeometrického rozdělení, $X \sim H(N, K, n) = H(30, 14, 3)$. Spočtěte pravděpodobnosti:
   (a) $P(X = 2)$

(b) $P(X > 1)$

Řešení: Protože $k \geq 0, k \leq K, k \leq n, n-k \leq N-K$, je $k \in \{0, ..., 3\}$

(a) $P(X = 2) = \dfrac{\binom{14}{2}\binom{16}{1}}{\binom{30}{3}} = \dfrac{91 \cdot 16}{4060} \doteq 0.359$

(b) $P(X > 1) = P(X = 2) + P(X = 3) = 0.359 + 0.090 = 0.448$

5.3 Literatura s dalšími příklady

- Trial KMA – kapitola 30.5.1. Hypergeometrické rozdělení
- Trial KMA – kapitola 30.5.2. Binomické rozdělení
- Brousek, Jan – Ryjáček, Zdeněk: Sbírka řešených příkladů z počtu pravděpodobnosti. Strana 22–32.

6 Cvičení 6 - Poissonovo rozdělení

6.1 Teoretická část

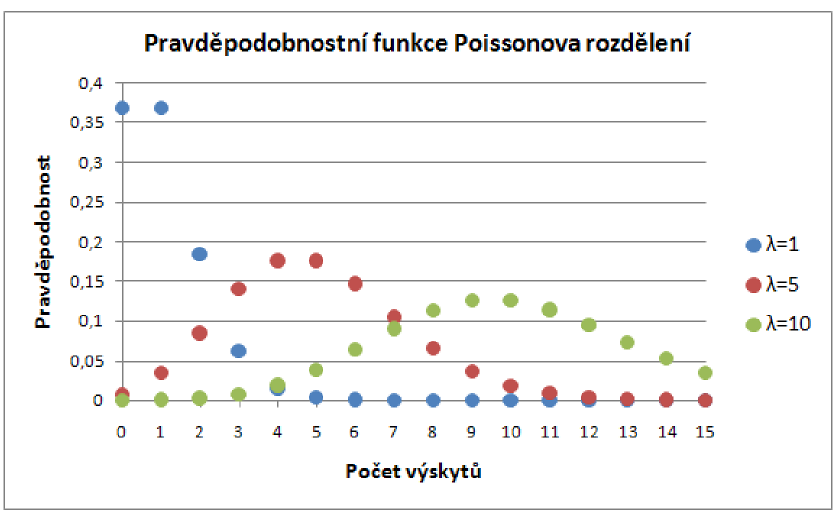
6.1.1 Poissonovo rozdělení

U tohoto rozdělení nabývá náhodná veličina X hodnot $\{0, 1, 2, 3, \dots\}$. Pokud je počet výskytů nějaké události během časového intervalu přímo úměrný délce časového intervalu a průměrný počet výskytů události za konstantní časovou jednotku je λ ($\lambda > 0$), potom náhodná veličina $X \sim Po(\lambda)$ charakterizuje počet výskytů události za konstantní časovou jednotku.
Pokud se počet výskytů v jednotkovém časovém intervalu řídí Poissonovo rozdělením s parametrem λ, tj. $Po(\lambda)$, potom se počet výskytů v časovém intervalu o délce t řídí Poissonovo rozdělením s parametrem λt, tj. $Po(\lambda t)$.
Pokud počet výskytů nějaké jednotky v dané oblasti je přímo úměrný velikosti oblasti a průměrný počet výskytů události v konstantní oblasti je λ ($\lambda > 0$), potom náhodná veličina $X \sim Po(\lambda)$ charakterizuje počet výskytů jednotky v konstantní oblasti.
Pokud se počet výskytů v oblasti o jednotkovém obsahu (objemu) řídí Poissonovo rozdělením s parametrem λ, tj. $Po(\lambda)$, potom se počet výskytů v oblasti o obsahu (objemu) S řídí Poissonovo rozdělením s parametrem λS, tj. $Po(\lambda S)$.

Lze jím popisovat např.: počet telefonátů v call centru, počet přístupů na server, počet autonehod, počet výskytů vzácných nemocí (např. leukemie), počet hvězd v dané oblasti vesmíru...

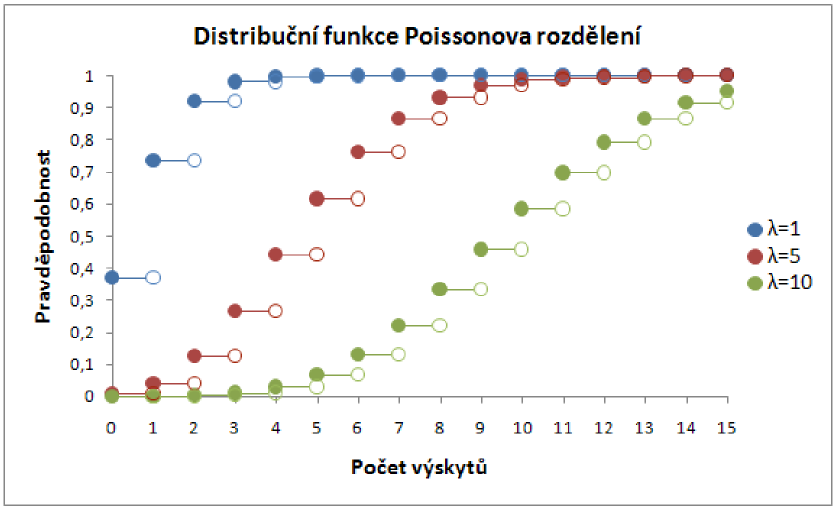
Pravděpobobnostní funkce

$$P(x) = e^{-\lambda}\frac{\lambda^x}{x!} \text{ pro } x = 0, 1, 2, \dots$$



Distribuční funkce

$$F(x) = \sum_{k=0}^{x} e^{-\lambda}\frac{\lambda^k}{k!} \text{ pro } x = 0, 1, 2, \dots$$



Hodnoty této funkce jsou tabelovány. Ukázka tabulek (udávají hodnoty distribuční funkce):

x	λ						
	0.1	0.2	0.5	1.0	1.5	2.0	3.0
0	0.9048	0.8187	0.6065	0.3679	0.2231	0.1353	0.0498
1	0.9953	0.9825	0.9098	0.7358	0.5578	0.4060	0.1991
2	0.9998	0.9989	0.9856	0.9197	0.8088	0.6767	0.4232
3	1.0000	0.9999	0.9982	0.9810	0.9344	0.8571	0.6472
4	1.0000	1.0000	0.9998	0.9963	0.9814	0.9473	0.8153
5	1.0000	1.0000	1.0000	0.9994	0.9955	0.9834	0.9161
6	1.0000	1.0000	1.0000	0.9999	0.9991	0.9955	0.9665
7	1.0000	1.0000	1.0000	1.0000	0.9998	0.9989	0.9881
8	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9962
9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9989
10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997
11	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999
12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Například pravděpodobnost toho, že náhodná veličina $X \sim Po(2)$ nabude hodnoty 6 a méně (tj. $P(X \leq 6) = F(6)$ je rovna 0.9955. Pokud chceme z tabulek zjistit hodnotu pravděpodobnostní funkce pro $X = x$, stačí od distribuční funkce pro hodnotu x odečíst distribuční funkci pro hodnotu $x - 1$, tj.

$$P(X = x) = P(X \leq x) - P(X \leq x - 1) = F(x) - F(x - 1)$$

Např. chceme-li zjistit pravděpodobnost toho, že náhodná veličina $X \sim Po(3)$ nabude přesně hodnoty 5

$$P(X = 5) = P(X \leq 5) - P(X \leq 4) = F(5) - F(4) = 0.9161 - 0.8153 = 0.1008$$

Střední hodnota

$$E(X) = \lambda$$

Rozptyl

$$D(X) = \lambda$$

Excel
V Excelu lze Poissonovo rozdělení nalézt pod **POISSON**. Zápis je následující:

$$= POISSON(x, \lambda, L) =$$

$$= POISSON(\text{nastane } x\text{,očekávaná hodnota } \lambda\text{,logická hodnota})$$

- Pokud je logická hodnota **L** rovna 0 (nepravda), potom je vrácena pravděpodobnostní funkce.
- Pokud je logická hodnota **L** rovna 1 (pravda), potom je vrácena distribuční funkce.

6.1.2 Aproximace binomického rozdělení Poissonovo rozdělením

Poissonovo rozdělení je limitním případem binomického rozdělení, tj. pro $n \to \infty$ a $p \to 0$ je $Bi(p, n) \approx Po(\lambda = np)$. Tato aproximace funguje dobře již pro hodnoty $n \geq 30$ a $p \leq 0.1$.

6.2 Příklady

1. Telefonní ústředna spojí průměrně 3 hovory za půl hodiny. Jaká je pravděpodobnost, že telefonní ústředna za půl hodiny:

   (a) spojí 3 hovory,

   (b) spojí 3 a více hovorů,

   (c) spojí více než 5 a méně než 10 hovorů.

   (d) Spojí za **hodinu** méně než 5 hovorů.

   Řešení:
   Jev X bude označovat počet spojených hovorů.

   (a) Chceme znát pravděpodobnost, že náhodná veličina nabude hodnoty 3, tj. $P(X = 3)$. Chceme tedy znát hodnotu pravděpodobnostní funkce pro hodnotu 3. K tomu můžeme použít tabulky pro distribuční funkci Poissonova rozdělení. Proto, abychom mohli použít tabulky, musíme zapsat $P(X = 3)$ jako

   $$P(X = 3) = P(X \leq 3) - P(X \leq 2) = F(3) - F(2)$$

   Nyní dosadíme:

   $$P(X = 3) = 0.6472 - 0.4232 = 0.224$$

   (b) $P(X \geq 3) = 1 - P(X < 3)$ protože se jedná o diskrétní náhodnou veličinu (pro tu platí $P(X < x) = P(X \leq x-1)$) tak je tento výraz roven $1-P(X \leq 2) = 1-F(2)$. Po dosazení:

   $$P(X \geq 3) = 1 - 0.4232 = 0.5768$$

   (c) $P(5 < X < 10) = P(5 < X \leq 9) = F(9) - F(5) = 0.9989 - 0.9161 = 0.0828$

   (d) Nyní chceme znát pravděpodobnost toho, že bude spojeno méně než 5 hovorů za hodinu. Víme, že pokud se počet výskytů náhodné veličiny X za časovou jednotku (pro nás je časovou jednotkou 1/2 hodiny) řídí rozdělením $Po(\lambda = 3)$, potom se počet výskytů náhodné veličiny X za dvě časové jednotky (tj. za hodinu) řídí rozdělením $Po(\lambda = 6)$.

   $$P(X < 5) = P(X \leq 4) = F(4) = 0.2851$$

2. Průměrná zmetkovitost výrobků je 1 %. Vybereme náhodně 100 výrobků (s vracením zpět).

   (a) S jakou pravděpodobností mezi nimi budou nejvýše dva vadné výrobky?

   (b) Vypočtěte totéž s použitím aproximace pomocí Poissonova rozdělení.

   Řešení:
   Jev X bude označovat počet vadných výrobků (zmetků) mezi 100 vybranými výrobky.

(a) Vzhledem k tomu, že se jedná o výběr s vracením zpět, tak $X \approx Bi(100, 0.01)$. Nejvýše dva vadné znamená $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$.
$P(X = 0) = \binom{100}{0} 0.01^0 (1 - 0.01)^{100} = 0.366$
$P(X = 1) = \binom{100}{1} 0.01^1 (1 - 0.01)^{99} = 0.370$
$P(X = 2) = \binom{100}{2} 0.01^2 (1 - 0.01)^{98} = 0.183$
Dohromady tedy $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = 0.919$

(b) Víme, že pro $n \to \infty$ a $p \to 0$ je $Bi(p, n) \approx Po(\lambda = np)$, a že tato aproximace funguje dobře již pro hodnoty $n \geq 30$ a $p \leq 0.1$. Obě tyto podmínky jsou v tomto příkladu splňeny, a proto můžeme tuto aproximaci použít:

$$Bi(100, 0.01) \approx Po(\lambda = 1)$$

Pravděpodobnost, že náhodná veličina X nabude hodnoty nejvýše 2 je pak z tabulek $P(X \leq 2) = F(2) = 0.9197$.

3. Na server přijde během hodiny průměrně 120 požadavků. Jaká je pravděpodobnost, že během 2 minut, po které je server restartován:

   (a) nepřijde žádný požadavek,

   (b) přijdou více jak 3 požadavky,

   (c) přijdou více jak 3 požadavky, ale méně než 7 požadavků.

   Řešení [(a) 0.0183; (b) 0.5665; (c) 0.4558]

4. Předpokládejme, že ústředna spojí v průměru 1 hovor během 100 sekund. Určete pravděpodobnost, že operátor zmešká nanejvýš jeden hovor, pokud si vezme 5minutovou přestávku na kávu.
   Řešení [0.9502]

6.3 Literatura s dalšími příklady

- Trial KMA – kapitola 30.5.2. Binomické rozdělení
- Trial KMA – kapitola 30.5.3. Poissonovo rozdělení
- Brousek, Jan – Ryjáček, Zdeněk: Sbírka řešených příkladů z počtu pravděpodobnosti. Strany 62, 63, 82 a 83.

7 Cvičení 7 - Rozdělení spojitého typu

7.1 Teoretická část

7.1.1 Spojitá náhodná veličina

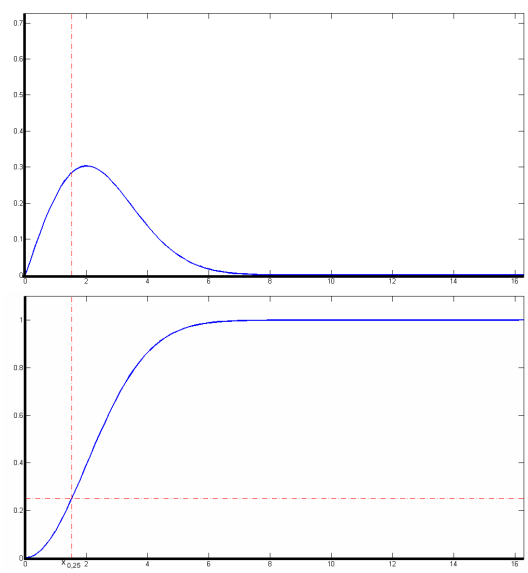
Náhodná veličina X s *distribuční funkcí* F má spojité rozdělení, existuje-li funkce $f(x)$ taková, že

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\ \mathtt{d}t, \qquad x \in \mathbf{R}.$$

Funkci f, pro kterou platí

$$f(x) \geq 0, \quad x \in \mathbf{R} \qquad a \qquad \int_{-\infty}^{+\infty} f(x)\ \mathtt{d}x = 1$$

nazýváme *hustotou* spojitého rozdělení s distribuční funkcí F. Platí $f(x) = F'(x)$ skoro všude.



