Pravděpodobnost a statistika – slajdy ke zkoušce

Miroslav Kolařík

Aktualizováno: 29. listopadu 2023

Použitá literatura

Studijní literatura, na které jsou postaveny tyto slajdy[1]:

1. L. Wasserman, *All of Statistics–A Concise Course in Statistical Inference*
2. S. Ross, *A First Course in Probability*
3. M. Mitzenmacher & E. Upfal, *Probability and Computing*
4. J. Slovák, M. Panák, M. Bulant & kolektiv, *Matematika drsně a svižně*

[1] Slajdy vytvořil Tomáš Masopust. S jeho svolením je upravil Miroslav Kolařík.

Pravděpodobnost

- Pravděpodobnost je matematický jazyk pro určení nejistoty.
- Zavedeme základní koncepty teorie pravděpodobnosti.

Výběrové prostory a jevy

Prostor elementárních jevů

- Výběrový prostor či prostor elementárních jevů Ω je množina možných výsledků náhodného pokusu.
- Prvky $\omega \in \Omega$ se nazývají elementární jevy.
- Podmnožiny Ω se nazývají (náhodné) jevy.

Příklad 1.

- Náhodný pokus = hod dvakrát mincí.
- Pak $\Omega = \{OO, OP, PO, PP\}$,
  kde „O“ je orel a „P“ je panna.
- Jev, že „na první hod padne orel“ je $A = \{OO, OP\}$.

Příklad 2.

- Nechť ω je výsledek měření nějaké fyzikální veličiny, například teploty, pak $\Omega = \mathbb{R} = (-\infty, +\infty)$.
- Zde $\Omega = \mathbb{R}$ není přesné, protože teplota má dolní hranici. Obvykle není na škodu vzít výběrový prostor větší než je potřeba.
- Jev „teplota je větší než 10 a menší nebo rovna 23“ je $A = (10, 23]$.

Příklad 3.

Jestliže házíme mincí do nekonečna, pak výběrový prostor je nekonečná množina

$$\Omega = \{\omega = (\omega_1, \omega_2, \ldots) \mid \omega_i \in \{O, P\}\}.$$

Nechť E je jev „první orel se objeví na třetí hod“. Pak

$$E = \{(\omega_1, \omega_2, \omega_3, \omega_4, \ldots) \mid \omega_1 = P, \omega_2 = P, \omega_3 = O, \omega_i \in \{O, P\} \text{ pro } i > 3\}.$$

Operace s jevy

- Pro jev A je $A^c = \{\omega \in \Omega \mid \omega \notin A\}$ jev, tzv. komplement jevu A.
  Komplement Ω je prázdná množina $\emptyset$.
- Sjednocení jevů A a B je jev

$$A \cup B = \{\omega \in \Omega \mid \omega \in A \text{ nebo } \omega \in B\}.$$

  Pro jevy $A_1, A_2, \ldots$ je $\bigcup_{i=1}^{+\infty} A_i = \{\omega \in \Omega \mid \exists i \text{ tak, že } \omega \in A_i\}$ jev.
- Průnik jevů A a B je jev

$$A \cap B = \{\omega \in \Omega \mid \omega \in A \text{ a } \omega \in B\}.$$

  $A \cap B$ budeme též zapisovat jako AB.
  Pro jevy $A_1, A_2, \ldots$ je $\bigcap_{i=1}^{+\infty} A_i = \{\omega \in \Omega \mid \omega \in A_i \text{ pro všechna } i\}$ jev.

- Rozdíl jevů je jev $A - B = \{\omega \mid \omega \in A, \omega \notin B\}$.
- Jestliže kadý prvek A je také v B, píšeme $A \subseteq B$ či $B \supseteq A$.
  Pro vlastní podmnožiny se používá značení $A \subset B$ či $B \supset A$.
- Jestliže A je konečná množina, značí $|A|$ počet prvků A.

Používané značení

Ω	výběrový prostor
ω	elementární jev (bod nebo prvek)
A	jev (podmnožina Ω)
A^c	komplement A
$A \cup B$	sjednocení
$A \cap B$ nebo AB	průnik
$A - B$	množinový rozdíl
$A \subseteq B$	množinová inkluze
$\emptyset$	nemožný jev
Ω	jistý jev

Disjunktní jevy, rozklad

- Jevy $A_1, A_2, \ldots$ jsou disjunktní nebo vzájemně neslučitelné, jestliže

$$A_i \cap A_j = \emptyset$$

  kdykoliv $i \neq j$.
- Například jevy $A_1 = [0, 1), A_2 = [1, 2), A_3 = [2, 3), \ldots$ jsou disjunktní.

- Rozklad Ω je posloupnost disjunktních množin $A_1, A_2, \ldots$ takových, že

$$\bigcup_i A_i = \Omega.$$

Pravděpodobnost

Definice 4.

Funkce $\mathbb{P}$ přiřazující reálné číslo $\mathbb{P}(A)$ každému jevu A je pravděpodobnostní míra, jestliže splňuje následující tři axiomy:

Axiom 1: $\mathbb{P}(A) \geq 0$ pro každý jev A

Axiom 2: $\mathbb{P}(\Omega) = 1$

Axiom 3: Jestliže $A_1, A_2, \ldots$ jsou disjunktní jevy, pak

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

Pozn.: Axiom 3 zahrnuje i případ, kdy jsou skoro všechna $A_i = \emptyset$.

Algebra jevů

- Obecně není možné přiřadit pravděpodobnost všem podmnožinám Ω (pokud je Ω nespočetná).
- Omezíme se proto na množinu jevů nazývanou σ-algebra, tedy na třídu $\mathcal{A}$ splňující
  - $\emptyset \in \mathcal{A}$
  - Jestliže $A_1, A_2, \ldots \in \mathcal{A}$, pak $\bigcup_{i=1}^{+\infty} A_i \in \mathcal{A}$
  - Jestliže $A \in \mathcal{A}$, pak též $A^c \in \mathcal{A}$.
- Množiny v $\mathcal{A}$ se nazývají měřitelné a dvojice $(\Omega, \mathcal{A})$ je tzv. měřitelný prostor.
- Je-li $\mathbb{P}$ pravděpodobnostní míra na $\mathcal{A}$, je trojice $(\Omega, \mathcal{A}, \mathbb{P})$ pravděpodobnostní prostor.
- Pokud je Ω reálná osa, vezmeme $\mathcal{A}$ jako nejmenší σ-algebru, která obsahuje všechny otevřené podmnožiny, tzv. Borelovská σ-algebra.
  - My se pro jednoduchost omezíme na případ, kdy otevřené množiny „znamená“ intervaly reálných čísel.

- Existuje mnoho interpretací pravděpodobnostní míry $\mathbb{P}(A)$.
  - Dvě základní jsou frekvence a stupeň důvěry.

- Frekvence: $\mathbb{P}(A)$ vyjadřuje poměr, kolikrát je A slpňeno při dlouhodobém opakování pokusu.
  - Například „pravděpodobnost, že padne orel je $1/2$“ znamená, že s rostoucím počtem hodů mincí „jde“ poměr hozených orlů vzhledem ke všem pokusům k $1/2$.
  - Nekonečně dlouhá, nepředvídatelná posloupnost hodů, jejíž mezní podíl směřuje ke konstantě, je idealizací, podobně jako představa přímky v geometrii.

- Stupeň důvěry: $\mathbb{P}(A)$ určuje pozorovatelovu intenzitu důvěry, že A je splněno.

- V obou interpretacích vyžadujeme platnost Axiomů 1 až 3.
- Rozdíl mezi interpretacemi hraje roli až ve statistické inferenci:
  - frekvencionistická vs. Bayesovská škola.

Vlastnosti pravděpodobnosti

Z axiomů lze odvodit mnoho vlastností $\mathbb{P}$:

- $\mathbb{P}(\emptyset) = 0$
  - $1 =_{A2} \mathbb{P}(\Omega) = \mathbb{P}(\Omega \cup \emptyset) =_{A3} \mathbb{P}(\Omega) + \mathbb{P}(\emptyset) = 1 + \mathbb{P}(\emptyset)$
- $A \subseteq B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(\mathbb{B})$
  - $B = A \cup (B - A) \Rightarrow \mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B - A) \Rightarrow_{A1} \mathbb{P}(A) \leq \mathbb{P}(B)$
- $0 \leq \mathbb{P}(A) \leq 1$
  - použije se předchozí tvrzení pro $A \subseteq \Omega$
- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
  - $\Omega = A \cup A^c \Rightarrow 1 = \mathbb{P}(\Omega) = \mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c)$
- $A \cap B = \emptyset \Rightarrow \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$
  - z Axiomu 3 dosazením $A_1 = A$, $A_2 = B$, $A_3 = A_4 = \cdots = \emptyset$.

Vlastnosti $\mathbb{P}$

Lemma 5.

Pro libovolné jevy A a B je $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(AB)$.

Důkaz.

$A \cup B = (AB^c) \cup (AB) \cup (A^c B)$ a uvedené jevy jsou disjunktní. Opakovaným použitím Axiomu 3 dostáváme

$$\begin{aligned}
\mathbb{P}(A \cup B) &= \mathbb{P}((AB^c) \cup (AB) \cup (A^c B)) \\
&= \mathbb{P}(AB^c) + \mathbb{P}(AB) + \mathbb{P}(A^c B) \\
&= \mathbb{P}(AB^c) + \mathbb{P}(AB) + \mathbb{P}(A^c B) + \mathbb{P}(AB) - \mathbb{P}(AB) \\
&= \mathbb{P}((AB^c) \cup (AB)) + \mathbb{P}((A^c B) \cup (AB)) - \mathbb{P}(AB) \\
&= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(AB).
\end{aligned}$$

□

Příklad 6.

- Uvažme dva hody mincí.
- Označme H_1 jev „orel padne v prvním hodu“ a H_2 jev „orel padne ve druhém hodu“.
- Jsou-li všechny výsledky stejně pravděpodobné, pak

$$\begin{aligned}\mathbb{P}(H_1 \cup H_2) &= \mathbb{P}(H_1) + \mathbb{P}(H_2) - \mathbb{P}(H_1 H_2) \\ &= \frac{1}{2} + \frac{1}{2} - \frac{1}{4} \\ &= \frac{3}{4}.\end{aligned}$$

Konečný výběrový prostor

Konečné výběrové prostory

- Je-li $\Omega = \{\omega_1, \ldots, \omega_n\}$, je Ω konečný a $|\Omega| = n$.

- Například pro „hod dvakrát kostkou“ je $|\Omega| = 36$, kde $\Omega = \{(i, j) \mid i, j \in \{1, \ldots, 6\}\}$.
  - Pokud je každý výsledek stejně pravděpodobný, je

$$\mathbb{P}(A) = \frac{|A|}{36},$$

    kde $|A|$ je počet prvků jevu A.
  - Například pravděpodobnost, že padne součet 11 je $\frac{2}{36}$, protože existují dva výsledky se sumou 11.
    - Které?
    - Jaká je pravděpodobnost, že padne součet 12?

Pravděpodobnost na konečném výběrovém prostoru

- Pokud je Ω konečný a každý jeho elementární jev je stejně pravděpodobný, pak

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}$$

je rovnoměrná (uniformní) pravděpodobnostní míra.

- K určení pravděpodobnosti tedy potřebujeme znát počet elementárních jevů příznivých jevu A. K tomu slouží kombinatorické metody.

Základní princip počítání

Věta 7 (Základní princip počítání (pravidlo součinu)).

Mějme dva experimenty. Pokud má první experiment m možných výsledků a pro každý výsledek existuje n možných výsledků druhého experimentu, pak počet možných výsledků obou experimentů společně je mn.

Důkaz.

Vyčísleme všechny možné výsledky obou experimentů:

$$\begin{gathered}(1,1),(1,2),\ldots,(1,n)\\(2,1),(2,2),\ldots,(2,n)\\\vdots\\(m,1),(m,2),\ldots,(m,n)\end{gathered}$$

kde (i,j) značí i-tý možný výsledek prvního experimentu a j-tý možný výsledek druhého. Množina všech možných výsledků tedy sestává z m řádků, kde každý má n prvků. □

Příklad 8.

- Mějme skupinu 10 žen, kde každá žena má 3 děti.
- Pokud bychom měli zvolit jednu ženu a jedno její dítě za matku a dítě roku, kolik možností máme?
- **Řešení:**
- První experiment je volba ženy, druhý experiment je volba jednoho z jejich dětí.
- Základní princip počítání (pravidlo součinu) dává $10 \cdot 3 = 30$ možností.

Zobecnění základního principu počítání

Věta 9 (Zobecněný základní princip počítání (zobecněné pravidlo součinu)).

Jestliže máme provést r experimentů, kde první má n_1 možných výsledků a pro každý z možných výsledků má druhý n_2 možných výsledků a pro každý z možných výsledků prvních dvou experimentů má třetí experiment n_3 možných výsledků atd., pak celkový počet možných výsledků těchto r experimentů je

$$n_1 \cdot n_2 \cdot n_3 \cdot \cdots \cdot n_r .$$

□

Příklad 10.

- Středoškolská komise se skládá ze
  - 3 prváků
  - 4 druháků
  - 5 třeťáků a
  - 2 čtvrťáků.
- Volíme 4 zastupitele tak, aby z každého ročníku byl přítomen jeden člen komise.
- Kolik různých zastupitelstev můžeme sestavit?

- **Řešení:** $3 \cdot 4 \cdot 5 \cdot 2 = 120$.

Příklad 11.

- Kolik různých 7-místných SPZ lze sestavit, jestliže první 3 místa jsou písmena anglické abecedy a další 4 jsou čísla?
- **Řešení:** $26 \cdot 26 \cdot 26 \cdot 10 \cdot 10 \cdot 10 \cdot 10 = 175\,760\,000$.

Příklad 12.

- Kolik SPZ by bylo možno sestavit, pokud se písmena ani čísla nesmí opakovat?
- **Řešení:** $26 \cdot 25 \cdot 24 \cdot 10 \cdot 9 \cdot 8 \cdot 7 = 78\,624\,000$.

S využitím pravidla součinu lze odvodit řadu často používaných vzorců. Patří k nim:

- **Permutace**
  - Permutace nějakých prvků je jejich seřazení.
- **Permutace s opakováním**
  - Seřazujeme-li objekty z nichž některé jsou stejné, provádíme tzv. permutace s opakováním.
- **Variace** – výběr, u kterého záleží na pořadí vybíraných prvků.
- **Variace s opakováním** – výběr, ve kterém záleží na pořadí vybíraných prvků a ve kterém se prvky mohou opakovat.
- **Kombinace** – výběr, u kterého nezáleží na pořadí vybíraných prvků.
- **Kombinace s opakováním** – výběr, ve kterém nezáleží na pořadí prvků a ve kterém se prvky mohou opakovat.

Permutace

- Kolik je různých uspořádání tří písmen a, b, c?
  - Přímým výpočtem dostaneme 6: *abc*, *acb*, *bac*, *bca*, *cab* a *cba*.
- Každé z těchto 6 uspořádání je permutace.
- Základní princip počítání dává, že první prvek permutace může být libovolný ze 3, druhý libovolný ze 2 a třetí ten jeden zbývající
- Tedy máme $3 \cdot 2 \cdot 1 = 6$ možných permutací.

Definice 13 (Permutace).

Mějme n různých prvků, pak existuje

$$n \cdot (n-1) \cdot (n-2) \cdots 3 \cdot 2 \cdot 1 = n!$$

různých permutací těchto n prvků.

- Zatímco $n!$ („n faktoriál“) je definován jako $1 \cdot 2 \cdots n$ pro celé $n \geq 1$, je vhodné dodefinovat $0! = 1$.

Příklad 14.

- Kolik možných uspořádání existuje v baseballovém týmu 9 hráčů?
- **Řešení:** $9! = 362\,880$ uspořádání.

Příklad 15.

- Ústní zkoušku skládá 6 mužů a 4 ženy.

(a) V kolika různých pořadích mohou postupně na zkoušku přijít?

(b) Kolik různých pořadí existuje, pokud jdou na zkoušku nejprve muži a poté ženy?

- **Řešení:**

(a) 10 lidí lze uspořádat $10! = 3\,628\,800$ způsoby.

(b) Jelikož 6 mužů lze uspořádat 6! způsoby a 4 ženy 4! způsoby, základní princip počítání dává celkem $(6!)(4!) = (720)(24) = 17\,280$ možných pořadí.

Příklad 16.

- Paní Nováková má 10 knih, které chce dát do jedné police knihovny. Z toho jsou
  - 4 o matematice
  - 3 o chemii
  - 2 o historii a
  - 1 jazyková učebnice.
- Paní Nováková chce knihy uspořádat tak, aby stejné obory byly pohromadě.
- Kolika způsoby může knihy uspořádat?

- **Řešení:**
- Paní Nováková má 4!3!2!1! možností, přičemž knihy o matematice jsou první a následují knihy o chemii, o historii a nakonec jazyková učebnice.
- Čtyři témata knih lze uspořádat 4! způsoby.
- Proto má paní Novákové celkem $4!(4!3!2!1!) = 6\,912$ možností.

Příklad 17.

- Kolik různých přesmyček slov PEPPER lze sestavit?

- Existuje 6! permutací písmen $P_1E_1P_2P_3E_2R$, pokud jsou ta tří písmenka P a dvě E od sebe navzájem rozlišitelná.
  - Uvažme libovolnou z těchto permutací, např. $P_1P_2E_1P_3E_2R$.
  - Pokud prohazujeme P-čka mezi sebou a E-čka mezi sebou, výsledek je stále $PPEPER$.
- Tedy 3!2! permutací jsou tvaru $PPEPER$:

$P_1P_2E_1P_3E_2R$	$P_3P_1E_1P_2E_2R$	$P_2P_1E_2P_3E_1R$
$P_1P_3E_1P_2E_2R$	$P_3P_2E_1P_1E_2R$	$P_2P_3E_2P_1E_1R$
$P_2P_1E_1P_3E_2R$	$P_1P_2E_2P_3E_1R$	$P_3P_1E_2P_2E_1R$
$P_2P_3E_1P_1E_2R$	$P_1P_3E_2P_2E_1R$	$P_3P_2E_2P_1E_1R$

- Celkem tedy máme $\dfrac{6!}{3!2!} = 60$ různých přesmyček písmen slova $PEPPER$.

Permutace s opakováním

Počet permutací n prvků, kde první prvek se vyskytuje k_1-krát, druhý k_2-krát, až n-tý k_n-krát je

$$\frac{n!}{k_1!k_2!\cdots k_n!},$$

přičemž $k_1 + k_2 + \ldots + k_n = n$.

Příklad 18.

- Kolik různých signálů lze vytvořit ze 4 bílých, 3 červených a 2 modrých vlajek, jestliže se každý symbol skládá z 9 vlajek zavěšených vedle sebe a víme, že vlajky stejné barvy jsou identické.
- **Řešení:** $\frac{9!}{4!3!2!} = 1\,260.$

Kombinace

- Často nás zajímá počet různých skupin (podmnožin) r prvků z n.
  - Např. kolik různých skupin 3 prvků lze vybrat z prvků A, B, C, D, E?
  - 5 způsobů jak vybrat první, 4 jak vybrat druhý a 3 jak vybrat poslední, proto $5 \cdot 4 \cdot 3$ způsobů, ale...
  - každá skupina 3 prvků (např. A, B, C) bude počítána 6-krát (počítáme všechny permutace ABC, ACB, BAC, BCA, CAB a CBA).
  - Celkový počet skupin je tedy
$$\frac{5 \cdot 4 \cdot 3}{3 \cdot 2 \cdot 1} = 10.$$

- Obecně máme $n(n-1)\cdots(n-r+1)$ různých způsobů, jak vybrat r-prvků z n prvků, kde záleží na pořadí prvků, a každá r-tice je počítána $r!$-krát, tedy máme
$$\frac{n(n-1)\cdots(n-r+1)}{r!} = \frac{n!}{r!(n-r)!}$$
r-prvkových podmnožin z n prvků.

Definice 19.

Definujeme číslo $\binom{n}{r}$ pro $r \leq n$ jako

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

a budeme jej nazývat kombinační číslo a číst jej „n nad r“.

- Z definice $0! = 1$ dostaneme, že $\binom{n}{n} = \binom{n}{0} = \frac{n!}{0!n!} = 1$.
- Pro $r > n$ nebo $r < 0$ dodefinujeme $\binom{n}{r} = 0$.

Příklad 20.

- Z 20 lidí má být vytvořena tříčlenná komise. Kolik různých komisí lze vytvořit?
- **Řešení:** $\binom{20}{3} = 1\,140$.

Příklad 21.

- Z 5 žen a 7 mužů máme vytvořit komisi 2 žen a 3 mužů. Kolik máme možností?
- **Řešení:** Máme $\binom{5}{2}$ možností jak vybrat 2 ženy z 5 a $\binom{7}{3}$ možností jak vybrat 3 muže ze 7. To je celkem $\binom{5}{2}\binom{7}{3} = 350$ možností.
- Co když se 2 muži nemají rádi a odmítají být spolu v komisi?
- **Řešení:** Počet skupin tří mužů, kde jsou oba, kteří se nemají rádi, je $\binom{2}{2}\binom{5}{1} = 5$, a proto máme celkem $(\binom{7}{3} - 5)\binom{5}{2} = 30 \cdot 10 = 300$ možností.