Obrázek 8: Funkce hustoty, distribuční funkce a dolni kvartil spojité náhodné veličiny

Pro náhodnou spojitou veličinu s hustotou pravděpodobnosti $f(x)$ a distribuční funkcí $F(x)$ platí

- $P(X = x) = 0 \quad x \in \mathbf{R}$ (plyne ze spojitosti distribuční funkce)
- $P(X < x) = P(X \leq x) = F(x) = \int_{-\infty}^{x} f(x)\ \mathtt{d}x \quad x \in \mathbf{R}$
- $P(X > x) = P(X \geq x) = 1 - P(X < x) = 1 - P(X \leq x) = 1 - F(x) =$
  $= \int_{x}^{+\infty} f(x)\ \mathtt{d}x \quad x \in \mathbf{R}$
- $P(a < X < b) = P(a \leq X \leq b) = F(b) - F(a) = \int_{a}^{b} f(x)\ \mathtt{d}x$ pro $a < b$

7.1.2 Charakteristky spojité náhodné veličiny

Střední hodnota

$$EX = \int_{\mathbf{R}} x \cdot f(x)\ \mathtt{d}x$$

Rozptyl

$$DX = E(X - E(X))^2,$$

$$\text{výpočtový tvar:}\quad DX = E\left(X^2\right) - (EX)^2 = \int_{\mathbf{R}} x^2 \cdot f(x)\ \mathtt{d}x - (EX)^2$$

$\alpha\%$ **kvantil** – hodnota x_α náhodné veličiny X, která splňuje podmínku

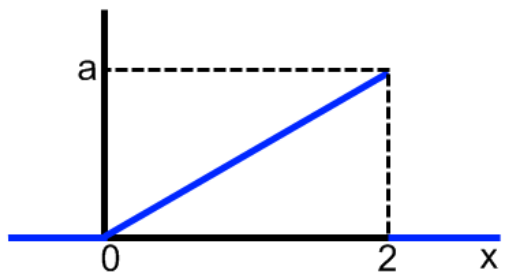
$$F(x_\alpha) = \alpha \quad \text{pro } 0 < \alpha < 1$$

Speciální označení

- medián – $x_{0.5}$
- dolní kvartil – $x_{0.25}$
- horní kvartil – $x_{0.75}$

7.2 Příklady

1. Určete konstantu a, distribuční funkci, střední hodnotu a rozptyl náhodné veličiny s následující funkcí hustoty.



Řešení: Z geometrického vyjádření vlastnosti $\int_{\mathbf{R}} f(x)\, \mathtt{d}x = 1$ funkce hustoty plyne podmínka, že plocha pod jejím grafem musí být jednotková. Pro danou funkci tedy plocha trojúhelníka $\frac{1}{2} \cdot a \cdot 2 = 1 \Rightarrow a = 1$ Body $[0;0]$ a $[2;1]$ určují lineární funkci hustoty pravděpodobnosti.

$$f(x) = \begin{cases} \frac{x}{2} & \text{pro } x \in \langle 0, 2 \rangle, \\ 0 & \text{jinde.} \end{cases}$$

$$F(x) = \begin{cases} \int_{-\infty}^{x} f(t)\, \mathtt{d}t = \int_0^x \frac{t}{2} \mathtt{d}t = \frac{x^2}{4} & \text{pro } x \in \langle 0, 2 \rangle, \\ 0 & \text{jinde.} \end{cases}$$

$$EX = \int_0^2 x \frac{x}{2} \mathtt{d}x = \frac{4}{3} \qquad DX = \int_0^2 x^2 \frac{x}{2} \mathtt{d}x - \left(\frac{4}{3}\right)^2 = \frac{2}{9}$$

2. Mějme náhodnou veličinu X s distribuční funkcí

$$F(x) = \begin{cases} 0 & x < 6 \\ -0.01x^2 + 0.32x - 1.56 & x \in \langle 6; 16 \rangle \\ 1 & x > 16 \end{cases}$$

Spočtěte

(a) $P(X = 12.3)$;

(b) $P(X > 6.58)$;

(c) $P(7.14 \leq X < 14.91)$;

(d) 95% kvantil $x_{0.95}$.

(e) Odvoďte hustotu náhodné veličiny X, nakreslete její graf a znázorněte v něm $P(7.14 \leq X \leq 14.91)$ a $x_{0.95}$.

Řešení:

(a) $P(X = 12.3) = 0$ – jedná se o spojitou náhodnou veličinu

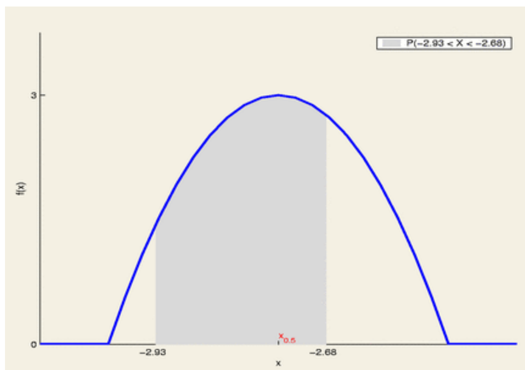
(b) $P(X > 6.58) = 1 - P(X \leq 6.58) = 1 - F(6.58) =$
$= 1 - \left(-0.01 \cdot 6.58^2 + 0.32 \cdot 6.58 - 1.56\right) = 0.877$

(c) $P(7.14 \leq X < 14.91) = F(14.91) - F(7.14) = 0.988 - 0.215 = 0.773$

(d) Z distribuční funkce F odvodíme kvantil $x_{0.95}$:

$$\begin{aligned} F(x_{0.95}) &= 0.95 \\ x_{0,95} &= -\sqrt{\frac{0.95}{-0.01} + 100} + 16 \\ x_{0,95} &= 13.764 \end{aligned}$$

(e) $f(x) = F'(x) = \begin{cases} -0{,}02x + 0{,}32 & x \in \langle -6; 16 \rangle \\ 0 & \text{jinak.} \end{cases}$



3. Nechť je dána náhodná veličina s hustotou pravděpodobnosti

$$f(x) = \begin{cases} 0 & x < -1 \\ x + 1 & x \in \langle -1, 0\ ) \\ -x + 1 & x \in \langle 0, 1\ ) \\ 0 & x \geq 1 \end{cases}$$

Spočtěte její střední hodnotu a rozptyl, určete distribuční funkci a najděte $c \in \mathbf{R}$ tak, aby $P(X \leq c) = 0.95$.

Řešení:

$$\begin{aligned} EX &= \int_{-1}^{0} x \cdot (x+1)\ \mathrm{d}x + \int_{0}^{1} x \cdot (-x+1)\ \mathrm{d}x = 0 \\ DX &= \int_{-1}^{0} x^2 \cdot (x+1)\ \mathrm{d}x + \int_{0}^{1} x^2 \cdot (-x+1)\ \mathrm{d}x - (EX)^2 = \frac{1}{6} \end{aligned}$$

$F(x) = \int_{-\infty}^{x} f(t)\ \mathtt{d}t =$

$$= \begin{cases} 0 & x < -1 \\ \displaystyle\int_{-1}^{x} (t+1)\ \mathtt{d}t = \frac{x^2}{2} + x & x \in \langle -1, 0\ ) \\ \displaystyle\int_{-1}^{0} (t+1)\ \mathtt{d}t + \int_{0}^{x} (-t+1)\ \mathtt{d}t = \frac{1}{2} - \frac{x^2}{2} + x & x \in \langle 0, 1\ ) \\ 1 & x \geq 1 \end{cases}$$

$$\begin{aligned} P(X \leq c) = F(c) = \int_{-\infty}^{c} f(x)\ \mathtt{d}x &= 0{,}95 \\ \int_{-1}^{0} (x+1)\ \mathtt{d}x + \int_{0}^{c} (-x+1)\ \mathtt{d}x &= 0{,}95 \quad c > 0 \\ c &= 0{,}6838 \end{aligned}$$

4. Je dána funkce

$$f(x) = \begin{cases} a \cdot x^2 & \text{pro } x \in \langle 0, 1\rangle, \\ 0 & \text{jinde.} \end{cases}$$

   (a) Určete hodnotu konstanty a, tak aby tato funkce byla hustotou pravděpodobnosti nějaké náhodné veličiny.
   (b) Vypočtěte EX, DX a $F(x)$.
   (c) Určete pravděpodobnost, že náhodná veličina X bude mezi 0.5 a 0.75.

   Řešení:

   (a) $\displaystyle\int_0^1 a \cdot x^2 \mathtt{d}x = 1 \Rightarrow a = 3$
   (b)

$$\begin{aligned} EX &= \int_0^1 x \cdot 3x^2 \mathtt{d}x = \left[3\frac{x^4}{4}\right]_0^1 = 0.75 \\ DX &= \int_0^1 x^2 \cdot 3x^2 \mathtt{d}x - 0.75^2 = \left[3\frac{x^5}{5}\right]_0^1 - 0.5625 = 0.6 - 0.5625 = 0.0375 \\ F(x) &= \begin{cases} \displaystyle\int_0^x 3t^2 dt = x^3 & x \in \langle 0; 1\rangle \\ 0 & \text{jinak.} \end{cases} \end{aligned}$$

   (c) $P(0.5 < X < 0.75) = F(0.75) - F(0.5) = 0.75^3 - 0.5^3 \doteq 0.297$

7.3 Literatura s dalšími příklady

- Trial KMA – kapitola 30.6.4. Obecné spojité rozdělení

8 Cvičení 8 - Rovnoměrné rozdělení. Exponenciální rozdělení.

8.1 Teoretická část

8.1.1 Rovnoměrné rozdělení

$$\mathrm{R}\,(a;b) \qquad a,b \in \mathbf{R}, \quad a<b$$

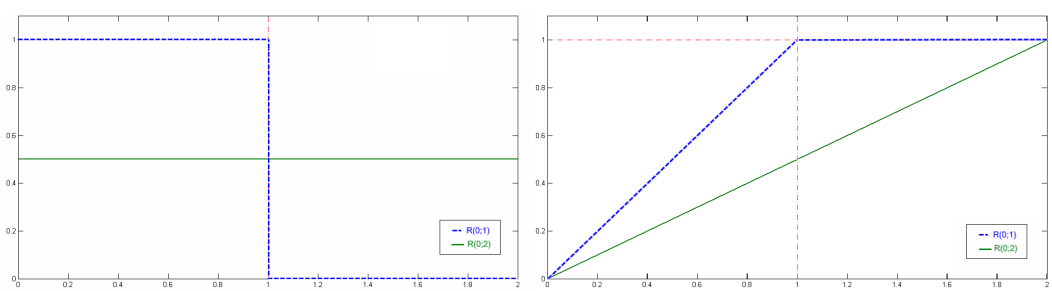
Náhodná veličina X může nabýt libovolné reálné hodnoty x z intervalu $(a;b)$ a její výskyt na celém intervalu $(a;b)$ je stejně možný. Pak X má rovnoměrné rozdělení na intervalu $(a;b)$ a plocha pod křivkou hustoty tvoří obdélník, jehož plocha je rovna 1. To znamená, že X jistě nabude hodnoty z intervalu $(a;b)$. Jelikož šířka tohoto intervalu je $(b-a)$, výška hustoty musí být rovna $\dfrac{1}{b-a}$ (neboť integrál přes hustotu dá 1).

Funkce hustoty $f\,(x) = \begin{cases} \dfrac{1}{b-a} & \text{pro } x \in \langle a;b\rangle \\ 0 & \text{jinde} \end{cases}$

Distribuční funkce $F\,(x) = \begin{cases} 0 & \text{pro } x \leq a \\ \dfrac{x-a}{b-a} & \text{pro } x \in (a;b) \\ 1 & \text{pro } x \geq b \end{cases}$

Střední hodnota a rozptyl

$$E(X) = \frac{a+b}{12} \qquad\qquad D(X) = \frac{(a-b)^2}{12}$$



Obrázek 9: Funkce hustoty a distribuční funkce rovnoměrného rozdělení

Použití – chyby při zaokrouhlování v numerických výpočtech

- výchozí rozdělení při simulaci náhodných veličin na počítači, ostatní náhodné veličiny lze získat pomocí různých transformací
- doba, která uplyne od náhodně zvoleného okamžiku do nastoupení jevu, který se pravidelně opakuje časovém intervalu $(a;b)$
- libovolná spojitá veličina z intervalu $(a;b)$, o jejímž chování na tomto intervalu není nic bližšího známo (nouzové řešení v případě neznalosti skutečného rozdělení)

8.1.2 Exponenciální rozdělení

$$\mathrm{Exp}\,(\lambda)\,,\quad \lambda > 0 \qquad\qquad \mathrm{Exp}\,(\delta)\,,\quad \delta = \frac{1}{\lambda}$$

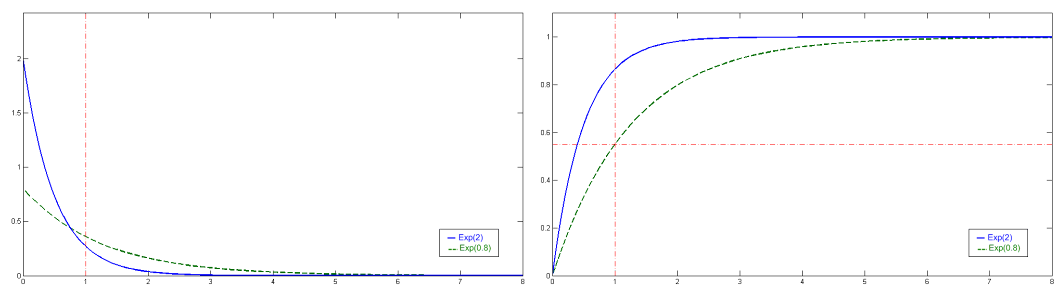
Náhodná veličina X může nabýt libovolné reálné hodnoty x z intervalu $[0, \infty\,)$.

Funkce hustoty $f\,(x) = \begin{cases} \lambda e^{-\lambda x} = \dfrac{1}{\delta} e^{-x/\delta} & \text{pro } x \geq 0 \\ 0 & \text{pro } x < 0 \end{cases}$

Distribuční funkce $F\,(x) = \begin{cases} 0 & \text{pro } x < 0 \\ 1 - e^{-\lambda x} = 1 - e^{-x/\delta} & \text{pro } x \geq 0 \end{cases}$

Střední hodnota a rozptyl

$$E(X) = \frac{1}{\lambda} = \delta \qquad\qquad D(X) = \frac{1}{\lambda^2} = \delta^2$$



Obrázek 10: Funkce hustoty a distribuční funkce exponenciálního rozdělení

Použití – doba čekání na určitou náhodnou událost, např. dobu životnosti součástek, které nepodléhají opotřebení

– λ označuje počet událostí za jednu časovou jednotku

– δ charakterizuje průměrnou dobu mezi výskytem dvou událostí

– jestliže se počet výskytů událostí během nějakého časového intervalu řídí Poissonovým rozdělením s parametrem λ, pak doba mezi výskytem dvou událostí se řídí exponenciálním rozdělením s parametrem λ

8.2 Příklady

1. Mějme náhodnou veličinu $X \sim \mathrm{R}\,(8; 12.5)$. Spočtěte

   (a) $P\,(X = 9.75)$;

   (b) $P\,(X > 11.3)$;

   (c) $P\,(8.8 < X < 10.1)$;

   (d) 50% kvantil $x_{0.5}$.