Příklad 22.

- Mějme n antén, z nichž je m vadných a $n-m$ funkčních. Antény jsou nerozlišitelné. Kolika způsoby je můžeme uspořádat tak, aby žádné dvě vadné nebyly vedle sebe?
  - Představme si, že $n-m$ funkčních antén jsou seřazeny vedle sebe.
  - Pokud žádné dvě vadné nemají být vedle sebe, obsahuje každé místo mezi dvěmi funkčními anténami nejvýše jednu vadnou (včetně krajních míst).
  - Tedy máme $n-m+1$ pozic mezi $n-m$ funkčními anténami, kam umístit m vadných antén.
  - To dává $\binom{n-m+1}{m}$ možných uspořádání.

- Pro $1 \leq r \leq n$ je užitečná následující kombinatorická identita:

$$\binom{n}{r} = \binom{n-1}{r-1} + \binom{n-1}{r}. \tag{1}$$

Důkaz.

- Mějme n prvků a fixujme jeden z nich, řekněme x.
- Dostaneme $\binom{n-1}{r-1}$ skupin r prvků obsahující x a $\binom{n-1}{r}$ skupin r prvků neobsahující x.
- Bylo zde použito pravidlo součtu: Lze-li úkol M provést m způsoby a lze-li úkol N provést n způsoby, přičemž žádný z m způsobů provedení úkolu M není totožný s žádným z n způsobů provedení úkolu N, pak provést úkol M **nebo** úkol N lze $m + n$ způsoby.

□

Věta 23 (Binomická věta).

Pro každé přirozené číslo n a libovolná reálná čísla a, b platí:

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k. \tag{2}$$

Příklad 24.

- Rozviňte $(x+y)^3$.
- **Řešení:** $(x+y)^3 = \binom{3}{0}x^3y^0 + \binom{3}{1}x^2y^1 + \binom{3}{2}x^1y^2 + \binom{3}{3}x^0y^3 = x^3 + 3x^2y + 3xy^2 + y^3$.

Příklad 25.

- Kolik podmnožin má n prvková množina?
- **Řešení:** k-prvkových podmnožin je $\binom{n}{k}$, proto všech je $\sum_{k=0}^{n}\binom{n}{k} = (1+1)^n = 2^n$.

Přehled vzorců pro permutace, variace a kombinace

Uspořádaný výběr		
Bez opakování	Variace bez opakování	$\frac{n!}{(n-k)!}$
	Permutace bez opakování	$n!$
S opakováním	Variace s opakováním	n^k
	Permutace s opakováním	$\frac{n!}{n_1! \cdot n_2! \cdots n_k!}$

Neuspořádaný výběr		
Bez opakování	Kombinace bez opakování	$\binom{n}{k}$
S opakováním	Kombinace s opakováním	$\binom{n+k-1}{k} = \binom{n+k-1}{n-1}$

Nezávislé jevy

Nezávislé jevy

- Pokud dvakrát hodíme férovou mincí, pravděpodobnost dvou orlů bude $\frac{1}{2} \cdot \frac{1}{2}$.
  - Pravděpodobnosti násobíme, protože hody považujeme za nezávislé.

Definice 26 (Nezávislé jevy).

Jevy A a B jsou nezávislé, jestliže

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

Množina jevů $\{A_i \mid i \in I\}$ je nezávislá, pokud

$$\mathbb{P}\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} \mathbb{P}(A_i)$$

pro každou konečnou podmnožinu $J \subseteq I$.

Nezávislost vzniká ve dvou případech:

1. Předpokládáme, že jevy jsou nezávislé
   - například při hodu mince dvakrát po sobě často předpokládáme, že hody jsou nezávislé, což odráží fakt, že mince nemá paměť na první hod.

2. Nezávislost odvodíme (ověřením $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$)
   - například při hodu férovou kostkou pro $A = \{2, 4, 6\}$ a $B = \{1, 2, 3, 4\}$ je $A \cap B = \{2, 4\}$ a $\mathbb{P}(AB) = 2/6 = \mathbb{P}(A)\mathbb{P}(B) = (1/2) \cdot (2/3)$, tedy A a B jsou nezávislé.

- Předpokládejme, že A a B jsou disjunktní jevy s nenulovou pravděpodobností.
- Mohou být nezávislé?

- **Ne**, protože $\mathbb{P}(A)\mathbb{P}(B) > 0$, ale $\mathbb{P}(AB) = \mathbb{P}(\emptyset) = 0$,
  - odkud $\mathbb{P}(AB) \neq \mathbb{P}(A)\mathbb{P}(B)$.

- Až na tento speciální případ neexistuje způsob, jak zjistit nezávislost pouze z Vennova diagramu.

Příklad 27.

- Házíme férovou mincí 10 krát.
- Označme jako A jev, že „padnul alespoň jednou orel“ a jako T_j jev, že „orel nepadnul v j-tém hodu“. Pak

$$\begin{aligned}
\mathbb{P}(A) &= 1 - \mathbb{P}(A^c) \\
&= 1 - \mathbb{P}(\text{„žádný orel“}) \\
&= 1 - \mathbb{P}(T_1 T_2 \cdots T_{10}) \\
&= 1 - \mathbb{P}(T_1)\mathbb{P}(T_2)\cdots\mathbb{P}(T_{10}) \qquad \text{(z nezávislosti)} \\
&= 1 - (1/2)^{10} \approx 0,999.
\end{aligned}$$

Shrnutí

- Jevy A a B jsou nezávislé, právě když $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$.
- Nezávislost se někdy předpokládá, někdy odvozuje.
- Disjunktní jevy s nenulovou pravděpodobností nejsou nezávislé.

Podmíněná pravděpodobnost

Podmíněná pravděpodobnost

Definice 28.

Jestliže $\mathbb{P}(B) > 0$, pak podmíněná pravděpodobnost jevu A za předpokladu, že nastal jev B je

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

- $\mathbb{P}(A|B)$ vyjadřuje, kolikrát nastal jev A mezi jevy, kde nastal jev B.
- Pro pevné B s $\mathbb{P}(B) > 0$ je $\mathbb{P}(\cdot|B)$ pravděpodobnost (splňuje axiomy)
  - $\mathbb{P}(A|B) \geq 0$
  - $\mathbb{P}(\Omega|B) = 1$ a
  - pokud jsou $A_1, A_2, \ldots$ disjunktní, tak $\mathbb{P}(\bigcup_{i=1}^{+\infty} A_i|B) = \sum_{i=1}^{+\infty} \mathbb{P}(A_i|B)$.
- Obecně neplatí $\mathbb{P}(A|B \cup C) = \mathbb{P}(A|B) + \mathbb{P}(A|C)$.
- Obecně neplatí $\mathbb{P}(A|B) = \mathbb{P}(B|A)$, například pravděpodobnost vyrážky u spalniček je 1, ale pravděpodobnost spalniček při vyrážce není 1.

Příklad 29 (Testování).

Test na nemoc n dává výsledky $+$ (pozitivní) a $-$ (negativní).
Otestování 1000 vzorků (10 s virem, 990 bez viru) dopadlo následovně:

	n	n^c
$+$	0,009	0,099
$-$	0,001	0,891

Z definice podmíněné pravděpodobnosti máme

$$\mathbb{P}(+|n) = \frac{\mathbb{P}(+\cap n)}{\mathbb{P}(n)} = \frac{0,009}{0,009+0,001} = 0,9 \quad \text{a} \quad \mathbb{P}(-|n^c) = \frac{0,891}{0,099+0,891} = 0,9.$$

Nemocný má pozitivní test (senzitivita) v 90 % a zdravý má negativní test (specificita) v 90 %.
Když si nechám udělat test a ten bude pozitivní, jaká je pravděpodobnost, že jsem nemocný?

$$\mathbb{P}(n|+) = \frac{\mathbb{P}(n\cap +)}{\mathbb{P}(+)} = \frac{0,009}{0,009+0,099} \approx 0,083 \qquad \text{asi 8,3 \%}$$

Lemma 30.

1. *Jestliže A a B jsou nezávislé jevy, pak $\mathbb{P}(A|B) = \mathbb{P}(A)$.*
2. *Pro libovolnou dvojici jevů A a B je*

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

Důkaz.

Přímo z definice. □

- Jiná interpretace nezávislosti tedy je, že znalost B nemění pravděpodobnost A.
- Rovnost $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B|A)$ je užitečná, budeme ji dále využívat.

Příklad 31.

- Bez opakování vybereme z balíčku 52 unikátních karet dvě karty.
  - Nechť A je jev, že první karta je křížové eso.
  - Nechť B je jev, že druhá karta je kárová dáma.
- Pak $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B|A) = (1/52) \cdot (1/51)$.

Shrnutí

- Jestliže $\mathbb{P}(B) > 0$, pak $\mathbb{P}(A|B) = \dfrac{\mathbb{P}(AB)}{\mathbb{P}(B)}$.
- $\mathbb{P}(\cdot|B)$ splňuje axiomy pravděpodobnosti pro fixní B.
- Obecně $\mathbb{P}(A|\cdot)$ nesplňuje axiomy pravděpodobnosti pro fixní A.
- Obecně $\mathbb{P}(A|B) \neq \mathbb{P}(B|A)$.
- A a B jsou nezávislé, právě když $\mathbb{P}(A|B) = \mathbb{P}(A)$.

Bayesova věta

Bayesova věta

Bayesova věta je základem expertních systémů a Bayesovských sítí.

Věta 32 (Věta o úplné pravděpodobnosti).

Nechť $A_1, \ldots, A_k$ je rozklad Ω. Pak pro libovolný jev B platí

$$\mathbb{P}(B) = \sum_{i=1}^{k} \mathbb{P}(B|A_i)\mathbb{P}(A_i).$$

Důkaz.

Definujme $C_j = B \cap A_j$, pak $C_1, \ldots, C_k$ jsou disjunktní a $B = \bigcup_{j=1}^{k} C_j$, tedy

$$\mathbb{P}(B) = \sum_{j=1}^{k} \mathbb{P}(C_j) = \sum_{j=1}^{k} \mathbb{P}(B \cap A_j) = \sum_{j=1}^{k} \mathbb{P}(B|A_j)\mathbb{P}(A_j),$$

protože $\mathbb{P}(B \cap A_j) = \mathbb{P}(B|A_j)\mathbb{P}(A_j)$ z definice podmíněné pravděpodobnosti. □

Věta 33 (Bayesova věta).

Nechť $A_1, \ldots, A_k$ je rozklad Ω takový, že $\mathbb{P}(A_i) > 0$ pro všechna i. Jestliže $\mathbb{P}(B) > 0$, pak pro každé $i = 1, \ldots, k$ platí

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{j=1}^{k} \mathbb{P}(B|A_j)\mathbb{P}(A_j)}.$$

Důkaz.

Z dvojího použití definice podmíněné pravděpodobnosti a věty o úplné pravděpodobnosti máme

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_i B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{j=1}^{k} \mathbb{P}(B|A_j)\mathbb{P}(A_j)}.$$

□

Příklad 34.

- Rozdělíme emaily do tří kategorií:
  - $A_1 =$ „spam“
  - $A_2 =$ „nedůležité“ a
  - $A_3 =$ „důležité“.
- Ze zkušenosti víme, že $\mathbb{P}(A_1) = 0,7$, $\mathbb{P}(A_2) = 0,2$ a $\mathbb{P}(A_3) = 0,1$.
  - Zajisté platí, že $0,7 + 0,2 + 0,1 = 1$.
- Nechť B je jev, že email obsahuje slovo „free“.
  - Ze zkušenosti víme, že $\mathbb{P}(B|A_1) = 0,9$, $\mathbb{P}(B|A_2) = 0,01$, $\mathbb{P}(B|A_3) = 0,01$.
    - Všimněme si, že $0,9 + 0,01 + 0,01 \neq 1$.
- Pokud obdržíme email se slovem „free“, jaká je pravděpodobnost, že jde o spam?
- Bayesova věta dává

$$\mathbb{P}(A_1|B) = \frac{0,9 \cdot 0,7}{(0,9 \cdot 0,7) + (0,01 \cdot 0,2) + (0,01 \cdot 0,1)} \approx 0,995.$$

Náhodná veličina

Náhodná veličina

- Statistika a data mining se zabývají daty.
- Jak spojit výběrové prostory a jevy s daty?
  - Pomocí konceptu náhodné veličiny.

Definice 35.

Náhodná veličina je funkce $X\colon \Omega \to \mathbb{R}$, která přiřazuje reálné číslo $X(\omega)$ každému výsledku $\omega \in \Omega$.

- Pravděpodobnostní míra $\mathbb{P}$ je definována na σ-algebře $\mathcal{A}$ prostoru Ω.
- Náhodná veličina X je měřitelná funkce $X\colon \Omega \to \mathbb{R}$.
  - Měřitelná znamená, že pro každé x je množina $\{\omega \in \Omega \mid X(\omega) \leq x\}$ jev, tedy $\{\omega \mid X(\omega) \leq x\} \in \mathcal{A}$.

Poznámka 36.

Ačkoli budeme pracovat přímo s náhodnými veličinami bez uvádění výběrového prostoru, je třeba si uvědomit, že výběrový prostor tam někde vždy je!

Příklad 37.

Uvažujme hod mincí desetkrát po sobě. Nechť $X(\omega)$ je počet orlů v posloupnosti ω, například pro $\omega = OOPOOPOOPP$ je $X(\omega) = 6$.

Příklad 38.

Nechť $\Omega = \{(x, y) \in \mathbb{R} \times \mathbb{R} \mid x + y \leq 1\}$. Náhodně zvolme bod z Ω. Typický výsledek je tvaru $\omega = (x, y)$. Příklady náhodných veličin:

- $X(\omega) = x$
- $Y(\omega) = y$
- $Z(\omega) = x + y$
- $W(\omega) = \sqrt{x^2 + y^2}$.

Pro danou náhodnou veličinu X a podmnožinu A reálné osy definujeme

$$X^{-1}(A) = \{\omega \in \Omega \mid X(\omega) \in A\}$$

a dále

$$\mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A)) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in A\})$$
$$\mathbb{P}(X = x) = \mathbb{P}(X^{-1}(x)) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) = x\}).$$

Všimněme si, že X značí náhodnou veličinu a x značí konkrétní hodnotu X.

Příklad 39.

Hodíme mincí dvakrát po sobě. Nechť X je počet orlů. Pak

- $\mathbb{P}(X = 0) = \mathbb{P}(\{PP\}) = 1/4$
- $\mathbb{P}(X = 1) = \mathbb{P}(\{OP, PO\}) = 1/2$
- $\mathbb{P}(X = 2) = \mathbb{P}(\{OO\}) = 1/4.$

Náhodnou veličinu a její distribuci lze zapsat tabulkou

ω	$\mathbb{P}(\{\omega\})$	$X(\omega)$
PP	$1/4$	0
PO	$1/4$	1
OP	$1/4$	1
OO	$1/4$	2

x	$\mathbb{P}(X = x)$
0	$1/4$
1	$1/2$
2	$1/4$

Distribuční a pravděpodobnostní funkce

Distribuční a pravděpodobnostní funkce

Definice 40.

(Kumulativní) distribuční funkce je funkce $F_X \colon \mathbb{R} \to [0, 1]$ definovaná jako

$$F_X(x) = \mathbb{P}(X \leq x)\,.$$

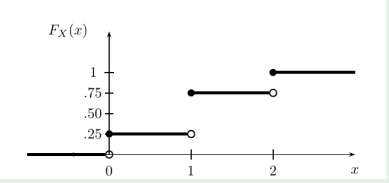
Distribuční funkce obsahuje veškerou informaci o náhodné veličině.

Občas píšeme distribuční funkce jako F místo F_X.

Příklad 41.

Hod férovou mincí dvakrát, X je počet orlů. Pak $\mathbb{P}(X = 0) = \mathbb{P}(X = 2) = 1/4$ a $\mathbb{P}(X = 1) = 1/2$. Distribuční funkce je

$$F_X(x) = \begin{cases} 0 & \text{pro } x < 0 \\ 1/4 & \text{pro } 0 \leq x < 1 \\ 3/4 & \text{pro } 1 \leq x < 2 \\ 1 & \text{pro } 2 \leq x. \end{cases}$$



Důkladně prostudujte, distribuční funkce mohou být komplikované. Funkce je zprava spojitá, neklesající a definovaná pro všechna x. Proč je $F_X(1,4) = 0,75$?

Distribuční funkce plně určuje rozložení náhodné veličiny.

Věta 42.

Nechť X má distribuční funkci F a Y má distribuční funkci G. Jestliže

$$F(x) = G(x)$$

pro všechna x, pak

$$\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$$

pro všechna A.[2]

[2] Přesněji máme, že $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$ pro každý měřitelný jev A.

Věta 43.

Funkce $F\colon \mathbb{R} \to [0,1]$ *je distribuční funkce pro nějakou pravděpodobnostní míru* $\mathbb{P}$*, právě když* F *splňuje následující tři podmínky:*

1. F *je neklesající:* $x_1 < x_2$ *implikuje, že* $F(x_1) \leq F(x_2)$
2. F *je normalizovaná:* $\lim_{x\to-\infty} F(x) = 0$ *a* $\lim_{x\to+\infty} F(x) = 1$
3. F *je zprava spojitá:* $F(x) = F(x^+)$ *pro všechna* x*, kde* $F(x^+) = \lim_{y\to x,\, x<y} F(y)$.

Diskrétní náhodná veličina

Diskrétní náhodná veličina

Definice 44.

Náhodná veličina X je diskrétní, jestliže nabývá spočetně mnoho hodnot $\{x_1, x_2, \ldots\}$.
Definujme pravděpodobnostní funkci pro X jako

$$f_X(x) = \mathbb{P}(X = x).$$

Pak $f_X(x) \geq 0$ pro všechna $x \in \mathbb{R}$ a $\sum_i f_X(x_i) = 1$.

Někdy píšeme f místo f_X.

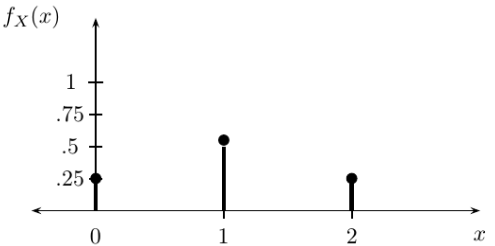
Distribuční funkce X souvisí s pravděpodobnostní funkcí f_X následovně:

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{x_i \leq x} f_X(x_i).$$

Příklad 45.

Hod férovou mincí dvakrát, X je počet orlů. Pak $\mathbb{P}(X = 0) = \mathbb{P}(X = 2) = 1/4$ a $\mathbb{P}(X = 1) = 1/2$. Pravděpodobnostní funkce je

$$f_X(x) = \begin{cases} 1/4 & \text{pro } x = 0 \\ 1/2 & \text{pro } x = 1 \\ 1/4 & \text{pro } x = 2 \\ 0 & \text{jinak.} \end{cases}$$



Vybrané diskrétní náhodné veličiny

- $X \sim F$ značí, že X má rozdělení F.
- $X \sim F$ tedy čteme jako „X má rozdělení F“, nikoli „X je přibližně F“.

Bodové rozdělení

X má bodové rozdělení v a, $X \sim \delta_a$, jestliže $\mathbb{P}(X = a) = 1$, přičemž

$$F(x) = \begin{cases} 0 & \text{pro } x < a \\ 1 & \text{pro } x \geq a. \end{cases}$$

Pravděpodobnostní funkce je $f(x) = 1$ pro $x = a$ a 0 jinak.

Diskrétní rovnoměrné rozdělení

Nechť $k > 1$ a nechť X má pravděpodobnostní funkci

$$f(x) = \begin{cases} \frac{1}{k} & \text{pro } x = 1, \ldots, k \\ 0 & \text{jinak.} \end{cases}$$

Pak X má rovnoměrné (uniformní) rozdělení na $\{1, \ldots, k\}$.

Jak vypadá distribuční funkce?

Bernoulliho rozdělení

Nechť X představuje hod mincí. Pak

$$\mathbb{P}(X = 1) = p$$

a

$$\mathbb{P}(X = 0) = 1 - p$$

pro $p \in [0, 1]$.
Řekneme, že X má Bernoulliho rozdělení, $X \sim Bernoulli(p)$.
Pravděpodobnostní funkce je

$$f(x) = p^x(1 - p)^{1-x} \text{ pro } x \in \{0, 1\}.$$

Jak vypadá distribuční funkce?

Binomické rozdělení

Mějme minci na níž padá orel s pravděpodobností p pro nějaké $0 \leq p \leq 1$. Házíme n krát a X je počet hodů, kdy padl orel. Předpokládejme, že hody jsou nezávislé. Pak pravděpodobnostní funkce je

$$f(x) = \mathbb{P}(X = x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{pro } x = 0, \ldots, n \\ 0 & \text{jinak.} \end{cases}$$

Taková náhodná veličina (NV) se nazývá binomická, $X \sim Binomial(n, p)$.