(e) Nakreslete graf hustoty náhodné veličiny X a znázorněte v něm $P(8.8 < X < 10.1)$.

(f) Nakreslete graf distribuční funkce náhodné veličiny X a znázorněte v něm $x_{0.5}$.

Řešení:
Hustota pravděpodobnosti je konstantní na intervalu $(8; 12.5)$, jinde je nulová. Tedy

$$f(x) = \begin{cases} \dfrac{1}{12.5-8} = \dfrac{1}{4.5} = 0.222 & \text{pro } x \in (8; 12.5) \\ 0 & \text{jinde.} \end{cases}$$

Distribuční funkce pro $x \in (8; 12.5)$ je pak $F(x) = \dfrac{x-8}{4.5}$.

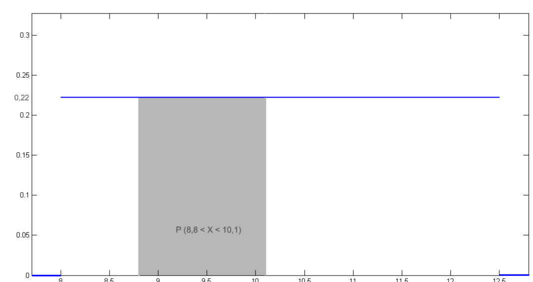
(a) $P(X = 9.75) = 0$ – jedná se o spojitou náhodnou veličinu

(b) $P(X > 11.3) = 1 - P(X \leq 11.3) = 1 - F(11.3) = 1 - \dfrac{11.3-8}{4.5} = 0.267$

(c) $P(8.8 < X < 10.1) = F(10.1) - F(8.8) = 0.467 - 0.178 = 0.289$

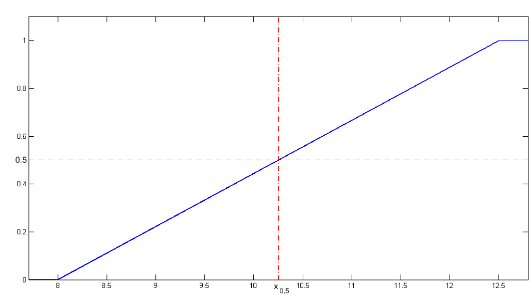
(d) $F(x_{0.5}) = 0.5 \Rightarrow x_{0.5} = 0.5 \cdot (12.5 - 8) + 8 = 10.25$

(e) Hustota rovnoměrného rozdělení $\mathrm{R}(8; 12.5)$



Obrázek 11: Funkce hustot rovnoměrného rozdělení

(f) Distribuční funkce rovnoměrného rozdělení $\mathrm{R}(8; 12.5)$



Obrázek 12: Distribuční funkce rovnoměrného rozdělení

2. Náhodná veličina má rovnoměrné rozdělení na intervalu $(0; 5)$. Určete:

(a) pravděpodobnost, že náhodná veličina X nabude hodnoty vyšší než 4, za předpokladu, že náhodná veličina již nabyla hodnoty 2.

(b) pravděpodobnost, že náhodná veličina nabude hodnoty nižší než 4, za předpokladu, že náhodná veličina již nabyla hodnoty 2.

Řešení:
Distribuční funkce této náhodné veličiny je $F(x) = \dfrac{x-0}{5-0} = \dfrac{x}{5}$.

(a) $P(X > 4|X > 2) = \dfrac{P(X > 4)}{P(X > 2)} = \dfrac{1 - F(4)}{1 - F(2)} = \dfrac{1 - 4/5}{1 - 2/5} = \dfrac{1}{5} \cdot \dfrac{5}{3} = \dfrac{1}{3}$

(b) $P(X < 4|X > 2) = \dfrac{P(2 < X < 4)}{P(X > 2)} = \dfrac{F(4) - F(2)}{1 - F(2)} = \dfrac{4/5 - 2/5}{1 - 2/5} = \dfrac{2}{5} \cdot \dfrac{5}{3} = \dfrac{2}{3}$

3. Předpokládejme, že průměrná doba zpracování zakázky je 30 sekund a řídí se exponenciálním rozdělením pravděpodobnosti.

(a) Určete pravděpodobnost, že zakázka se zpracuje do 1 minuty.

(b) Určete dobu, do níž se zakázka zpracuje s pravděpodobností 0.95.

Řešení:
Doba zpracování zakázky (v sekundách) $X \sim \text{Exp}(\delta = 30) = \text{Exp}\left(\lambda = \dfrac{1}{30}\right)$

(a) $P(X < 60) = P(X \leq 60) = F(60) = 1 - e^{-60/30} = 0.865$

(b) $F(t) = 0.95 \Rightarrow t = -30 \cdot \ln 0.05 = 89.87[s]$

4. Výrobce udává, že střední doba životnosti určité součástky je 4 roky. Za předpokladu, že životnost součástky se řídí exponenciálním rozdělením pravděpodobnosti a údaj daný výrobcem je pravdivý, spočtěte pravděpodobnost, že životnost náhodně vybrané součástky bude kratší, než půl roku.

Řešení:
Životnost součástky $X \sim \text{Exp}(\delta = 4)$. Platí

$$P(X < 0.5) = P(X \leq 0.5) = F(0.5) = 1 - e^{-0.5/4} = 0.118.$$

5. Mějme náhodnou veličinu $X \sim \text{Exp}(\delta = 11)$. Spočtěte

(a) $P(X = 27.5)$;

(b) $P(X < 9.9)$;

(c) $P(18.5 \leq X \leq 54.7)$;

(d) 10% kvantil $x_{0.1}$.

(e) Nakreslete graf hustoty náhodné veličiny X a znázorněte v něm $P(18.5 \leq X \leq 54.7)$ a $x_{0.1}$.

Řešení:
Distribuční funkce této náhodné veličiny je $F(x) = 1 - e^{-x/11}$ pro $x \geq 0$.

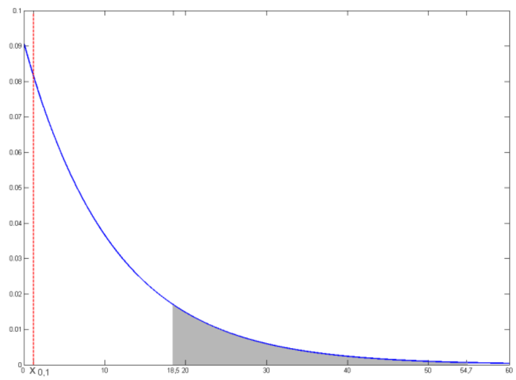
(a) $P(X = 27.5) = 0$ – jedná se o spojitou náhodnou veličinu

(b) $P(X < 9.9) = F(9.9) = 1 - e^{-9.9/11} = 0.593$

(c) $P(18.5 \leq X \leq 54.7) = F(54.7) - F(18.5) = 0.993 - 0.814 = 0.179$

(d) $F(x_{0.1}) = 0.1 \Rightarrow x_{0.1} = -11 \cdot \ln(1 - 0.1) = 1.159$

(e) Hustota exponenciálního rozdělení $\mathrm{Exp}(\delta = 11)$



Obrázek 13: Funkce hustoty exponenciálního rozdělení

8.3 Literatura s dalšími příklady

- Trial KMA – kapitola 30.6.1. Rovnoměrné rozdělení
- Trial KMA – kapitola 30.6.2. Exponenciální rozdělení
- Reif, Jiří – Kobeda, Zdeněk: Úvod do pravděpodobnosti a spolehlivosti. Strana 40–42, 55

9 Cvičení 9 - Normální rozdělení

9.1 Teoretická část

9.1.1 Normální (Gaussovo) rozdělení $N(\mu;\sigma^2)$

$$X \sim N(\mu;\sigma^2), \qquad \sigma^2 > 0, \qquad \mu;\sigma^2 \in \mathbf{R}$$

Náhodná veličina X může nabývat hodnot $x \in \mathbf{R}$.

Funkce hustoty: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$,

pro $x \in \mathbf{R}$ je symetrická kolem bodu μ.

Distribuční funkce: $F(x) = \int\limits_{-\infty}^{x} f(t)\mathrm{d}t$

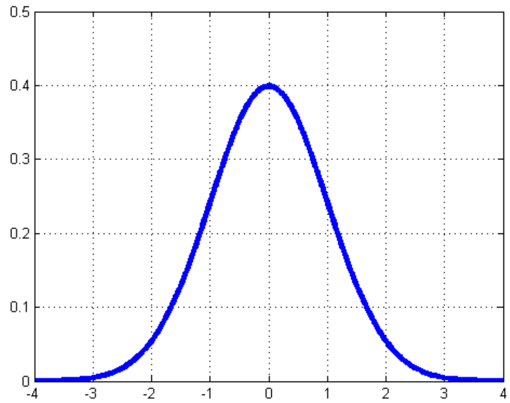
Distribuční funkci nelze vyjádřit pomocí elementárních funkcí. Hodnoty distribuční funkce určujeme pomocí tabelovaných hodnot pro normální normované rozdělení viz další odstavec. Dále platí, že pro $x \in \mathbf{R}$ je

$$F(\mu - x) = 1 - F(\mu + x).$$

Střední hodnota a rozptyl: $E(X) = \mu$, $D(X) = \sigma^2$

9.1.2 Normované normální rozdělení $N(\mu = 0; \sigma^2 = 1)$

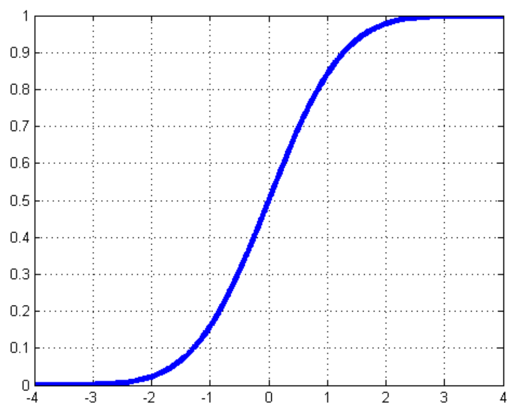
Funkce hustoty: $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ Funkci hustoty značíme φ.



Obrázek 14: Hustota pravděpodobnosti normovaného normálního rozdělení

Distribuční funkci značíme Φ. Její hodnoty jsou tabelovány.
Pro hodnoty distribuční funkce platí, že

$$\Phi(-x) = 1 - \Phi(x)$$



Obrázek 15: Distribuční funkce normovaného normálního rozdělení

Kvantily normovaného rozdělení značíme u_p. Pro p% kvantily $(0 < p < 1)$ platí, že

$$u_p = -u_{1-p}.$$

Hodnoty těchto kvantilů jsou tabelovány.

Vztah mezi Normálním a Normálním normovaným rozdělením:

$$X \sim N(\mu; \sigma^2) \Rightarrow U = \frac{X - \mu}{\sigma} \sim N(0; 1)$$

Má-li náhodná proměnná X rozdělení $N(\mu; \sigma^2)$ s distribuční funkcí $F(x)$, pak příslušná normovaná proměnná má normované normální rozdělení. Platí tedy

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

Pro kvantily normálního a normálního normovaného rozdělení platí:

$$x_p = u_p \cdot \sigma + \mu \text{ je } p\% \text{ kvantil } N(\mu; \sigma^2)$$

9.1.3 Použití normálního rozdělení

Normálním rozdělením se v praxi řídí náhodné proměnné, jejichž hodnota je součtem velkého množství vzájemně nezávislých vlivů, z nichž žádný nemá dominantní význam - např. chyby měření. Dále se pak využívá k aproximaci jiných diskrétních i spojitých náhodných proměnných.

Příklady náhodných veličin s normálním rozdělením:

- Náhodné chyby fyzikálních (obecně jakýchkoli) měření
- Veličiny utvářející se pod vlivem balistických zákonů (výsledky střelby)
- Znaky v biologických populacích podléhající zákonům genetiky
- Náhodné veličiny vznikající jako součty či průměry jiných náhodných veličin (spojitých ale i diskrétních) s libovolným rozdělením

Aproximace jiných typů rozdělení

- pro $\lambda \to \infty$ platí $Po(\lambda) \approx N(\mu = \lambda; \sigma^2 = \lambda)$
- pro $n \to \infty$ platí $Bi(p; n) \approx N(\mu = n\ p; \sigma^2 = n\ p\ (1 - p))$

Použití těchto aproximací je doporučeno pro $D(X) = \sigma^2 \geq 9$. Dále pak

- $P(a \leq X \leq b) \doteq \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)$
- pro nepříliš velké hodnoty n (řádově stovky) používáme přesnější aproximaci (tzv. korekce na spojitost)
  $P(a \leq X \leq b) \doteq \Phi\left(\frac{b + 0.5 - \mu}{\sigma}\right) - \Phi\left(\frac{a - 0.5 - \mu}{\sigma}\right)$

9.1.4 Centrální limitní věta

Nechť $X_i,\ i = 1, 2, \ldots, n$ jsou vzájemně nezávislé náhodné veličiny se stejným rozdělením, $E(X_i) = \mu_0$, $D(X_i) = \sigma_0^2$. Pak platí

$$\begin{aligned} X &= \sum_{i=1}^{n} X_i \approx N(n\mu_0; n\sigma_0^2) \\ \overline{X} &= \frac{1}{n}\sum_{i=1}^{n} X_i \approx N(\mu_0; \sigma_0^2/n) \end{aligned}$$

9.2 Příklady

1. Čekáme na autobus v horské vesnici. Dlouhodobým pozorováním bylo zjištěno, že *zpoždění* odjezdu autobusu ze zastávky se přibližně řídí normálním rozdělením se střední hodnotou 10 min. a rozptylem 25 (min^2). Spočtěte:

   (a) ppst, že autobus bude mít zpoždění více než 20 min.;

   (b) ppst, že autobus odjede dříve;

   (c) ppst, že autobus odjede o 0 až 2.5 min. dříve;

   (d) ppst, že autobus bude mít zpoždění více než 20 min., jestliže již má zpoždění 15 min.;

   (e) čas, ve který bychom měli být na zastávce, aby nám autobus neujel alespoň na 90%.

   (f) nakreslete graf hustoty pravděpodobnosti a v něm znázorněte ppst, že autobus odjede o 0 až 2.5 min. dříve;

   Řešení
   X … zpoždění autobusu $X \sim N(10;\ 25)$

   (a) $P(X > 20) = 1 - P(X \leq 20) = 1 - \overbrace{\phi\underbrace{\left(\frac{20-10}{5}\right)}_{2}}^{0.9772} = 0.0228$

   (b) $P(X < 0) = P(X \leq 0) = F(0) = \phi\left(\frac{0-10}{5}\right) = \phi(-2) = 1 - \phi(2) =$
   $= 0.0228$

   (c) $P(-2.5 < X < 0) = F(0) - F(-2.5) = 0.0228 - \phi\left(\frac{-2.5-10}{5}\right) =$
   $= 0.0228 - \phi\,(-2.5) = 0.0228 - 1 + \phi(2.5) = 0.0228 - 1 + 0.9938 = 0.0166$

   (d) $P(X > 20|X > 15) = \frac{P(X>20)}{P(X>15)} = \frac{1-F(20)}{1-F(15)} = \frac{1-\phi\left(\frac{10}{5}\right)}{1-\phi\left(\frac{5}{5}\right)} = \frac{1-0.9772}{1-0.8413} = \frac{0.0228}{0.1587} =$
   $0.1436 = 14.36\%$

   (e) $x_{0.1} = \mu + \sigma u_{0.1} = 10 - 5 \cdot 1.2816 = 3.592$

2. Náhodná proměnná X má normální rozdělení s parametry μ, σ_0^2. Zjistěte následující pravděpodobnosti

(a) $P(X \in (\mu - \sigma; \mu + \sigma))$

(b) $P(X \in (\mu - 2\sigma; \mu + 2\sigma))$

(c) $P(X \in (\mu - 3\sigma; \mu + 3\sigma))$

Řešení

(a) $P(\mu - \sigma < X < \mu + \sigma) = F(\mu + \sigma) - F(\mu - \sigma) = \phi\left(\frac{\mu + \sigma - \mu}{\sigma}\right) - \phi\left(\frac{\mu - \sigma - \mu}{\sigma}\right) = \phi(1) - \phi(-1) = \phi(1) - 1 + \phi(1) = 0.68268$

(b) $P(\mu - 2\sigma < X < \mu + 2\sigma) = \phi(2) - \phi(-2) = 0.9545$

(c) $P(\mu - 3\sigma < X < \mu + 3\sigma) = \phi(3) - \phi(-3) = 0.9973$

3. Pro náhodnou proměnnou s normálním rozdělením platí, že

$$P(X \leq 4) = 0.6, \qquad P(X \geq 0) = 0.8$$

Zjistěte hodnoty parametrů μ, σ_0^2.

Řešení:
$u_{0.6} = 0.2533 = \frac{4 - \mu}{\sigma}$ a současně $u_{0.8} = 0.8416 = \frac{\mu}{\sigma}$. Řešením soustavy dvou rovnic o dvou neznámých získáme řešení $\mu = 3.08$, $\sigma_0^2 = 13.35$.

4. Telefonní ústředna spojí průměrně 76 hovorů za minutu a jejich počet se řídí Poissonovým rozdělením. Spočtěte pravděpodobnost, že ústředna za minutu spojí více než 80 hovorů.

Řešení:

$$X \sim Po(76) \qquad X \in \{0, 1, ...\}$$

$P(X > 80) = 1 - P(X \leq 80) = 1 - P(0) - P(1) - \ldots - P(80)$. Výpočet standardním způsobem je velice náročný. Prověříme předpoklady možných aproximací. Rozptyl náhodné veličiny má hodnotu 76, tzn. podmínka aproximace Poissonova rozdělení rozdělením normálním je splněna ($\sigma^2 \geq 9$). Platí tedy

$$X \sim N(76; 76)$$

$P(X > 80) = 1 - P(X \leq 80) = 1 - F_{Poisson}(80) = 1 - F_{Norm.}(80.5) = 1 - \phi\left(\dfrac{80.5 - 76}{8.718}\right) =$ $1 - 0.6985 = 0.3015$.

5. Zaokrouhlovací chyba na celé jednotky má rovnoměrné rozložení na intervalu (-0.5; 0.5). Spočtěte pravděpodobnost, že součet 100 zaokrouhlovacích chyb (nezávislých) bude v absolutní hodnotě menší než 5.

Řešení:

$$\text{Zaokrouhlovací chyba } X_i \sim R(-0.5; 0.5)$$

Označme $S = \sum_{i=1}^{100} X_i \sim N(100 \cdot 0; \frac{100}{12})$

Máme zjistit $P(-5 < S < 5) = F(5) - F(-5) = \phi\left(\dfrac{5}{\sqrt{\frac{100}{12}}}\right) - 1 + \phi\left(\dfrac{5}{\sqrt{\frac{100}{12}}}\right) = 0.9164$

9.3 Literatura s dalšími příklady

- Trial KMA – kapitola 30.6.3. Normální rozdělení
- Brousek, Jan – Ryjáček, Zdeněk: Sbírka řešených příkladů z počtu pravděpodobnosti. Strana 86–88.

10 Cvičení 10 - Statistický soubor. Náhodný výběr a výběrové statistiky. Odhady parametrů.

Statistický soubor. Náhodný výběr a výběrové statistiky (aritmetický průměr, geometrický průměr, výběrový rozptyl,...). Bodové odhady parametrů. Intervalové odhady parametrů. Jednostranné a oboustranné odhady. Intervalový odhad střední hodnoty, rozptylu, relativní četnosti.

10.1 Teoretická část

10.1.1 Statistika

Statistika je matematická disciplína, která vychází z empirických dat (pozorování), ze kterých pak dělá obecné závěry. Zabývá se řešením problémů náhodných situací - např. odhady hodnot platné s určitou ppstí, ohodnocení rizik při rozhodování, aj.. V teorii statistiky je náhodnost a neurčitost modelována pomocí teorie pravděpodobnosti. Statistika nám také poskytuje soubor matematických metod (postupů) pro plánování experimentů, získávání dat a jejich analýzu a následnou interpretaci závěrů. Závěry a rozhodnutí učiněné na základě statistických modelů mohou, ale nemusí odpovídat realitě. Statistické postupy můžeme rozdělit na:

- Konfirmační analýzu, která se zabývá testováním předem přesně formulovaných hypotéz. Zjednodušeně řečeno, úkolem konfirmační analýzy je dávat odpovědi na otázky typu: Je pravda, že ...?
- Explorační analýzu, při které není dostatečně jasné, co vše může být výsledkem. Jejím cílem je vyčíst z dat maximum informace, inspirace, poučení - to vše vzhledem k nějakému obecnému, často vágně formulovanému problému (např. analýza příčin poruchovosti).

Jako statistiku také označujeme hodnoty, které získáme provedením náhodného výběru.

10.1.2 Základní soubor

Základní soubor představuje množinu všech prvků s konkrétními sledovanými vlastnostmi, které jsou podrobeny zkoumání (např. obyvatelstvo ČR ke dni ..., výrobky vyrobené v závodě Z v době od ...do ...). Obvykle je tento soubor velmi rozsáhlý - může být konečný i nekonečný. Základní soubor je charakterizován charakteristikami - střední hodnota, rozptyl, variační rozpětí, ...

10.1.3 Výběrový soubor (statistický soubor)

Výběrový soubor představuje konečnou podmnožinu základního souboru - $n-$tice reálných čísel, získanou na základě výsledků statistického experimentu.

- **Uspořádaný statistický soubor** - Statistický soubor s uspořádanými prvky podle velikosti. Hodnoty v souboru se mohou opakovat.

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

- **Popisná statistika** - definuje výběrové charakteristiky (statistiky, míry) výběrového souboru: charakteristiky (míry) polohy, charakteristiky (míry) variability, ...

10.1.4 Popisná statistika

- Charakteristiky polohy
  - Aritmetický průměr
    $$\overline{x} = x_1 + x_2 + \cdots + x_n$$
    Dále platí.
    $$\sum_{i=1}^{n}(x_i - \overline{x}) = 0.$$
    Pro libovolné a $a \neq \overline{x}$ platí: $\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2 < \sum_{i=1}^{n}(x_i - a)^2$
    Nechť jsou $a, b \in R$ a položme $y_i = a \cdot x_i + b$ pro $i = 1, 2, \ldots, n$, pak $\overline{y} = a \cdot \overline{x} + b$.
    Aritmetický průměr je citlivý na hrubé chyby
    (př. 8,00; 12,00; 15,00; 23,00; 1500) $\Rightarrow \overline{x} = 311{,}60$).
  - Geometrický průměr
    $$x_G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots x_n}$$
    Geometrický průměr je používán pouze pro kladné hodnoty x_i. Využívá se zejména pro určení průměrné hodnoty tzv. řetězových indexů. Tj. nechť $x_0, x_1, \ldots, x_n$ udávají počet prodaných výrobků v i- tém časovém období. Vývoj prodeje charakterizujeme pomocí tzv. řetězových indexů $i_1 = \dfrac{x_1}{x_0}, i_2 = \dfrac{x_2}{x_1}, \ldots, i_n = \dfrac{x_n}{x_{n-1}}$.
    Pak lze vyjádřit $x_n = x_0 \cdot i_1 \cdot i_2 \cdot \cdots \cdot i_n$. „Průměrnou " hodnotu indexu i_k charakterizuje nejlépe geometrický průměr. $x_n = x_0 \cdot \overline{i}_G$
  - Harmonický průměr
    $$\overline{x}_H = \frac{n}{x_1^{-1} + x_2^{-1} + \cdots + x_n^{-1}}$$
    Auto jede do kopce rychlosti v_1 a po stejné dráze z kopce rychlosti v_2. Jaká je jeho průměrná rychlost ?
    Řešení: Délku tratě označme d, dobu jízdy do kopce $t_1 = d/v_1$, dobu jízdy z kopce $t_2 = d/v_2$, průměrná rychlost je $\dfrac{2d}{t_1 + t_2} = \dfrac{2}{v_1^{-1} + v_2^{-1}} = \overline{v}_H$
    Pro jednotlivé typy průměrů platí:
    $$x_{(1)} \leq \overline{x}_H \leq \overline{x}_G \leq \overline{x} \leq x_{(n)}$$
    Rovnost je splněna když jsou všechny prvky x_i shodné.
  - Medián $\widetilde{x}$ představuje prvek, který se ve statistickém uspořádaném souboru nachází v polovině. Představuje robustní míru polohy tzn. není citlivý na hrubé chyby.
    $$\begin{aligned} \widetilde{x} &= x_{(m)} && \text{pro } n \text{ liché, } n = 2m - 1 \\ &= \frac{1}{2}\left(x_{(m)} + x_{(m+1)}\right) && \text{pro } n \text{ sudé, } n = 2m \end{aligned}$$
    Medián není citlivý na hrubé chyby
    (př. $8, 12, 15, 23, 1500 \Rightarrow \widetilde{x} = 15$)

– Modus $\hat{x}$ je nejčastěji se vyskytující hodnoty v souboru $x_1, x_2, \ldots, x_n$. modus není určen jednoznačně

- Charakteristiky variability

10.2 Příklady

1. Pro zadaná data vypočtěte výběrové statistické charakteristiky

Data 1	2	3	2	4	15	2
Data 2	20	22	19	2	21	20

	Data 1	Data 2
průměr	4.67	17.33
geom. průměr	3.36	13.84
harm. průměr	2.79	8.05
modus	2.00	20.00
medián	2.50	20.00
rozptyl	21.89	47.89
směrodatná odchylka	4.68	6.92
výběrový rozptyl	26.27	57.47
výb. sm. odchylka	5.13	7.58
variační rozpětí	13.00	20.00
variační koeficient	2.79	1.15

2. Pro zadaná data odhadněte základní charakteristiky

Data a)	8	6	11	7	9	9	12	13		
Data b)	175	186	189	169	170	184				
Data c)	3	3	3	3	6	0	0	1	2	3
Data d)	5.4	9.4	22.4	1.6	4.9	14.1	34.1	9.3	1.4	

a)	odhad λ	9.38
b)	odhad μ	178.83
	odhad σ^2	74.17
c)	odhad p	0.02
d)	odhad λ	0.09

3. Pro zadaná data určete intervalový odhad střední hodnoty při známém rozptylu

Data	175	186	189	169	170	184
σ^2	49					

odhad μ	178.83
dolní mez	173.23
horní mez	184.43

4. Pro zadaná data určete intervalový odhad střední hodnoty při neznámém rozptylu a intervalový odhad rozptylu

	A	B	C	D	E	F	G	H
1								
2		Data	175	186	189	169	170	184
3								
4								
5		odhad μ	178.83					
6		dolní mez	170.23					
7		horní mez	187.44					
8		odhad σ^2	74.17					
9		dolní mez	28.90					
10		horní mez	446.14					
11								

10.3 Literatura s dalšími příklady

- Brousek, Jan – Ryjáček, Zdeněk: Sbírka řešených příkladů z počtu pravděpodobnosti.

11 Cvičení 11 - Testování statistických hypotéz

11.1 Teoretická část

11.1.1 Testování statistických hypotéz

Testování statistických hypotéz slouží k ověření, zda experimentálně získaná data vyhovují představám, které byly na základě těchto dat získány.
Při testování statistických hypotéz porovnáváme dvě hypotézy:

- **Nulová (testovaná) hypotéza**: jedná se o hypotézu, kterou testujeme, označuje se H_0 (například hypotéza, že lék **nemá** námi žádaný účinek).
- **Alternativní hypotéza**: jedná se o hypotézu, oproti které provádíme test, označuje se H_1 (například hypotéza, že lék **má** námi žádaný účinek).

11.1.2 Základní pojmy

- **Chyba 1. druhu** (α): hypotéza H_0 **platí**, ale my ji na základě experimentu **zamítneme**. Parametr α se nazývá **hladina významnosti** testu, obvykle se volí malé hodnoty, např. $\alpha = 0.05$, $\alpha = 0.01$.
- **Chyba 2. druhu** (β): hypotéza H_0 **neplatí**, ale my ji na základě experimentu **přijmeme**. Hodnota $1 - \beta$ udává **sílu testu**, tj. pravděpodobnost, že neplatná hypotéza bude zamítnuta. Platí, že za jinak stejných podmínek vede snižování α ke zvyšování β a naopak.