Jestliže $X_1 \sim Binomial(n_1, p)$ a $X_2 \sim Binomial(n_2, p)$, pak $X_1 + X_2 \sim Binomial(n_1 + n_2, p)$.

Jak vypadá distribuční funkce si ukážeme na následujícím příkladu.

Příklad 46.

Házíme pětkrát férovou mincí a předpokládáme, že hody jsou nezávislé. Jaká je pravděpodobnostní funkce počtu padlých orlů? (Jak vypadá odpovídající distribuční funkce?) Nechť X je NV vyjadřující počet úspěchů, tj. počet padlých orlů. Pak X má binomické rozdělení s parametry $n = 5$ a $p = 1/2$:

$$\mathbb{P}(X = 0) = \binom{5}{0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^5 = \frac{1}{32}$$
$$\mathbb{P}(X = 1) = \binom{5}{1}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^4 = \frac{5}{32}$$
$$\mathbb{P}(X = 2) = \binom{5}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^3 = \frac{10}{32}$$
$$\mathbb{P}(X = 3) = \binom{5}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^2 = \frac{10}{32}$$
$$\mathbb{P}(X = 4) = \binom{5}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^1 = \frac{5}{32}$$
$$\mathbb{P}(X = 5) = \binom{5}{5}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^0 = \frac{1}{32}.$$

Příklad 47.

Jistý výrobce vyrábí produkt, o němž je známo, že je špatný s pravděpodobností 0,01, nezávisle na ostatních. Tento produkt výrobce prodává v balení po 10 kusech a nabízí výměnu balení, pokud je více než jeden produkt špatný. Jaké množství prodaných balení by mělo být výrobcem vyměněno?

Řešení: Pokud X značí počet špatných produktů, pak X je binomická NV s parametry $(10; 0,01)$. Pravděpodobnost, že balení bude obsahovat alespoň dva špatné produkty je tedy

$$1 - (\mathbb{P}(X = 0) + \mathbb{P}(X = 1)) = 1 - \binom{10}{0}(0,01)^0(0,99)^{10} - \binom{10}{1}(0,01)^1(0,99)^9 \approx 0,004.$$

Výrobce tedy očekává oprávnění na výměnu u 0,4 procenta balení.

Poznámka

- X je náhodná veličina a x je konkrétní hodnota náhodné veličiny.
- n a p jsou parametry, nějaká fixní reálná čísla.
- Parametr p je obvykle neznámý a musí být odhadnut z dat (třeba na základě statistické inference).
- V mnoha statistických modelech jsou NV a parametry – nezaměňovat!

Poissonovo rozdělení

NV X má Poissonovo rozdělení s parametrem λ, $X \sim Poisson(\lambda)$, jestliže

$$f(x) = \mathbb{P}(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}, \text{ kde } x \geq 0.$$

Platí

$$\sum_{x=0}^{+\infty} f(x) = e^{-\lambda}\sum_{x=0}^{+\infty}\frac{\lambda^x}{x!} = e^{-\lambda}e^{\lambda} = 1.$$

Poznámka 48.

Poissonovo rozdělení se často používá jako model pro počítání vyjímečných jevů (radioaktivní rozklad, dopravní nehody, atd.).

Jestliže $X_1 \sim Poisson(\lambda_1)$ a $X_2 \sim Poisson(\lambda_2)$, pak $X_1 + X_2 \sim Poisson(\lambda_1 + \lambda_2)$.

Aproximace binomického rozdělení

Poissonovo rozdělení má mnoho aplikací – lze jej použít jako aproximaci binomického rozdělení s parametry (n, p), kde n je velké a p je dostatečně malé, aby np bylo přiměřené.

Nechť X je binomická NV s parametry (n, p) a $\lambda = np$. Pak

$$\begin{aligned}\mathbb{P}(X = i) &= \frac{n!}{(n-i)!i!}p^i(1-p)^{n-i} = \frac{n!}{(n-i)!i!}\left(\frac{\lambda}{n}\right)^i\left(1-\frac{\lambda}{n}\right)^{n-i} \\ &= \frac{n(n-1)\cdots(n-i+1)\lambda^i(1-\lambda/n)^n}{n^i i!(1-\lambda/n)^i}.\end{aligned}$$

Pro velké n a přiměřené λ dostáváme

$$\left(1-\frac{\lambda}{n}\right)^n \approx e^{-\lambda}, \qquad \frac{n(n-1)\cdots(n-i+1)}{n^i} \approx 1, \qquad \left(1-\frac{\lambda}{n}\right)^i \approx 1,$$

odkud

$$\mathbb{P}(X = i) \approx e^{-\lambda}\frac{\lambda^i}{i!}.$$

Příklady použití

Pokud provedeme n nezávislých pokusů, každý s pravděpodobností úspěchu p, a pokud n je velké a p dostatečně malé, aby np bylo přiměřené, tak počet výskytů úspěchů je přibližně Poissonova NV s parametrem $\lambda = np$.

Hodnota λ se obvykle zjistí empiricky.

Příklady NV, které mají Poissonovo rozdělení:

- Počet překlepů na stránce knihy.
- Počet lidí v komunitě, kteří se dožijí 100 let.
- Počet špatně zadaných telefonních čísel denně.
- Počet balíků psích sucharů prodaných v daném obchodě za den.
- Počet zákazníků, kteří přijdou daný den na poštu.
- Počet α-částic uvolněných z radioaktivního materiálu během fixního časového období.

Příklad 49.

Nechť počet typografických chyb na jedné stránce knihy má Poissonovo rozdělení s parametrem $\lambda = 1/2$. Určete pravděpodobnost, že na dané stránce je chyba.

Řešení: Nechť X značí počet chyb na dané stránce. Pak máme

$$\mathbb{P}(X \geq 1) = 1 - P(X = 0) = 1 - e^{-1/2} \approx 0,393.$$

Příklad 50.

Nechť pravděpodobnost, že produkt vyrobený jistým strojem bude vadný je 0,1. Určete pravděpodobnost toho, že vzorek 10 produktů bude obsahovat nejvýše jeden vadný.

Řešení: Hledaná pravděpodobnost je

$$\binom{10}{0}(0,1)^0(0,9)^{10} + \binom{10}{1}(0,1)^1(0,9)^9 = 0,7365.$$

Pro srovnání, aproximace pomocí Poissonova rozdělení dává hodnotu

$$e^{-1} + e^{-1} \approx 0,7358.$$

Poznámka

- NV jsou funkce z Ω do $\mathbb{R}$, ale v rozděleních nezmiňujeme výběrový prostor Ω.
  - Ten tam vždy je a lze ho zkonstruovat.
- Zkonstruujme výběrový prostor například pro Bernoulliho NV.
  - Nechť $\Omega = [0, 1]$ a definujme $\mathbb{P}([a, b]) = b - a$ pro $0 \leq a \leq b \leq 1$.
  - Fixujme $p \in [0, 1]$ a definujme

$$X(\omega) = \begin{cases} 1 & \text{pro } \omega \leq p \\ 0 & \text{pro } \omega > p. \end{cases}$$

  - Pak $\mathbb{P}(X = 1) = \mathbb{P}(\omega \leq p) = \mathbb{P}([0, p]) = p$ a $\mathbb{P}(X = 0) = 1 - p$.
  - Tedy $X \sim \mathit{Bernoulli}(p)$.
- Podobně lze postupovat pro všechna definovaná rozdělení.
- V praxi bereme NV jako náhodná čísla, ale formálně jde o funkce na nějakém výběrovém prostoru.

Spojitá náhodná veličina

Spojitá náhodná veličina

Definice 51.

Náhodná veličina X je spojitá, jestliže existuje funkce f_X taková, že

1. $f_X(x) \geq 0$ pro všechna x
2. $\int_{-\infty}^{+\infty} f_X(x)\,dx = 1$
3. pro každé $a \leq b$ platí

$$\mathbb{P}(a < X < b) = \int_a^b f_X(x)\,dx.$$

Funkce f_X se nazývá hustota. Platí

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$$

a $f_X(x) = F_X'(x)$ ve všech bodech x, ve kterých je F_X diferencovatelná.

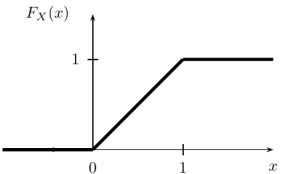
Příklad 52.

Nechť X má hustotu

$$f_X(x) = \begin{cases} 1 & \text{pro } 0 \leq x \leq 1 \\ 0 & \text{jinak.} \end{cases}$$

Pak $f_X(x) \geq 0$ a $\int f_X(x)\,dx = 1$. NV s touto hustotou má uniformní (rovnoměrné) rozdělení na intervalu $(0, 1)$, tj. náhodný výběr bodu mezi 0 a 1. Distribuční funkce je pak dána jako

$$F_X(x) = \begin{cases} 0 & \text{pro } x < 0 \\ x & \text{pro } 0 \leq x \leq 1 \\ 1 & \text{pro } x > 1. \end{cases}$$



Příklad 53.

Nechť X má hustotu

$$f_X(x) = \begin{cases} 0 & \text{pro } x < 0 \\ \frac{1}{(1+x)^2} & \text{jinak.} \end{cases}$$

Jelikož $\int f_X(x)\,dx = 1$, jde skutečně o hustotu.

Pozor!

- Pokud je X spojitá, tak $\mathbb{P}(X = x) = 0$ pro každé x!
  - Neuvažujte tedy o $f(x)$ jako o $\mathbb{P}(X = x)$, to platí pouze pro diskrétní NV.
  - Pravděpodobnosti získáme z hustoty integrací.

- Hustota může být větší než 1 (narozdíl od pravděpodobnosti).
  - Například pro

$$f(x) = \begin{cases} 5 & \text{pro } x \in [0, 1/5] \\ 0 & \text{jinak.} \end{cases}$$

  je $f(x) \geq 0$ a $\int f(x)\,dx = 1$, tudíž jde o hustotu, ačkoli $f(x) = 5$.

- Hustota může být i neomezená.
  - Například pro

$$f(x) = \begin{cases} (2/3)x^{-1/3} & \text{pro } 0 < x < 1 \\ 0 & \text{jinak.} \end{cases}$$

Příklad 54.

Nechť

$$f(x) = \begin{cases} 0 & \text{pro } x < 0 \\ \frac{1}{1+x} & \text{jinak.} \end{cases}$$

Pak nejde o hustotu, protože $\int f(x)\,dx = \int_0^{+\infty} \frac{dx}{1+x} = \ln +\infty = +\infty$.

Lemma 55.