11.1.3 Postup při testování

1. Formulujeme nulovou a alternativní hypotézu.
2. Zvolíme vhodné testovací kritérium, pomocí kterého budeme hypotézu testovat.
3. Zvolíme hladinu významnosti testu α.
4. S ohledem na alternativní hypotézu vymezíme **kritický obor testu** W tak, aby pravděpodobnost toho, že bude zamítnuta platná hypotéza byla nejvýše rovna hodnotě α. Doplňkem W je **obor přijetí** V. Platí tedy $W \cap V = \emptyset$.
5. Zjistíme hodnotu testovacího kritéria T. Pokud platí $T \in W$, tak H_0 zamítáme (na hladině významnosti α). V opačném případě H_0 nezamítáme (na hladině významnosti α). Platí:

$$\alpha = P(T \in W \mid H_0)$$

$$\beta = P(T \in V \mid H_1)$$

Pokud provádíme test hypotézy v nějakém SW, tak není potřeba zadávat α, ale bývá ve většině případů k dispozici p-hodnota testu. Testovanou hypotézu lze zamítnout na hladině významnosti α pokud je p-hodnota testu menší než α.

11.1.4 Test hypotézy $\mu = \mu_0$ při známém rozptylu (z-test)

Mějme náhodný výběr $x_1, x_2, ..., x_n$ z normálního rozdělení $N(\mu, \sigma^2)$, kde rozptyl σ^2 známe. Na hladině významnosti α testujeme hypotézu:

$$H_0 : \mu = \mu_0$$

Oproti hypotéze:

- Oboustranná alternativa $H_1 : \mu \neq \mu_0$
- Jednostranná alternativa $H_1 : \mu < \mu_0$ resp. $H_1 : \mu > \mu_0$

Testovací statistika má tvar

$$z = \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n}$$

Za předpokladu platnosti hypotézy H_0 je z realizací náhodné veličiny s rozdělením $N(0, 1)$. Obor kritických hodnot W je pak dán:

- V případě oboustranné alternativy $H_1 : \mu \neq \mu_0$: $W = (-\infty, u_{\frac{\alpha}{2}}) \cup (u_{1-\frac{\alpha}{2}}, +\infty)$
- V případě jednostranné alternativy $H_1 : \mu < \mu_0$: $W = (-\infty, u_\alpha)$
- V případě jednostranné alternativy $H_1 : \mu > \mu_0$: $W = (u_{1-\alpha}, +\infty)$

V případě, že $z \in W$, tak hypotézu H_0 zamítáme.

11.1.5 Test hypotézy $\mu = \mu_0$ při neznámém rozptylu (t-test)

Mějme náhodný výběr $x_1, x_2, ..., x_n$ z normálního rozdělení $N(\mu, \sigma^2)$, kde rozptyl σ^2 neznáme. Na hladině významnosti α testujeme hypotézu:

$$H_0 : \mu = \mu_0$$

Oproti hypotéze:

- Oboustranná alternativa $H_1 : \mu \neq \mu_0$
- Jednostranná alternativa $H_1 : \mu < \mu_0$ resp. $H_1 : \mu > \mu_0$

Testovací statistika má tvar

$$t = \frac{\bar{x} - \mu_0}{s}\sqrt{n}$$

kde s je výběrová směrodatná odchylka:

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

Za předpokladu platnosti hypotézy H_0 je t realizací náhodné veličiny se Studentovo rozdělením (t-rozdělením) $t(\nu)$ s počtem stupňů volnosti $\nu = n - 1$. Obor kritických hodnot W je pak dán:

- V případě oboustranné alternativy $H_1 : \mu \neq \mu_0$: $W = (-\infty, t_{\frac{\alpha}{2}}(n-1)) \cup (t_{1-\frac{\alpha}{2}}(n-1), +\infty)$
- V případě jednostranné alternativy $H_1 : \mu < \mu_0$: $W = (-\infty, t_\alpha(n-1))$
- V případě jednostranné alternativy $H_1 : \mu > \mu_0$: $W = (t_{1-\alpha}(n-1), +\infty)$

11.1.6 Párový t-test

Mějme $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ náhodný výběr z dvourozměrného normálního rozdělení. Označíme $\mu_1 = E(X)$ a $\mu_2 = E(Y)$. Na hladině významnosti α testujeme hypotézu:

$$H_0 : \mu_1 - \mu_2 = d, d \in R$$

Pozn.: Pro případ testování shody středních hodnot testujeme $\mu_1 - \mu_2 = 0$.
Oproti hypotéze:

- Oboustranná alternativa $H_1 : \mu_1 - \mu_2 \neq d$
- Jednostranná alternativa $H_1 : \mu_1 - \mu_2 < d$ resp. $H_1 : \mu_1 - \mu_2 > d$

Tento test lze převést na předchozí jednovýběrový t-test (11.1.5) tak, že vytvoříme $z_1 = x_1 - y_1$, $z_2 = x_2 - y_2, ..., z_n = x_n - y_n$ a testujeme hypotézu $H_0 : \mu_z = d$.

11.1.7 t-test pro dva nezávislé výběry z normálních rozdělení se stejnými rozptyly

Mějme $x_1, x_2, ..., x_n$ náhodný výběr rozsahu n z normálního rozdělení $N(\mu_1, \sigma^2)$, kde rozptyl σ^2 neznáme a $y_1, y_2, ..., y_m$ náhodný výběr rozsahu m z normálního rozdělení $N(\mu_2, \sigma^2)$, kde rozptyl σ^2 neznáme, ale je stejný jako v prvním případě. Předpokládáme, že oba výběry jsou nezávislé. Na hladině významnosti α testujeme hypotézu:

$$H_0 : \mu_1 - \mu_2 = d, d \in R$$

Pozn.: Pro případ testování shody středních hodnot testujeme $\mu_1 - \mu_2 = 0$.
Oproti hypotéze:

- Oboustranná alternativa $H_1 : \mu_1 - \mu_2 \neq d$
- Jednostranná alternativa $H_1 : \mu_1 - \mu_2 < d$ resp. $H_1 : \mu_1 - \mu_2 > d$

Testovací statistika má tvar

$$t = \frac{\bar{x} - \bar{y} - d}{\sqrt{(n-1)s_x^2 + (m-1)s_y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}}$$

kde s_x^2 a s_y^2 jsou výběrové rozptyly. Za předpokladu platnosti hypotézy H_0 je t realizací náhodné veličiny se Studentovo rozdělením (t-rozdělením) $t(\nu)$ s počtem stupňů volnosti $\nu = n + m - 2$. Obor kritických hodnot W je pak dán:

- V případě oboustranné alternativy $H_1 : \mu_1 - \mu_2 \neq d$:

$$W = (-\infty, t_{\frac{\alpha}{2}}(n+m-2)) \cup (t_{1-\frac{\alpha}{2}}(n+m-2), +\infty)$$

- V případě jednostranné alternativy $H_1 : \mu_1 - \mu_2 < d$: $W = (-\infty, t_\alpha(n+m-2))$
- V případě jednostranné alternativy $H_1 : \mu_1 - \mu_2 > d$: $W = (t_{1-\alpha}(n+m-2), +\infty)$

11.1.8 Test o rozptylu normálního rozdělení

Mějme náhodný výběr $x_1, x_2, ..., x_n$ z normálního rozdělení $N(\mu, \sigma^2)$. Na hladině významnosti α testujeme hypotézu:

$$H_0 : \sigma^2 = \sigma_0^2$$

Oproti hypotéze:

- Oboustranná alternativa $H_1 : \sigma^2 \neq \sigma_0^2$
- Jednostranná alternativa $H_1 : \sigma^2 < \sigma_0^2$ resp. $H_1 : \sigma^2 > \sigma_0^2$

Testovací statistika má tvar

$$t = \frac{(n-1)s^2}{\sigma_0^2}$$

Za předpokladu platnosti hypotézy H_0 je t realizací náhodné veličiny s χ^2-rozdělením $\chi^2(\nu)$ s počtem stupňů volnosti $\nu = n - 1$. Obor kritických hodnot W je pak dán:

- V případě oboustranné alternativy $H_1 : \sigma^2 \neq \sigma_0^2$:

$$W = (-\infty, \chi^2_{\frac{\alpha}{2}}(n-1)) \cup (\chi^2_{1-\frac{\alpha}{2}}(n-1), +\infty)$$

- V případě jednostranné alternativy $H_1 : \sigma^2 < \sigma_0^2$: $W = (-\infty, \chi^2_\alpha(n-1))$
- V případě jednostranné alternativy $H_1 : \sigma^2 > \sigma_0^2$: $W = (\chi^2_{1-\alpha}(n-1), +\infty)$

11.1.9 Test shody dvou rozptylů

Mějme dva nezávislé náhodné výběry $x_1, x_2, ..., x_n \sim N(\mu_1, \sigma_1^2)$ a $y_1, y_2, ..., y_m \sim N(\mu_2, \sigma_2^2)$. Na hladině významnosti α testujeme hypotézu:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

Oproti hypotéze:

- Oboustranná alternativa $H_1 : \sigma_1^2 \neq \sigma_2^2$
- Jednostranná alternativa $H_1 : \sigma_1^2 < \sigma_2^2$ resp. $H_1 : \sigma_1^2 > \sigma_2^2$

Testovací statistika má tvar

$$Z = \frac{s_1^2}{s_2^2}$$

Za předpokladu platnosti hypotézy H_0 je Z realizací náhodné veličiny s Fisherovo rozdělením $F(\nu_1, \nu_2)$ s počtem stupňů volnosti $\nu_1 = n - 1$ a $\nu_2 = m - 1$. Obor kritických hodnot W je pak dán:

- V případě oboustranné alternativy $H_1 : \sigma_1^2 \neq \sigma_2^2$:

$$W = (-\infty, F_{\frac{\alpha}{2}}(n-1, m-1)) \cup (F_{1-\frac{\alpha}{2}}(n-1, m-1), +\infty)$$

- V případě jednostranné alternativy $H_1 : \sigma_1^2 < \sigma_2^2$: $W = (-\infty, F_\alpha(n-1, m-1))$
- V případě jednostranné alternativy $H_1 : \sigma_1^2 > \sigma_2^2$: $W = (F_{1-\alpha}(n-1, m-1), +\infty)$

11.2 Příklady

1. Stroj má vyrábět výrobky o délce 2 metry. Náhodně bylo vybráno 20 výrobků a byla u nich zjištěna průměrná délka $\bar{x} = 2.06m$. Za předpokladu, že náhodný výběr pochází z normálního rozdělení s rozptylem $\sigma^2 = 0.025$ testujte na hladině významnosti $\alpha = 0.05$ hypotézu $H_0 : \mu = 2$ oproti hypotéze:

   (a) $H_1 : \mu \neq 2$

   (b) $H_1 : \mu > 2$

   Řešení:
   Jelikož známe rozptyl, budeme v obou případech používat z-test (viz 11.1.4). Dosadíme do vzorce pro statistiku z:
   $$z = \frac{\bar{x} - \mu_0}{\sigma}\sqrt{n} = \frac{2.06 - 2}{\sqrt{0.025}}\sqrt{20} = 1.6971$$
   Pro oba případy teď určíme kritický obor:

   (a) Jedná se o oboustranný test, "vadí"nám tedy jak odchylky směrem dolů, tak i směrem nahoru. Musíme tedy rozdělit 5% na dvě části. Kritický obor je dán:
   $$W = (-\infty, u_{\frac{\alpha}{2}}) \cup (u_{1-\frac{\alpha}{2}}, +\infty) = (-\infty, u_{0.025}) \cup (u_{0.975}, +\infty)$$
   V tabulkách pro normální normované rozdělení najdene hodnotu 97.5% kvantilu.
   $$u_{0.975} = 1.96$$
   Pro normální normované rozdělení platí, že $u_p = -u_{1-p}$. Proto
   $$u_{0.025} = -u_{0.975} = -1.96$$
   Kritický obor je tedy:
   $$W = (-\infty, -1.96) \cup (1.96, +\infty)$$
   Jelikož $z \notin W$, tak uvedenou hypotézu na hladině významnosti $\alpha = 5\%$ nezamítáme.

   (b) V tomto případě je alternativní hypotéza pouze jednostranná, proto nebudeme 5% rozdělovat. Kritický obor je:
   $$W = (u_{1-\alpha}, +\infty) = (u_{0.95}, +\infty) = (1.6449, +\infty)$$
   V tomto případě platí $z \in W$ a proto uvedenou hypotézu na hladině významnosti $\alpha = 5\%$ zamítáme.

2. U 5 sáčků kávy byly zjištěny následující váhy: 249g, 247g, 252g, 244g a 247g. Na hladině významnosti $\alpha = 10\%$ testujte hypotézu, že průměrná váha je 250g oproti hypotéze:

   (a) $H_1 : \mu \neq 250g$

   (b) $H_1 : \mu < 250g$

Řešení:
Jelikož neznáme rozptyl, budeme v obou případech používat t-test (viz 11.1.5). Nejdříve musíme spočítat průměr a rozptyl:

$$\bar{x} = 247.8$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{4}\sum_{i=1}^{5}(x_i - 247.8)^2 = 8.7$$

Dosadíme do vzorce pro statistiku t:

$$t = \frac{\bar{x} - \mu_0}{s}\sqrt{n} = \frac{247.8 - 250}{\sqrt{8.7}}\sqrt{5} = -1.6678$$

Pro oba případy teď určíme kritický obor:

(a) Jedná se o oboustranný test, "vadí"nám tedy jak odchylky směrem dolů, tak i směrem nahoru. Musíme tedy rozdělit 10% na dvě části. Kritický obor je dán:

$$W = (-\infty, t_{\frac{\alpha}{2}}(n-1)) \cup (t_{1-\frac{\alpha}{2}}(n-1), +\infty) = (-\infty, t_{0.05}(4)) \cup (t_{0.95}(4), +\infty)$$

V tabulkách pro Studentovo rozdělení najdene hodnotu 95% kvantilu.

$$t_{0.95}(4) = 2.1318$$

Pro Studentovo rozdělení platí, že $t_p = -t_{1-p}$. Proto

$$t_{0.05}(4) = -t_{0.95}(4) = -2.1318$$

Kritický obor je tedy:

$$W = (-\infty, -2.1318) \cup (2.1318, +\infty)$$

Jelikož $t \notin W$, tak uvedenou hypotézu na hladině významnosti $\alpha = 10\%$ nezamítáme.

(b) V tomto případě je alternativní hypotéza pouze jednostranná, proto nebudeme 10% rozdělovat. Kritický obor je:

$$W = (-\infty, t_\alpha(n-1)) = (-\infty, t_{0.1}(4)) = (-\infty, -1.5332)$$

V tomto případě platí $t \in W$ a proto uvedenou hypotézu na hladině významnosti $\alpha = 10\%$ zamítáme a přijmeme alternativní hypotézu.

3. U stroje pozorujeme následující dvojice rozměrů výrobků v mm (před opravou, po opravě): (100, 97), (105, 102), (96, 101), (92, 98) a (101, 100). Na hladině významnosti 5% chceme testovat vliv opravy na stroj.

Řešení:
Budeme tedy testovat hypotézu, že oprava neměla vliv (měla nulový efekt) $H_0 : \mu_1 - \mu_2 = 0$ (viz 11.1.6). Alternativa je, že oprava vliv měla, proto bude alternativní hypotéza $H_1 : \mu_1 - \mu_2 \neq 0$. Tento test převedeme na jednovýběrový t-test (viz 11.1.5) pro hodnoty 3, 3, -5, -6 a 1. Hypotézy jsou převedeny na:

$$H_0 : \mu_z = 0$$

$$H_1 : \mu_z \neq 0$$

Hodnota statistiky:

$$t = -0.4083$$

Kritický obor W:

$$W = (-\infty, -2.7765) \cup (2.7765, +\infty)$$

Jelikož $t \notin W$, tak uvedenou hypotézu na hladině významnosti $\alpha = 5\%$ nezamítáme. Tedy, hypotéza o tom, že střední hodnoty se neliší nebyla zamítnuta, znamená to, že oprava neměla vliv.

4. Výsledky dvou skupin studentů při písemce jsou následující:

   - 1. skupina: Počet studentů=29, průměr=6.97, směrodatná odchylka=2.38
   - 2. skupina: Počet studentů=20, průměr=7.48, směrodatná odchylka=1.77

   Na hladině významnosti 5% chceme otestovat, zda jsou hodnoty stejné, tedy

   $$H_0 : \mu_1 - \mu_2 = 0$$

   oproti hypotéze, že hodnoty jsou různé, tedy

   $$H_1 : \mu_1 - \mu_2 \neq 0$$

   Řešení:
   Budeme řešit dle 11.1.7. Předpokládáme tedy, že se jedná o výběry z normálních rozdělení se stejným rozptylem a že oba výběry jsou nezávislé. Testovací statistika:

   $$t = \frac{\bar{x} - \bar{y} - d}{\sqrt{(n-1)s_x^2 + (m-1)s_y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}} =$$

   $$= \frac{6.97 - 7.48}{\sqrt{28 \cdot 2.38^2 + 19 \cdot 1.77^2}} \sqrt{\frac{29 \cdot 20(29+20-2)}{29+20}} = -0.8145$$

   Kritický obor W:

   $$W = (-\infty, t_{\frac{\alpha}{2}}(n+m-2)) \cup (t_{1-\frac{\alpha}{2}}(n+m-2), +\infty) = (-\infty, -2.012) \cup (2.012, +\infty)$$

   Jelikož $t \notin W$, tak uvedenou hypotézu na hladině významnosti $\alpha = 5\%$ nezamítáme. Tedy, hypotéza o tom, že střední hodnoty se neliší nebyla zamítnuta, znamená to, že střední hodnoty v obou skupinách se významně neliší.

5. Pevnost vlákna bavlněné příze lze pokládat za náhodnou veličinu s rozdělením $N(\mu, \sigma^2)$. Je-li $\sigma^2 > 0.36kg^2$, vznikají potíže při tkaní. Při zkoušce pevnosti 11 náhodně vybraných vláken byly zjištěny tyto hodnoty jejich pevnosti:
   5.3, 3.0, 4.8, 3.6, 4.1, 2.5, 4.7, 2.4, 3.2, 3.8 a 4.4
   Na hladině významnosti $\alpha = 5\%$ chceme testovat hypotézu $H_0 : \sigma^2 = 0.36$ oproti alternativě $H_1 : \sigma^2 > 0.36$.

Řešení:
Budeme používat test o rozptylu normálního rozdělení (viz 11.1.8). Nejdříve si musíme vypočítat průměr a výběrový rozptyl:

$$\bar{x} = 3.8$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{10}\sum_{i=1}^{11}(x_i - 3.8)^2 = 0.92$$

Určíme hodnotu statistiky:

$$t = \frac{(n-1)s^2}{\sigma_0^2} = \frac{10 \cdot 0.92}{0.36} = 25.5556$$

Jelikož jde o jednostrannou alternativu, bude kritický obor:

$$W = (\chi^2_{1-\alpha}(n-1), +\infty) = (18.31, +\infty)$$

Protože $t \in W$, tak uvedenou hypotézu na hladině významnosti $\alpha = 5\%$ zamítáme a přijmeme alternativní hypotézu. Příze je tedy nevyhovující.

6. Pro náhodný výběr o rozsahu 16 z rozdělení $N(\mu_1, \sigma_1^2)$ byl zjištěn výběrový rozptyl $s_1^2 = 1.8$ a pro náhodný výběr o rozsahu 30 z rozdělení $N(\mu_2, \sigma_2^2)$ byl zjištěn výběrový rozptyl $s_2^2 = 2.4$ Na hladině významnosti $\alpha = 5\%$ testujeme hypotézu:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

Oproti hypotéze:

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Testovací statistika (v případě platnosti H_0 by měla nabývat hodnot kolem 1):

$$Z = \frac{s_1^2}{s_2^2} = \frac{1.8}{2.4} = 0.75$$

Kritický obor:
Pro Fisherovo rozdělení opět existují tabulky, ve kterých lze příslušné kvantily vyhledat, popř. v Excelu pomocí funkce FINV.

$$W = (-\infty, F_{\frac{\alpha}{2}}(n-1, m-1)) \cup (F_{1-\frac{\alpha}{2}}(n-1, m-1), +\infty) = (-\infty, 0.3771) \cup (2.3248, +\infty)$$

Jelikož $Z \notin W$, tak uvedenou hypotézu o shodě rozptylů na hladině významnosti $\alpha = 5\%$ nezamítáme.

11.3 Literatura s dalšími příklady

- Brousek, Jan – Ryjáček, Zdeněk: Sbírka řešených příkladů z počtu pravděpodobnosti.

12 Cvičení 12 - χ^2 test dobré shody, kontingenční tabulky, kovariance a korelace

12.1 Teoretická část

12.1.1 χ^2 test dobré shody

Jedná se o jeden z testů dobré shody (další např. Kolmogorovovův test a Lillieforsův test). Slouží k tomu, abychom pro náhodný výběr o rozsahu n z rozdělení nějaké náhodné veličiny X ověřili (na hladině významnosti α) hypotézu, že se řídí určitým rozdělením, až na hodnotu m neznámých parametrů. Postup je následující:

- Rozdělíme obor hodnot náhodné veličiny X na k nepřekrývajících se tříd.
- Zjistíme, kolik hodnot realizovaného náhodného výběru se nachází v jednotlivých třídách. Počty prvků v jednotlivých třídách označíme n_i
- Pokud je $m > 0$, tj. některé parametry rozdělení jsou neznámé, tak je odhadneme (k dispozici máme tedy po tomto kroku pravděpodobnosti p_i dané tímto rozdělením).
- Pro každou třídu spočteme očekávaný počet hodnot v této třídě, ozn. o_i. Platí $o_i = np_i$.
- V případě, že je v některé třídě počet očekávaných hodnot menší než 5, pak musíme tuto třídu sdružit s jinou.
- Testovací statistika má tvar
$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - o_i)^2}{o_i}$$
- Testovací statistika má za platnosti nulové hypotézy asymptoticky χ^2-rozdělení se stupněm volnosti $\nu = k - 1 - m$. Obor kritických hodnot W je pak dán:
$$W = (\chi^2_{1-\alpha}(\nu), +\infty)$$
kde $\chi^2_{1-\alpha}(\nu)$ je kvantil χ^2 rozdělení. Hodnoty kvantilů lze najít v tabulkách.
- Hypotézu, že se náhodná veličina řídí předpokládaným modelem, zamítáme na hladině významnosti α, je-li $\chi^2 \in W$.

12.1.2 Test nezávislosti v dvourozměrných kontingenčních tabulkách

Pomocí dvourozměrných kontingenčních tabulek lze testovat nezávislost dvou náhodných veličin X, Y. Test se používá především pro diskrétní náhodné veličiny. Počet variant náhodné veličiny X se označuje I a počet variant náhodné veličiny Y se označuje J. Je tedy $I \cdot J$ různých variant, kterých může dvourozměrná náhodná veličina (X, Y) nabývat. Četnosti v jednotlivých kategoriích se označují n_{ij}. Ukázka kontingenční tabulky:

	y_1	y_2	y_3	součty
x_1	n_{11}	n_{12}	n_{13}	$n_{1.}$
x_2	n_{21}	n_{22}	n_{23}	$n_{2.}$
součty	$n_{.1}$	$n_{.2}$	$n_{.3}$	n

Hodnota n je součet všech pozorování. Hodnoty $n_{i.}$, resp. $n_{.j}$ představují součty v jednotlivých řádcích, resp. sloupcích. Pomocí těchto hodnot lze vypočítat očekávané hodnoty v jednotlivých kategoriích:

$$o_{ij} = \frac{n_{i.}n_{.j}}{n}$$

Stejně jako v minulém testu musí platit $o_{ij} \geq 5$, pokud tomu tak není, tak musí být některé kategorie sloučeny. Vždy však musí platit $I, J \geq 2$. Testovací statistika má tvar:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - o_{ij})^2}{o_{ij}}$$

Testovací statistika má za platnosti nulové hypotézy (nezávislost obou veličin) asymptoticky χ^2-rozdělení se stupněm volnosti $\nu = (I-1)(J-1)$. Obor kritických hodnot W je pak dán:

$$W = (\chi^2_{1-\alpha}(\nu), +\infty)$$

Hypotézu, že se náhodné veličiny jsou nezávislé, zamítáme na hladině významnosti α, je-li $\chi^2 \in W$.

12.1.3 Kovariance

Kovariance dvou náhodných veličin X, Y se označuje $cov(X,Y)$, popř. σ_{XY}. Je definována takto:

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

Výpočetní tvar kovariance:

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

Výběrová kovariance (statistický odhad kovariance):

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} \sum_{i=1}^{n} x_i y_i - \frac{n}{n-1} \bar{x}\bar{y}$$

Pro kovaranci platí:

$$cov(X,Y) = cov(Y,X)$$

$$cov(X,X) = D(X)$$

Pokud jsou veličiny nezávislé, tak platí $E(X \cdot Y) = E(X)E(Y)$ a tedy $cov(X,Y) = 0$. POZOR, toto nelze obrátit, pokud je $cov(X,Y) = 0$, tak z toho neplyne, že náhodné veličiny jsou nezávislé!

12.1.4 Korelace

Korelace dvou náhodných veličin X, Y se označuje ρ_{XY}. Je definována takto:

$$\rho_{XY} = \frac{cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

Výběrová korelace (statistický odhad korelace):

$$r_{XY} = \frac{S_{XY}}{\sqrt{S^2(X)}\sqrt{S^2(Y)}}$$

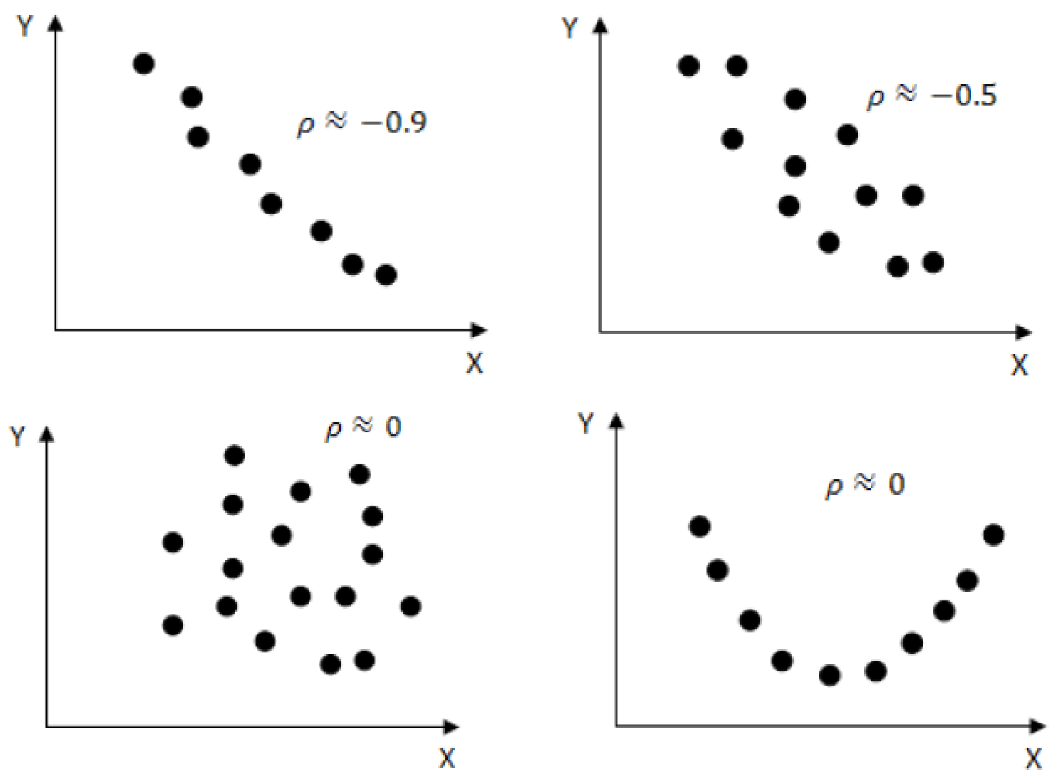
Pro korelaci platí:

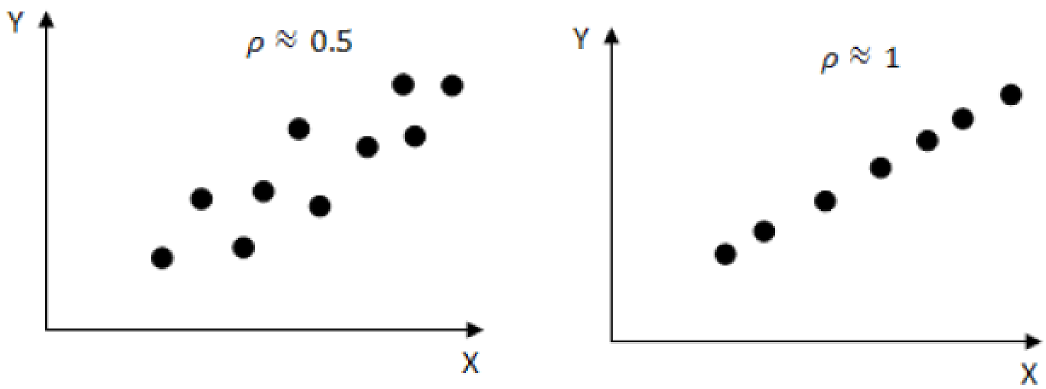
$$\rho_{XY} = \rho_{YX}$$

$$\rho_{XX} = 1$$

$$\rho \in [-1, 1]$$

Korelace vyjadřuje míru lineární závislosti mezi X a Y. Mezní hodnoty (-1 a +1) nastávají, pokud všechny body (x_i, y_i) leží na přímce. Pokud jsou veličiny nezávislé, tak platí $cov(X, Y) = 0$ a tedy $\rho_{XY} = 0$. POZOR, toto nelze obrátit, pokud je $\rho_{XY} = 0$, tak z toho neplyne, že náhodné veličiny jsou nezávislé! Ukázka hodnot korelace:





12.1.5 Test nezávislosti

V případě, že (X,Y) má dvourozměrné normální rozdělení pravděpodobnosti, pak X a Y jsou nezávislé, pokud $\rho = 0$ (výše bylo uvedeno, že z $\rho_{XY} = 0$ neplyne nezávislost, pokud ovšem (X,Y) má dvourozměrné normální rozdělení, tak výroky "$\rho_{XY} = 0$" a "X a Y jsou nezávislé" jsou ekvivalentní). Při testu nezávislosti dvou náhodných veličin s dvourozměrným normálním rozdělením se testuje $H_0 : \rho = 0$ oproti $H_1 : \rho \neq 0$ (popř. jednostranná alternativa). Testovací statistika má tvar:

$$T = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

kde r je výběrový korelační koeficient. Testovací statistika má za platnosti nulové hypotézy Studentovo t-rozdělení s počtem stupňů volnosti $\nu = n-2$. Obor kritických hodnot pro test na hladině významnosti α je:

$$W = (-\infty, t_{\frac{\alpha}{2}}(n-2)) \cup (t_{1-\frac{\alpha}{2}}(n-2), +\infty)$$

Hypotéza o nezávislosti se zamítá, pokud $T \in W$.