Nechť F je distribuční funkce náhodné veličiny X. Pak

1. $\mathbb{P}(X = x) = F(x) - F(x^-)$, *kde* $F(x^-) = \lim_{y \to x^-} F(y)$
2. $\mathbb{P}(x < X \leq y) = F(y) - F(x)$
3. $\mathbb{P}(X > x) = 1 - F(x)$
4. *Pokud je X spojitá, tak*

$$\begin{aligned} F(b) - F(a) &= \mathbb{P}(a < X < b) = \mathbb{P}(a \leq X < b) \\ &= \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X \leq b). \end{aligned}$$

Vybrané spojité náhodné veličiny

Rovnoměrné rozdělení

NV X má rovnoměrná rozdělení na intervalu (a, b), $X \sim Uniform(a, b)$, jestliže je pro $a < b$

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

Distribuční funkce má tvar

$$F(x) = \begin{cases} 0 & \text{pro } x < a \\ \frac{x-a}{b-a} & \text{pro } x \in [a, b] \\ 1 & \text{pro } x > b. \end{cases}$$

Příklad 56.

Nechť X je NV s rovnoměrným rozdělením na intervalu $(0, 10)$. Určete pravděpodobnost, že

- $X < 3$
- $X > 6$
- $3 < X < 8$.

Řešení:

$$\mathbb{P}(X < 3) = \int_0^3 \frac{1}{10} dx = \frac{3}{10}$$

$$\mathbb{P}(X > 6) = \int_6^{10} \frac{1}{10} dx = \frac{4}{10}$$

$$\mathbb{P}(3 < X < 8) = \int_3^8 \frac{1}{10} dx = \frac{1}{2}.$$

Normální (Gaussovo) rozdělení

NV X má normální (Gaussovo) rozdělení s parametry μ a σ, $X \sim N(\mu, \sigma^2)$, jestliže

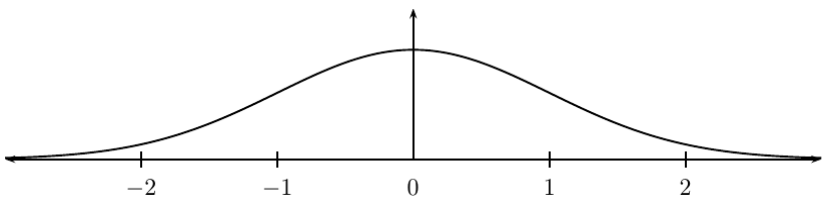
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

kde $x \in \mathbb{R}$, $\mu \in \mathbb{R}$ a $\sigma > 0$.

- Parametr μ je střední hodnota rozdělení a σ je směrodatná odchylka rozdělení.
  - Střední hodnotu a směrodatnou odchylku definujeme později.
- Normální rozdělení hraje důležitou roli v pravděpodobnosti a statistice.
  - Mnoho přírodních fenoménů má přibližně normální rozdělení.
- Později budeme studovat centrální limitní větu, která říká, že rozdělení sumy náhodných veličin lze aproximovat normálním rozdělením.

Standardní normální rozdělení

- NV X má standardní normální rozdělení pokud je $\mu = 0$ a $\sigma = 1$.
- Standardní normální NV budeme značit Z.
- Hustota a distribuční funkce standardní NV se značí $\varphi(z)$ a $\Phi(z)$.[3]



[3] Hodnoty $\Phi(z)$ hledáme v tabulkách.

Vlastnosti standardní NV

1. Jestliže $X \sim N(\mu, \sigma^2)$, pak $Z = \dfrac{(X-\mu)}{\sigma} \sim N(0,1)$.
2. Jestliže $Z \sim N(0,1)$, pak $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$.
3. Jestliže $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$ jsou nezávislé, pak

$$\sum_{i=1}^{n} X_i \sim N\Big(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\Big).$$

Pokud je $X \sim N(\mu, \sigma^2)$, pak

$$\mathbb{P}(a < X < b) = \mathbb{P}\left(\frac{a-\mu}{\sigma} < Z \ < \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

- Platí, že $\Phi(-x) = 1 - \Phi(x)$.

Příklad 57.

Nechť $X \sim N(3, 9)$. Určete $\mathbb{P}(2 < X < 5)$, $\mathbb{P}(X > 0)$ a $\mathbb{P}(|X - 3| > 6)$.

Řešení:

$$\mathbb{P}(2 < X < 5) = \mathbb{P}\left(\frac{2-3}{3} < \frac{X-3}{3} < \frac{5-3}{3}\right) = \mathbb{P}\left(-\frac{1}{3} < Z < \frac{2}{3}\right) = \Phi\left(\frac{2}{3}\right) - \Phi\left(-\frac{1}{3}\right)$$
$$= \Phi\left(\frac{2}{3}\right) - \left(1 - \Phi\left(\frac{1}{3}\right)\right) \approx 0,3779$$

$$\mathbb{P}(X > 0) = \mathbb{P}\left(\frac{X-3}{3} > \frac{0-3}{3}\right) = \mathbb{P}\left(Z > -1\right) = 1 - \Phi\left(-1\right) = \Phi\left(1\right) \approx 0,8413$$

$$\mathbb{P}(|X - 3| > 6) = \mathbb{P}\left(X > 9\right) + \mathbb{P}\left(X < -3\right) = \mathbb{P}\left(Z > 2\right) + \mathbb{P}\left(Z < -2\right)$$
$$= 1 - \Phi\left(2\right) + \Phi\left(-2\right) = 2(1 - \Phi\left(2\right)) \approx 0,0456.$$

Exponenciální rozdělení

NV X má exponenciální rozdělení s parametrem β, $X \sim Exp(\beta)$, jestliže

$$f(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}},$$

kde $x > 0$, $\beta > 0$.

$$F(x) = \begin{cases} 1 - e^{\frac{-x}{\beta}} & \text{pokud } x > 0 \\ 0 & \text{jinak.} \end{cases}$$

Exponenciální rozdělení se používá na modelování doby čekání mezi vzácnými jevy:

- doba mezi nehodami na jisté křižovatce
- doba životnosti počítače
- doba čekání ve frontě.

Příklad 58.

Předpokládejme, že délka odbavení zákazníka kupujícího nový mobil v minutách je exponenciální NV s parametrem $\beta = 10$. Pokud někdo přijde těsně před vámi, jaká je pravděpodobnost, že budete čekat
(a) více jak 10 minut?
(b) mezi 10 a 20 minutami?

Řešení: Nechť X značí délku odbavení zákazníka před vámi. Pak

$$\text{(a)} \quad \mathbb{P}(X > 10) = 1 - \mathbb{P}(X \leq 10) = 1 - \int_0^{10} \frac{1}{10} e^{-\frac{x}{10}} dx = 1 - (1 - e^{-1}) = e^{-1} \approx 0,368.$$

$$\text{(b)} \quad \mathbb{P}(10 < X < 20) = \int_{10}^{20} \frac{1}{10} e^{-\frac{x}{10}} dx = -e^{-2} + e^{-1} \approx 0,233.$$

Studentovo t rozdělení a Cauchyho rozdělení

NV X má *t rozdělení* s ν stupni volnosti[4], $X \sim t_\nu$, jestliže

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \cdot \frac{1}{(1+\frac{x^2}{\nu})^{\frac{1+\nu}{2}}}.$$

Gamma funkce je pro $\alpha > 0$ definována jako

$$\Gamma(\alpha) = \int_0^{+\infty} y^{\alpha-1} e^{-y}\, dy.$$

Cauchyho rozdělení je speciální případ t rozdělení pro $\nu = 1$. Hustota je

$$f(x) = \frac{1}{\pi(1+x)^2}.$$

[4] Normální rozdělení odpovídá t rozdělení s $\nu = +\infty$.

χ^2 rozdělení

NV X má χ^2 rozdělení s p stupni volnosti, $X \sim \chi^2_p$, jestliže je pro $x > 0$

$$f(x) = \frac{1}{\Gamma(\frac{p}{2}) 2^{\frac{p}{2}}} x^{\frac{p}{2}-1} e^{-\frac{x}{2}}.$$

Jsou-li $Z_1, \ldots, Z_p$ nezávislé standardní normální NV, pak $\sum_{i=1}^{p} Z_i^2 \sim \chi^2_p$.

Rozdělení χ^2 se také nazývá Pearsonovo rozdělení. Využívá se ve statistice a má velký význam pro určování, zda množina dat vyhovuje dané distribuční funkci.

Sdružená rozdělení

Sdružená rozdělení

- Pro dvě diskrétní NV X a Y definujeme sdruženou pravděpodobnostní funkci

$$f(x, y) = \mathbb{P}(X = x \text{ a } Y = y).$$

- $\mathbb{P}(X = x \text{ a } Y = y)$ budeme stručně zapisovat $\mathbb{P}(X = x, Y = y)$.

- Pokud budeme chtít specifikovat NV, budeme psát $f_{X,Y}$.

Příklad 59.

Mějme sdružené rozdělení NV X a Y, kde každá NV nabývá hodnot 0 nebo 1:

	$Y=0$	$Y=1$
$X=0$	$1/9$	$2/9$
$X=1$	$2/9$	$4/9$

Pak například $f(1,1) = \mathbb{P}(X=1, Y=1) = \frac{4}{9}$.

Sdružená funkce hustoty

Definice 60.

Ve spojitém případě je $f(x, y)$ hustota sdružené NV (X, Y), jestliže

1. $f(x, y) \geq 0$ pro všechna (x, y)
2. $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y)\, dx\, dy = 1$
3. pro libovolnou množinu $A \subseteq \mathbb{R} \times \mathbb{R}$ je $\mathbb{P}((X, Y) \in A) = \iint_A f(x, y)\, dx\, dy$.

V diskrétním i spojitém případě je sdružená distribuční funkce definována jako

$$F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y).$$

Příklad 61.

Nechť sdružená NV (X, Y) má rovnoměrné rozdělení na jednotkovém čtverci. Pak

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

Určete $\mathbb{P}(X < \frac{1}{2}, Y < \frac{1}{2})$.

Řešení: Jev $A = \{X < \frac{1}{2}, Y < \frac{1}{2}\}$ odpovídá podmnožině jednotkového čtverce.

Integrace funkce f přes A odpovídá obsahu A, který je $\frac{1}{4}$, tedy

$$\mathbb{P}\left(X < \frac{1}{2}, Y < \frac{1}{2}\right) = \frac{1}{4}.$$

Příklad 62.

Nechť (X, Y) má hustotu

$$f(x,y) = \begin{cases} x+y & \text{pro } 0 \le x, y \le 1 \\ 0 & \text{jinak.} \end{cases}$$

Pak

$$\int_0^1 \int_0^1 (x+y)\,dx\,dy = \int_0^1 \frac{1}{2}\,dy + \int_0^1 \frac{1}{2}\,dx = 1,$$

což je v souladu s tím, že f je skutečně hustota.