12.2 Příklady

1. Chceme testovat, zda hrací kostka je korektní. Provedli jsme 600x hod kostkou a získali jsme následující četnosti:

Číslo	1	2	3	4	5	6
n_i	122	61	98	115	79	125

Pokud je kostka korektní, měly by se očekávané četnosti řídit diskrétním rovnoměrným rozdělením. Budeme tedy testovat shodu získaných hodnot s diskrétním rovnoměrným rozdělením na hladině významnosti 5%.
Řešení:
H_0: Kostka je korektní
H_1: Kostka není korektní
Budeme se řídit postupem uvedeným v první části tohoto cvičení:

- Obor hodnot je již rozdělen na 6 nepřekrývajících se tříd, tedy $k = 6$.
- Počty prvků n_i jsou uvedeny již v zadání.
- Není potřeba odhadovat parametry, tj. $m = 0$.
- Spočteme očekávané hodnoty v jednotlivých třídách $o_i = np_i = 600 \cdot \frac{1}{6} = 100$ pro $i = 1, 2, ..., 6$
- V žádné třídě není $o_i < 5$, nebudeme tedy žádné třídy slučovat.
- Vypočteme hodnotu testovací statistiky:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - o_i)^2}{o_i} = \chi^2 = \sum_{i=1}^{6} \frac{(n_i - 100)^2}{100} = 33$$

- Kritický obor je dán χ^2-rozdělením s $\nu = k - 1 = 5$ stupni volnosti:

$$W = (\chi^2_{0.95}(5), +\infty) = (11.1, +\infty)$$

- Jelikož $\chi^2 \in W$, tak hypotézu o tom, že kostka je korektní zamítáme (na hladině významnosti $\alpha = 5\%$.

2. Po provedení 60 pokusů s diskrétní náhodnou veličinou X, která může nabývat hodnot 0 až 4 (tj. v každém z pokusů nastane buď 0, 1, 2, 3 nebo 4krát sledovaný jev) jsou získány následující četnosti.

Hodnota	0	1	2	3	4
n_i	3	12	21	20	4

Tedy například hodnota 12 znamená, že při 12 pokusech z 60 nabyla náhodná veličina X hodnoty 1. Otestujte na hladině významnosti $\alpha = 2.5\%$, zda se náhodná veličina X řídí binomickým rozdělením.

Řešení:

H_0: Náhodná veličina se řídí binomickým rozdělením

H_1: Náhodná veličina se neřídí binomickým rozdělením

Budeme se řídit postupem uvedeným v první části tohoto cvičení:

- Obor hodnot je již rozdělen na 5 nepřekrývajících se tříd, tedy $k = 5$.
- Počty prvků n_i jsou uvedeny již v zadání.
- Ze zadání víme, že parametr n binomického rozdělení je 4, ten tedy odhadovat nemusíme. Je ale potřeba odhadnout parametr p binomického rozdělení. Ten lze odhadnout přes střední hodnotu. U binomického rozdělení víme, že $E(X) = np$. n známe, střední hodnotu lze odhadnout pomocí průměru a pak již jen vyjádříme neznámý parametr p:

  $$\bar{x} = \frac{3 \cdot 0 + 12 \cdot 1 + 21 \cdot 2 + 20 \cdot 3 + 4 \cdot 4}{60} = 2.1667$$

  Dosadíme:

  $$2.1667 = 4 \cdot \hat{p}$$

  A odtud:

  $$\hat{p} = 0.5417$$

  Předpokládáme, že náhodná veličina se řídí rozdělením $Bi(4, 0.5417)$. Odhadovali jsme jeden parametr, takže $m = 1$.
- Spočteme očekávané pravděpodobnosti p_i a následně očekávané hodnoty v jednotlivých třídách $o_i = np_i$ pro $i = 0, 1, ..., 4$:

Hodnota	0	1	2	3	4
p_i	0.0441	0.2086	0.3698	0.2914	0.0861
o_i	2.65	12.51	22.19	17.48	5.17

- V první třídě je $o_i < 5$, sloučíme tedy tuto třídu se sousední. V poslední třídě je sice $n_i < 5$, ale očekávaná hodnota splňuje podmínku a slučovat tedy nebudeme. Po sloučení obdržíme:

Hodnota	0 a 1	2	3	4
n_i	15	21	20	4
o_i	15.16	22.19	17.48	5.17

  Stejným způsobem musí být sloučeny i naměřené hodnoty.

- Vypočteme hodnotu testovací statistiky:

  $$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - o_i)^2}{o_i} = 0.6936$$

- Kritický obor je dán χ^2-rozdělením s $\nu = k - 1 - m = 2$ stupni volnosti:

$$W = (\chi^2_{0.975}(2), +\infty) = (7.38, +\infty)$$

- Jelikož $\chi^2 \notin W$, tak hypotézu o tom, že náhodná veličina se řídí rozdělením Bi(4, 0.5417) (na hladině významnosti $\alpha = 2.5\%$) nezamítáme.

3. Z průzkumu provedeného u 1 000 osob, který měl zjistit efektivnost očkování proti chřipce, byly získány tyto výsledky:

	Bez očkování	Jedno očkování	Dvě očkování	Celkem
Chřipka	24	9	13	46
Bez chřipky	289	100	565	954
Celkem	313	109	578	1 000

Na hladině významnosti $\alpha = 5\%$ testujte, zda má očkování vliv na výskyt chřipky. **Řešení:**
H_0: Očkování vliv nemá (veličiny jsou nezávislé)
H_1: Očkování vliv má (mezi veličinami existuje závislost)
Použijeme tedy test nezávislosti:
Hodnoty $n, n_{i.}$ a $n_{.j}$ jsou uvedeny již v tabulce. Pomocí těchto hodnot vypočteme očekávané hodnoty:

$$o_{ij} = \frac{n_{i.}n_{.j}}{n}$$

Např.:

$$o_{12} = \frac{n_{1.}n_{.2}}{n} = \frac{46 \cdot 109}{1000} = 5.014$$

Celá tabulka s očekávanými hodnotami:

	Bez očkování	Jedno očkování	Dvě očkování
Chřipka	14.40	5.01	26.59
Bez chřipky	298.60	103.99	551.41

Ve všech kategoriích platí $o_{ij} \geq 5$.
Testovací statistika:

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(n_{ij} - o_{ij})^2}{o_{ij}} = 17.32$$

Obor kritických hodnot W:

$$W = (\chi^2_{0.95}(1 \cdot 2), +\infty) = (5.99; +\infty)$$

Protože $\chi^2 \in W$, tak hypotézu o nezávislosti (na hladině významnosti $\alpha = 5\%$) zamítáme a očkování má tedy vliv.

4. Chceme otestovat vliv nové technologie. Máme k dispozici následující výsledky:

	I. jakost	II. jakost	III. jakost	Zmetek	Celkem
Stará technologie	503	105	33	7	648
Nová technologie	553	95	35	3	686
Celkem	1 056	200	68	10	1334

Na hladině významnosti $\alpha = 5\%$ testujte, zda má nová technologie vliv na výrobu. **Řešení:**
H_0: Technologie nemá vliv (veličiny jsou nezávislé)
H_1: Technologie má vliv (mezi veličinami existuje závislost)
Použijeme tedy test nezávislosti v dvourozměrné kontingenční tabulce:
Hodnoty $n, n_{i.}$ a $n_{.j}$ jsou uvedeny již v tabulce. Pomocí těchto hodnot vypočteme očekávané hodnoty:

	I. jakost	II. jakost	III. jakost	Zmetek
Stará technologie	512.96	97.15	33.03	4.86
Nová technologie	543.03	102.85	34.97	5.14

Jelikož $o_{14} < 5$, tak musíme sloučit poslední dva sloupce (řádky slučovat nemůžeme, musí platit $I, J \geq 2$). Máme tedy:

	I. jakost	II. jakost	III. jakost + Zmetek
Stará technologie	512.96	97.15	37.89
Nová technologie	543.03	102.85	40.11

Stejným způsobem musí být sloučeny i naměřené hodnoty.
Testovací statistika:

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(n_{ij} - o_{ij})^2}{o_{ij}} = 1.84$$

Obor kritických hodnot W:

$$W = (\chi^2_{0.95}(1 \cdot 2), +\infty) = (5.99; +\infty)$$

Protože $\chi^2 \notin W$, tak hypotézu o nezávislosti (na hladině významnosti $\alpha = 5\%$) nezamítáme a nová technologie tedy nemá vliv.

5. U 5 lidí byla zjišťována váha (ozn. X) a výška (ozn. Y). Výsledky jsou následující:

Výška	170	183	192	164	196
Váha	70	72	88	60	82

Předpokládáme, že dvourozměrná náhodná veličina (X, Y) má dvourozměrné normální rozdělení. Otestujte na hladině významnosti $\alpha = 10\%$, zda jsou X a Y nezávislé. **Řešení:**
Jelikož se jedná o dvourozměrné normální rozdělení, tak stačí testovat nulovost korelačního koeficientu. Testujeme tedy:

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

Musíme vypočítat průměry, výběrové rozptyly, hodnotu výběrové kovariance a následně výběrové korelace:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = 181$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = 74.4$$

$$S_x^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 = 190$$

$$S_y^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 118.8$$

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} x_i y_i - \frac{n}{n-1} \bar{x}\bar{y} = \frac{1}{4} \cdot 67884 - \frac{5}{4} \cdot 181 \cdot 74.4 = 138$$

$$r_{XY} = \frac{S_{XY}}{\sqrt{S^2(X)}\sqrt{S^2(Y)}} = \frac{138}{\sqrt{190}\sqrt{118.8}} = 0.9185$$

Testovací statistika má tvar:

$$T = \frac{r}{\sqrt{1 - r^2}} \sqrt{n-2} = \frac{0.9185}{\sqrt{1 - 0.9185^2}} \sqrt{5 - 2} = 4.0242$$

Obor kritických hodnot pro test na hladině významnosti $\alpha = 10\%$ je:

$$W = (-\infty, -2.353) \cup (2.353, +\infty)$$

Hypotézu o nezávislosti lze zamítnout na hladině významnosti $\alpha = 10\%$, protože $T \in W$. Přijmeme tedy alternativní hypotézu, že veličiny jsou závislé.

12.3 Literatura s dalšími příklady

- Brousek, Jan – Ryjáček, Zdeněk: Sbírka řešených příkladů z počtu pravděpodobnosti.

13 CVičení 13 - Regresní analýza. Jednoduchá a vícenásobná regrese. Koeficient determinace.

13.1 Teoretická část

Regrese je snad nejčastěji používaná statistická metoda. Regrese se zabývá problémem vysvětlení změn jedné náhodné veličiny (vysvětlovaná, závislá , endogenní proměnná, regresand) na jedné nebo více jiných veličinách (regresory, vysvětlující proměnné, exogenní proměnné). V případě, že závislost je popsána lineárními vztahy, mluvíme o lineárním regresním modelu. Pokud modelujeme chování vysvětlované proměnné pomocí jedné vysvětlující proměnné, mluvíme o jednoduché regresi, v opačném případě se jedná o regresi vícenásobnou.

Označme $\boldsymbol{X}$ nezávisle proměnné a Y závislou proměnnou. Regresní funkcí se pak rozumí $\mu(x) = E\left(Y|\boldsymbol{X} = \boldsymbol{x}\right)$. Regresní funkce tedy udává, jaká je střední hodnota náhodné veličiny Y při dané hodnotě x.

V dalším textu tedy budeme pracovat s modelem

$$Y_i = f(\boldsymbol{X}, \beta_1, \ldots, \beta_k) + \varepsilon_i,$$

kde

- $\beta_1 \ldots, \beta_k$ jsou neznámé parametry modelu; počet parametrů je k;
- ε_i jsou náhodné veličiny, který modelují nesystematické chyby měření;
- $\boldsymbol{X}$ je matice nezávislých proměnných;
- Y_i jsou náhodné veličiny reprezentující vysvětlovanou proměnnou.

Předpokládejme, že máme k dispozici naměřené hodnoty $y_1, y_2, \ldots, y_n$ pro jednotlivé kombinace vysvětlujících proměnných $(x_{11}, x_{12}, \ldots, x_{1k}), (x_{21}, x_{22}, \ldots, x_{2k}), \ldots, (x_{n1}, x_{n2}, \ldots, x_{nk})$.

Cílem regresní analýzy je odhadnout parametry $\beta_1 \ldots, \beta_k$ tak, aby $f(\boldsymbol{X}, \widehat{\beta_1}, \ldots, \widehat{\beta_k})$ „co nejvíce odpovídala k empiricky naměřeným hodnotám y_i“.

Funkce $y = f(\boldsymbol{,}\beta_1, \ldots, \beta_k)$ se nazývá teoretická regresní funkce závislosti proměnné y na $\boldsymbol{X}$, její grafické vyjádření se nazývá teoretická regresní křivka. Regresní funkce, v níž jsou nahrazeny neznámé parametry β jejich odhady $\widehat{\beta}$ (resp. b) se nazývá empirická regresní funkce a její grafický obraz je empirická regresní křivka.