Pozn.: $\iint_I (f(x,y) + g(x,y))\,dx\,dy = \iint_I f(x,y)\,dx\,dy + \iint_I g(x,y)\,dx\,dy$

Marginální rozdělení

Marginální rozdělení

Definice 63.

Jestliže sdružená NV (X, Y) má sdružené rozdělení s pravděpodobnostní funkcí $f_{X,Y}$, pak marginální pravděpodobnost X je

$$f_X(x) = \mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x, Y = y) = \sum_y f(x, y)$$

a podobně marginální pravděpodobnost Y je

$$f_Y(y) = \mathbb{P}(Y = y) = \sum_x \mathbb{P}(X = x, Y = y) = \sum_x f(x, y).$$

Příklad 64.

Nechť $f_{X,Y}$ je dána tabulkou

	$Y=0$	$Y=1$	
$X=0$	$1/10$	$2/10$	$3/10$
$X=1$	$3/10$	$4/10$	$7/10$
	$4/10$	$6/10$	1

Pak

- marginální rozdělení X odpovídá sumě řádků a
- marginální rozdělení Y sumě sloupců.

Například $f_X(0) = \frac{3}{10}$ a $f_X(1) = \frac{7}{10}$.

Definice 65.

Pro spojité NV je marginální hustota

$$f_X(x) = \int f(x,y)\,dy \qquad \text{a} \qquad f_Y(y) = \int f(x,y)\,dx.$$

Marginální distribuční funkce se značí F_X a F_Y.

Příklad 66.

Nechť je pro $x, y \geq 0$

$$f_{X,Y}(x,y) = e^{-(x+y)}.$$

Pak

$$f_X(x) = e^{-x} \int_0^{+\infty} e^{-y}\,dy = e^{-x}.$$

Příklad 67.

Nechť

$$f(x,y) = \begin{cases} x+y & \text{pro } 0 \le x, y \le 1 \\ 0 & \text{jinak.} \end{cases}$$

Pak

$$f_Y(y) = \int_0^1 (x+y)\,dx = \frac{1}{2} + y.$$

Nezávislé náhodné veličiny

Nezávislé náhodné veličiny

Definice 68.

Dvě náhodné veličiny X a Y jsou nezávislé, jestliže pro každé A a B platí

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B).$$

Ověřit, zda X a Y jsou nezávislé, znamená ověřit podmínku pro všechny podmnožiny A a B. Následující trzení dává zjednodušení.

Věta 69.

Nechť X a Y mají sdruženou pravděpodobnost (hustotu) $f_{X,Y}$. Pak X a Y jsou nezávislé, jestliže

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$$

pro všechny hodnoty x a y.[5]

[5] Tvrzení není úplně přesné, hustota je definována pouze pro množiny nenulové míry.

Příklad 70.

Nechť X a Y mají následující rozdělení

	$Y=0$	$Y=1$	
$X=0$	$1/4$	$1/4$	$1/2$
$X=1$	$1/4$	$1/4$	$1/2$
	$1/2$	$1/2$	1

Pak $f_X(0) = f_X(1) = \frac{1}{2}$ a $f_Y(0) = f_Y(1) = \frac{1}{2}$.

X a Y jsou nezávislé, protože

- $f_X(0)f_Y(0) = f(0,0)$
- $f_X(0)f_Y(1) = f(0,1)$
- $f_X(1)f_Y(0) = f(1,0)$
- $f_X(1)f_Y(1) = f(1,1)$.

Příklad 71.

Pokud by X a Y měly následující rozdělení

	$Y=0$	$Y=1$	
$X=0$	$1/2$	0	$1/2$
$X=1$	0	$1/2$	$1/2$
	$1/2$	$1/2$	1

tak by nebyly nezávislé, protože

$$f_X(0)f_Y(1) = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4},$$

přičemž

$$f(0,1) = 0.$$

Příklad 72.

Nechť X a Y jsou nezávislé a mají stejnou hustotu

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

Určete $\mathbb{P}(X + Y \leq 1)$.

Řešení: Z nezávislosti máme, že sdružená hustota je

$$f(x, y) = f_X(x) f_Y(y) = \begin{cases} 4xy & \text{pro } 0 \leq x, y \leq 1 \\ 0 & \text{jinak.} \end{cases}$$

Pak

$$\mathbb{P}(X + Y \leq 1) = \iint_{x+y \leq 1} f(x, y)\, dy\, dx = 4 \int_0^1 x \left(\int_0^{1-x} y\, dy \right) dx = \frac{1}{6}.$$

Náhodné vektory

Náhodné vektory

Nechť $X = (X_1, \ldots, X_n)$, kde $X_1, \ldots, X_n$ jsou náhodné veličiny. Pak X je náhodný vektor.

Nechť $f(x_1, \ldots, x_n)$ je hustota náhodného vektoru X. Pak je možné definovat marginální a podmíněné pravděpodobnosti podobně jako ve sdruženém případě.

Řekneme, že $X_1, \ldots, X_n$ jsou nezávislé, jestliže pro každé $A_1, \ldots, A_n$ je

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i).$$

Opět stačí ověřit, že $f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$.

Definice 73.

Jestliže $X_1, \ldots, X_n$ jsou nezávislé NV a každá má stejné marginální rozdělení s distribuční funkcí F, pak říkáme, že $X_1, \ldots, X_n$ jsou IID (independent and identically distributed) a píšeme

$$X_1, \ldots, X_n \sim F.$$

Jestliže F má hustotu f, píšeme také $X_1, \ldots, X_n \sim f$.

$X_1, \ldots, X_n$ nazýváme také náhodný výběr velikosti n z F.

Více o IID později.

Střední hodnota

Definice 74.

Očekávaná či střední hodnota (či první moment) NV X je definována jako

$$\mathbb{E}[X] = \int x \, dF(x) = \begin{cases} \sum\limits_{x} x f(x) & X \text{ je diskrétní} \\ \\ \int\limits_{-\infty}^{+\infty} x f(x) \, dx & X \text{ je spojitá} \end{cases}$$

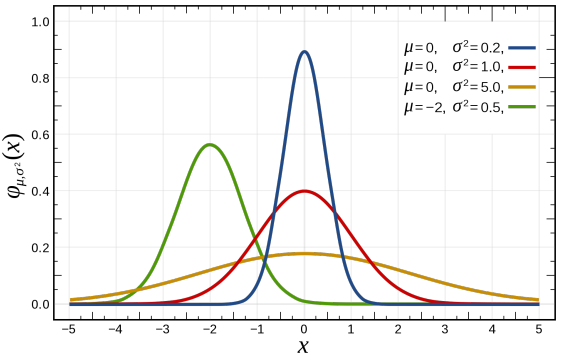
za předpokladu, že daná suma/integrál existuje (je absolutně konvergentní).

Notace:

$$\mathbb{E}[X] = \mathbb{E}(X) = \mathbb{E}X = \int x \, dF(x) = \mu = \mu_X$$

Poznámka: Zápis $\int x \, dF(x)$ zde slouží pouze jako notace pro zjednodušení, abychom nemuseli rozlišovat diskrétní a spojitý případ. (V analýze má svůj vlastní význam!)

- Střední hodnota je jednočíselný souhrn rozdělení.
  - Udává hodnotu, kolem které náhodná veličina kolísá.
- Uvažujme o $\mathbb{E}(X)$ jako o průměru $\sum_{i=1}^{n} X_i / n$ velkého počtu IID[6] výběrů $X_1, \ldots, X_n$.
  - Fakt, že $\mathbb{E}(X) \approx \sum_{i=1}^{n} X_i / n$ je ve skutečnosti věta nazývaná zákon velkých čísel (později).
- $\mathbb{E}(X)$ existuje, jestliže $\int |x| dF_X(x) < +\infty$; jinak neexistuje.



[6] Jestliže $X_1, \ldots, X_n$ jsou nezávislé NV a každá má stejné marginální rozdělení s distribuční funkcí F, pak říkáme, že $X_1, \ldots, X_n$ jsou IID (independent and identically distributed).

Příklad 75.

Nechť $X \sim Bernoulli(p)$, pak

$$\mathbb{E}(X) = \sum_{x=0}^{1} xf(x) = 0(1-p) + 1p = p.$$

Příklad 76.

Hodíme dvakrát férovou mincí. Nechť X je počet orlů. Pak

$$\begin{aligned}\mathbb{E}(X) &= \sum_{x} xf(x) = 0f(0) + 1f(1) + 2f(2) \\ &= 0(1/4) + 1(1/2) + 2(1/4) = 1.\end{aligned}$$

Příklad 77.

Nechť $X \sim Uniform(-1, 3)$, pak

$$\mathbb{E}(X) = \int x f(x)\, dx = \frac{1}{4} \int_{-1}^{3} x\, dx = 1.$$

Příklad 78.

Náhodná veličina má Cauchyho rozdělení, jestliže má hustotu $f_X(x) = (\pi(1 + x^2))^{-1}$. Integrací dostaneme, že

$$\int_{-\infty}^{+\infty} |x| dF(x) = \frac{2}{\pi} \int_{0}^{+\infty} \frac{x\, dx}{1 + x^2} = +\infty,$$

tedy střední hodnota neexistuje.

Kdykoliv dále mluvíme o střední hodnotě, tak předpokládáme, že existuje (je abs. konv.).

Definice 79.

Pro NV X je *k-tý moment* X definován jako $\mathbb{E}(X^k)$, pokud $\mathbb{E}(|X|^k) < +\infty$.

Věta 80.

Jestliže k-tý moment existuje a $j < k$, pak existuje i j-tý moment.

Důkaz.

Platí, že

$$\begin{aligned}\mathbb{E}(|X|^j) &= \int_{-\infty}^{+\infty} |x|^j f_X(x)\,dx = \int_{|x|\le 1} |x|^j f_X(x)\,dx + \int_{|x|>1} |x|^j f_X(x)\,dx \\ &\le \int_{|x|\le 1} f_X(x)\,dx + \int_{|x|>1} |x|^k f_X(x)\,dx \le 1 + \mathbb{E}(|X|^k) < +\infty.\end{aligned}$$

□

Definice 81.

Pro NV X je *k-tý centrální moment* definován jako $\mathbb{E}((X - \mathbb{E}(X))^k)$.

Speciálně máme: $\mathbb{E}(X^0) = 1 = \mathbb{E}(X - \mathbb{E}(X))^0$ a $\mathbb{E}(X - \mathbb{E}(X)) = 0$.

Vlastnosti střední hodnoty

Věta 82.

Jestliže $X_1, \ldots, X_n$ jsou náhodné veličiny a $a_1, \ldots, a_n$ jsou konstanty, pak

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

Příklad 83.