Pro hodnoty $\boldsymbol{X}$ můžeme na základě empirické regresní křivky určit hodnotu $\widehat{y_i} = f(\boldsymbol{X}, \widehat{\beta_1}, \ldots, \widehat{\beta_k})$, tyto hodnoty nazýváme vyrovnanými hodnotami y_i a rozdíl mezi $y_i - \widehat{y_i}$ nazýváme chyby odhadu (značíme e_i).

13.1.1 Jednoduchá a vícenásobná regrese

O jednoduché regresi mluvíme v situacích, kdy uvažujme tento základní jednoduchý model :

$$Y_i = \beta_1 + \beta_2 \cdot x_i + \varepsilon_i,$$

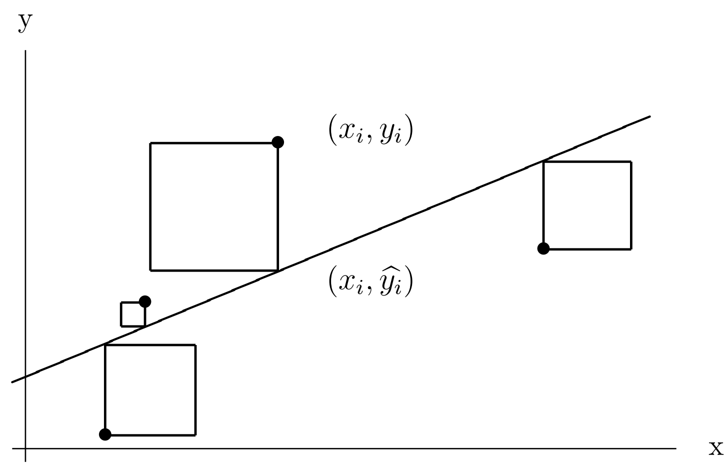
kde β_0, β_1 jsou neznámé parametry, které odhadujeme a ε_i jsou neznámé náhodné odchylky, které splňují následující podmínky:

- $E(\varepsilon_i) = 0$ pro $i = 1, 2, \ldots, n$
- $D(\varepsilon_i) = \sigma^2$ pro $i = 1, 2, \ldots, n$ a $\sigma^2 > 0$ je neznámá konstanta (homoskedasticita)
- ε_i jsou nezávislé pro $i = 1, 2, \ldots, n$

Princip metody nejmenších čtverců (MNČ) je založen na jednoduchém volbě optimalizačního kritéria, kdy minimalizuji kvadrát odchylek naměřených y_i a vyrovnaných hodnot $\widehat{y_i}$. Název MNČ se odvozuje od toho, že se při této metodě minimalizuje součet druhých mocnin typu:

$$SSE = \sum_{i=1}^{n} (b_0 + b_1 x_i - y_i)^2$$

Graficky lze MNČ znázornit následujícím způsobem



V případě vícenásobná regrese pracujeme s modelem

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i,$$

kdy se snažíme vysvětlit proměnnou Y pomocí více vysvětlujících proměnných $x_1, x_2, \ldots, x_k$.
U vícenásobné regrese minimalizujeme výraz

$$SSE = \sum_{i=1}^{n} - (b_0 + b_1 x_{1i} + \cdots + b_k x_{ki} - y_i)^2$$

Při minimalizaci výrazu SSE, který chápeme jako funkci proměnných β_j, vycházíme ze známého faktu, že funkce nabývá svého minima v bodech, kdy derivace je rovna nule, tj. při hledání minima řešíme soustavu p lineárních rovnic tvaru

$$\left.\frac{\partial Q}{\partial \beta_k}\right|_{\beta_k = b_k} = 0 \qquad \text{pro } k = 0, 1, \ldots, k$$

Takto vzniklou soustavu $k + 1$ rovnic nazýváme soustavou normálních rovnic.
Soustava normálních rovnic pro jednoduchou regresi má tedy tvar

$$
\begin{array}{llll}
b_0\cdot & n & +b_1\cdot & \sum_{i=1}^{n} x_i & = \sum_{i=1}^{n} y_i \\
b_0\cdot & \sum_{i=1}^{n} x_i & +b_1\cdot & \sum_{i=1}^{n} x_i^2 & = \sum_{i=1}^{n} x_i\, y_i
\end{array}
$$

a řešením výše uvedených soustav dostáváme příslušné odhady b_0 a b_1, které minimalizují výraz SSE:

$$
b_0 = \widehat{\beta}_0 = \frac{\sum\limits_{i=1}^{n} y_i \sum\limits_{i=1}^{n} x_i^2 - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} x_i\, y_i}{n \cdot \sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2}
$$

$$
b_1 = \widehat{\beta}_1 = \frac{n \cdot \sum\limits_{i=1}^{n} x_i\, y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{n \cdot \sum\limits_{i=1}^{T} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2}
$$

V praktických příkladech metodu nejmenších čtverců pro regresní analýzu zpracováváme pomocí vhodného softwaru. Například je v Excelu lze lineární regresi a metodu nejmenších čtverců aplikovat pomocí funkce LINREGRESE, popř. lze užít doplňku Analýza dat, v něm pak Regrese.

13.1.2 Maticový zápis regrese a metody nejmenších čtverců

Budeme uvažovat následující maticový zápis

$$
\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \vdots \\ \varepsilon_n \end{bmatrix}
$$

kde

- $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_n)^T$ je vektor hodnot vysvětlované proměnné;
- $\boldsymbol{X} = [x_{ij}]_{i=1,\dots,n;\; j=1,\dots,k}$ je matice typu $n \times k$ hodnot vysvětlující proměnné;
- $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_k)^T$ je vektor hledaných k neznámých parametrů;
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)^T$ je vektor náhodné složky.

Pokud první sloupec matice $\boldsymbol{X}$ jsou jednotky, mluvíme o lineárním regresním modelu s absolutním členem.

Model tedy můžeme zapsat v maticovém vyjádření

$$
\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}
$$

a předpoklady pro metodu nejmenších čtverců lze také zapsat v maticovém tvaru

Předpoklady řešení pomocí metody nejmenších čtverců

(P1) $E(\boldsymbol{\varepsilon}) = \mathbf{0}$;

(P2) $var(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_T$;

(P3) v některých případech uvažujeme též silnější podmínku zahrnující předcházející $\boldsymbol{\varepsilon} \sim N_n\left(0; \sigma^2 \boldsymbol{I}_T\right)$.

(P4) $\boldsymbol{X}$ je nestochastická matice, která má plnou hodnost.

Neznámé parametry lze v maticovém zápisu odhadnout podle vztahu takto:

$$\boldsymbol{b} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{Y}$$

13.1.3 Hodnocení kvality regrese a koeficient determinace R^2

Kvalitu regresního vztahu lze poměřovat podle toho, jak odhadnuté hodnoty $\widehat{y_i}$ odpovídají realizacím y_i. Kvalitu odhadu však významným vlivem ovlivňuje též variabilita dat, resp. variabilita náhodné složky modelu.

Při hodnocení modelu vycházíme především ze získaných reziduí $e_i = \widehat{y_i} - y_i$, které zachycují rozdíl mezi naměřenou a vyrovnanou hodnotou.

Pro naměřené y_i a vyrovnané $\widehat{y_i}$ hodnoty vysvětlované proměnné obvykle počítáme

celkový součet čtverců $S_T^2 = \sum_{i=1}^{n} (y_i - \overline{y})^2$

vysvětlený (regresní) součet čtverců $S_V^2 = \sum_{i=1}^{n} (\widehat{y_i} - \overline{y})^2$

nevysvětlený (residuální) součet čtverců $SSE = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$

Při použití metody nejmenších čtverců platí $S_T^2 = S_V^2 + SSE$. Jako vhodnější je volen ten model, který má menší hodnotu nevysvětlených (residuálních) součtů čtverců.

Na základě výše uvedených součtů čtverců lze pro vícenásobnou regresi s absolutním členem určit **koeficient determinace** R^2 podle vzorce

$$R^2 = \frac{S_V^2}{S_T^2}$$

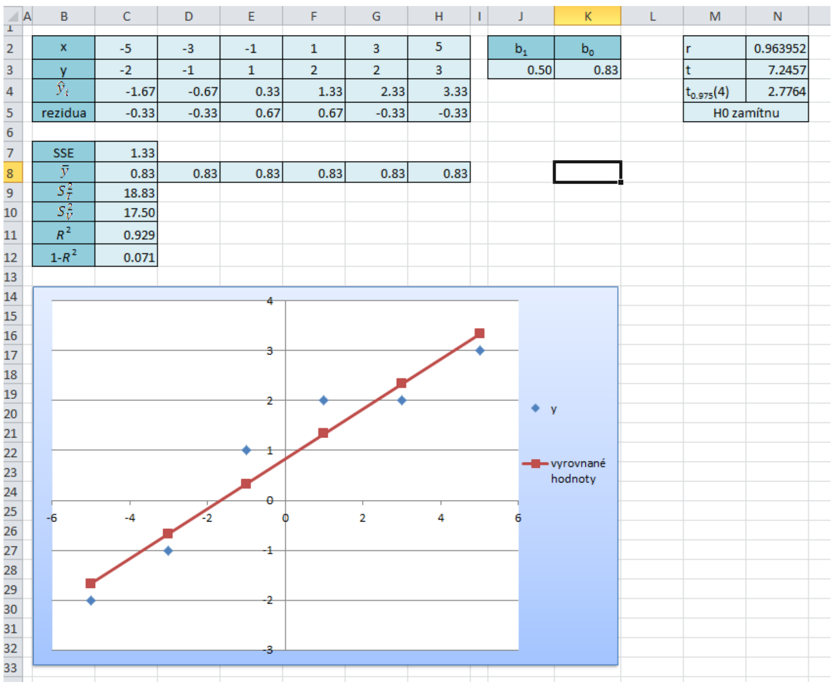
Pro koeficient determinace platí $R^2 \in [0, 1]$. Jde o podíl rozptylu hodnot y_i, který se podařilo vysvětlit pomocí regresního modelu.

Hodnota $1 - R^2$ určuje podíl rozptylu hodnot y_i, který se vysvětlit nepodařilo.

13.2 Příklady

1. Pro následující data odhadněte koeficienty regresní přímky $y = \beta_0 + \beta_1 x$, vypočtěte přes soustavu normálních rovnic.

x	-5	-3	-1	1	3	5
y	-2	-1	1	2	2	3



x	-5	-3	-1	1	3	5
y	-2	-1	1	2	2	3
$\hat{y}_i$	-1.67	-0.67	0.33	1.33	2.33	3.33
rezidua	-0.33	-0.33	0.67	0.67	-0.33	-0.33
SSE	1.33					
$\bar{y}$	0.83	0.83	0.83	0.83	0.83	0.83
S_T^2	18.83					
S_V^2	17.50					
R^2	0.929					
$1-R^2$	0.071					

b_1	b_0
0.50	0.83

r	0.963952
t	7.2457
$t_{0.975}(4)$	2.7764
H0 zamítnu	

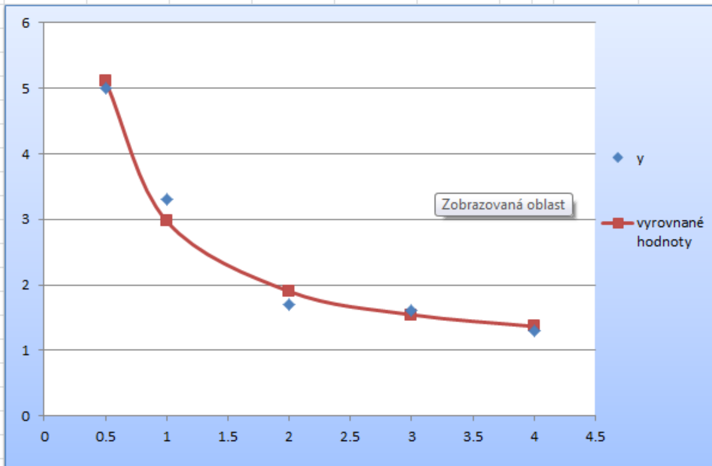
2. Pro následující data odhadněte koeficienty regresní funkce $y = \beta_0 + \beta_1 \dfrac{1}{x}$, vypočtěte přes soustavu normálních rovnic.

x	0.5	1	2	3	4
y	5.0	3.3	1.7	1.6	1.3

x	0.5	1	2	3	4
1/x	2.00	1.00	0.50	0.33	0.25
y	5	3.3	1.7	1.6	1.3
$\hat{y}_i$	5.11	2.97	1.90	1.54	1.37
rezidua	-0.11	0.33	-0.20	0.06	-0.07

b_1	b_0
2.14	0.83

SSE	0.17				
$\bar{y}$	2.58	2.58	2.58	2.58	2.58
S_T^2	9.75				
$S_{\hat{Y}}^2$	9.58				
R^2	0.983				
$1-R^2$	0.017				

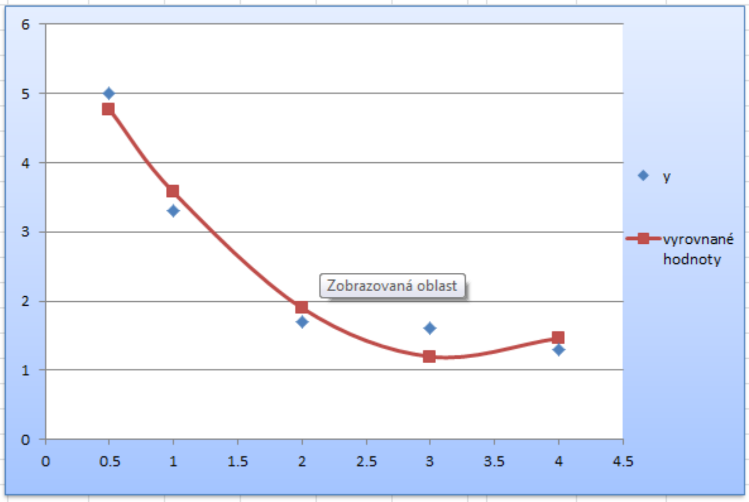


3. Pro data z předchozího příkladu odhadněte koeficienty regresní funkce $y = \beta_0 + \beta_1 x + \beta_2 x^2$

x	0.5	1	2	3	4
x^2	0.25	1	4	9	16
y	5	3.3	1.7	1.6	1.3
$\hat{y}_i$	4.77	3.57	1.90	1.20	1.46
rezidua	0.23	-0.27	-0.20	0.40	-0.16

b_2	b_1	b_0
0.48	-3.12	6.21

SSE	0.36				
$\bar{y}$	2.58	2.58	2.58	2.58	2.58
S_T^2	9.75				
S_V^2	9.39				
R^2	0.964				
$1-R^2$	0.036				



4. Pro předchozí příklady spočtěte S_V^2, S_T^2, SSE a R^2. Získané výsledky interpretujte.

13.3 Literatura s dalšími příklady

- Brousek, Jan – Ryjáček, Zdeněk: Sbírka řešených příkladů z počtu pravděpodobnosti.