Nechť $X \sim Binomial(n, p)$. Jaká je střední hodnota X?

Řešení: Z definice je $\mathbb{E}(X) = \sum_x x f_X(x) = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x}$, což není snadné určit.

Místo toho si všimněme, že $X = \sum_{i=1}^{n} X_i$ pro $X_i = 1$ když na i-tý hod padl orel a $X_i = 0$ jinak. Pak

$$\mathbb{E}(X_i) = p \cdot 1 + (1-p) \cdot 0 = p,$$

a proto je $\mathbb{E}(X) = \mathbb{E}(\sum_i X_i) = \sum_i \mathbb{E}(X_i) = np$.

Důsledek 84.

Nechť X je NV a c konstanta. Pak $\mathbb{E}(cX) = c\mathbb{E}(X)$.

Důkaz.

Například pro diskrétní NV: $\mathbb{E}(cX) = \sum_x cxf_X(x) = c\sum_x xf_X(x) = c\mathbb{E}(X)$. □

Důsledek 85.

Nechť X a Y jsou NV. Pak $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.

Důsledek 86.

Nechť X je NV, a, b jsou konstanty. Pak $\mathbb{E}(aX+b) = a\mathbb{E}(X) + b$.

Věta 87.

Nechť $X_1, \ldots, X_n$ jsou nezávislé náhodné veličiny. Pak

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

Důsledek 88.

Nechť X_1, X_2 jsou nezávislé náhodné veličiny. Pak

$$\mathbb{E}(X_1 X_2) = \mathbb{E}(X_1)\mathbb{E}(X_2).$$

Poznámka. Pravidlo součtu nevyžaduje nezávislost, pravidlo součinu ano!

Aplikace: průměrná časová složitost Quicksortu

Algoritmus 1: Quicksort

Input: Seznam n různých čísel $S = \{x_1, \ldots, x_n\}$

Output: Seřazený seznam S

1 **if** $|S| \leq 1$ **then return** S
2 $p \leftarrow$ náhodně (uniformě, rovnoměrně) zvolený prvek S
3 $S_1 = \{x \in S \mid x < p\}$
4 $S_2 = \{x \in S \mid x > p\}$
5 Zavolej Quicksort na S_1 a S_2
6 **return** S_1, p, S_2

Poznámka. Toto je tzv. Randomized Quicksort. Při jiné volbě pivota (například prvního prvku seznamu) je analýza průměrné časové složitosti analogická.

Věta 89.

Nechť je pivot p vybírán nezávisle a rovnoměrně ze všech možností. Pak očekávaný počet porovnání dvou čísel pro libovolný vstup je $2n \ln n + \Theta(n)$.

Důkaz

- Nechť $y_1, \ldots, y_n$ jsou hodnoty vstupů $x_1, \ldots, x_n$ seřazené vzestupně.
- Pro $i < j$ je NV X_{ij} rovna 1 pokud jsou y_i a y_j porovnány algoritmem; 0 jinak.
- Celkový počet X porovnání dvou čísel splňuje

$$X = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} X_{ij}$$

a

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} X_{ij}\right) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{E}(X_{ij}).$$

Důkaz pokračování

- Jelikož je X_{ij} indikátor (charakteristická funkce), je $\mathbb{E}(X_{ij}) = \mathbb{P}(X_{ij} = 1)$.
- Potřebujeme tedy určit pravděpodobnost, že prvky y_i a y_j budou porovnány.
- Prvky y_i a y_j budou porovnány právě tehdy, když y_i či y_j je pivot vybraný z množiny $Y^{ij} = \{y_i, y_{i+1}, \ldots, y_{j-1}, y_j\}$:
  - Pokud je y_i (či y_j) pivot z Y^{ij}, pak y_i a y_j musí být ve stejném podseznamu, a tedy budou porovnány.
  - Pokud ani jedno není vybráno jako pivot, pak y_i a y_j budou rozděleny do dvou různých podseznamů, a tedy nebudou nikdy porovnány.
- Jelikož vybíráme pivoty nezávisle a rovnoměrně z každého podseznamu, má každý prvek Y^{ij} stejnou pravděpodobnost, že bude vybrán jako pivot.
- Tedy pravděpodobnost, že y_i či y_j je vybráno jako pivot z Y^{ij}, tedy pravděpodobnost, že $X_{ij} = 1$, je $2/(j - i + 1)$.

Důkaz pokračování.

- Za použití substituce $k = j - i + 1$ dostáváme

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{2}{j-i+1} = \sum_{i=1}^{n-1} \sum_{k=2}^{n-i+1} \frac{2}{k} = \sum_{k=2}^{n} \sum_{i=1}^{n+1-k} \frac{2}{k} \\
&= \sum_{k=2}^{n} (n+1-k)\frac{2}{k} = \sum_{k=2}^{n} (n+1)\frac{2}{k} - \sum_{k=2}^{n} k\frac{2}{k} \\
&= \left((n+1) \sum_{k=2}^{n} \frac{2}{k} \right) - 2(n-1) \\
&= \left(2(n+1) \sum_{k=2}^{n} \frac{1}{k} \right) - 2n + 2 + 2(n+1) - 2(n+1) = 2(n+1) \sum_{k=1}^{n} \frac{1}{k} - 4n.
\end{aligned}$$

- Jelikož platí, že $\sum_{k=1}^{n} \frac{1}{k} = \ln n + \Theta(1)$, dostáváme $\mathbb{E}(X) = 2n \ln n + \Theta(n)$.

Variance a kovariance

Variance

Variance (též rozptyl) je charakteristika variability rozdělení pravděpodobnosti náhodné veličiny. Vyjadřuje variabilitu rozdělení souboru náhodných hodnot kolem své střední hodnoty.

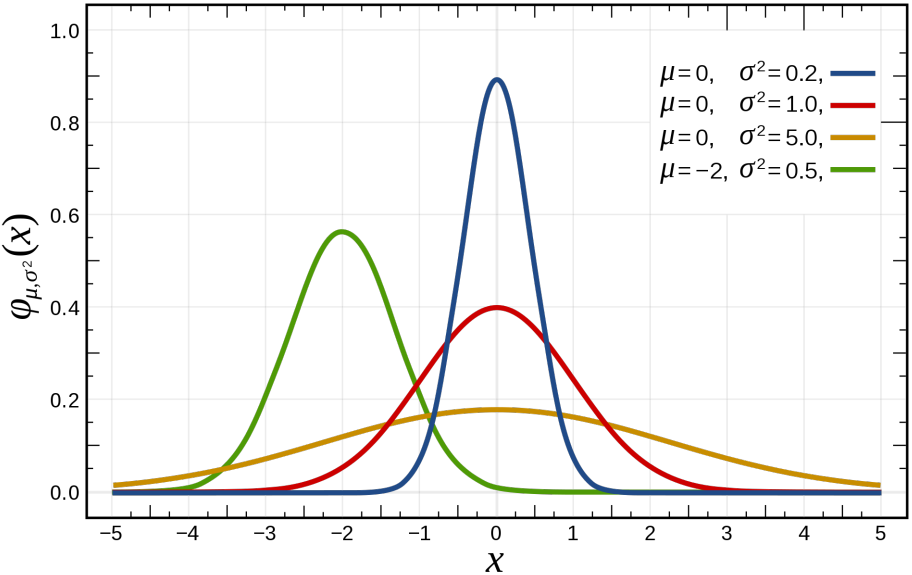
Definice 90.

Nechť X je náhodná veličina se střední hodnotou $\mu = \mathbb{E}[X]$.
Variance X (značeno σ^2, σ_X^2 či $Var(X)$) je definována jako

$$Var(X) = \sigma^2 = \mathbb{E}[(X-\mu)^2] = \int (x-\mu)^2 dF(x),$$

pokud střední hodnota existuje.

Poznámka. Všimněme si, že nelze použít $\mathbb{E}(X-\mu)$ jako míru rozptylu, neboť $\mathbb{E}(X-\mu) = \mathbb{E}(X) - \mu = \mu - \mu = 0$. Občas se používá $\mathbb{E}|X-\mu|$, častěji však variance.


1.0
0.8
0.6
0.4
0.2
0.0
$\varphi_{\mu,\sigma^2}(x)$
−5
−4
−3
−2
−1
0
1
2
3
4
5
x
$\mu = 0, \quad \sigma^2 = 0.2,$
$\mu = 0, \quad \sigma^2 = 1.0,$
$\mu = 0, \quad \sigma^2 = 5.0,$
$\mu = -2, \quad \sigma^2 = 0.5,$

Směrodatná odchylka

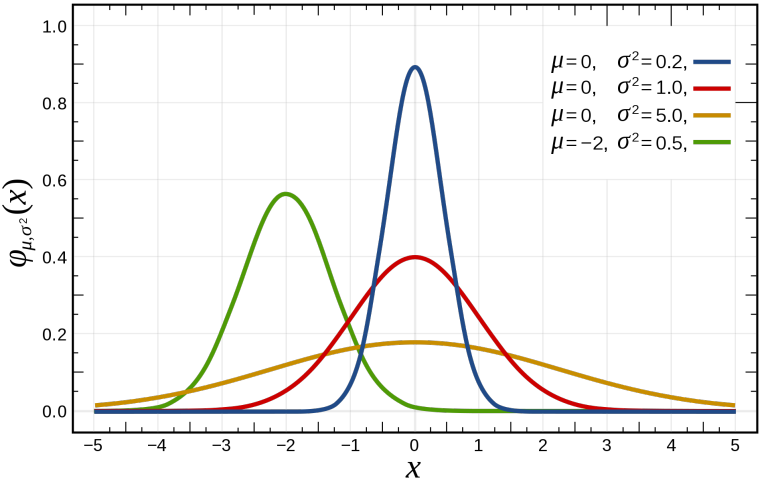
Definice 91.

Směrodatná odchylka náhodné veličiny X se značí jako σ, σ_X či $sd(X)$ a je definovaná jako

$$\sqrt{Var(X)}\,.$$

Poznámka. Směrodatná odchylka vypovídá o tom, nakolik se od sebe navzájem typicky liší jednotlivé případy v souboru zkoumaných hodnot.

- Je-li malá, jsou si prvky souboru většinou navzájem podobné.
- Je-li velká, signalizuje to velké vzájemné odlišnosti.



Směrodatná odchylka σ je postupně $0,447$; 1; $2,236$ a $0,707$.

Věta 92.

Variance má následující vlastnosti:

1. $Var(X) = \mathbb{E}(X^2) - \mu^2$
2. *Pokud jsou* a, b *konstanty, pak* $Var(aX + b) = a^2 Var(X)$.
3. *Pokud jsou* $X_1, \ldots, X_n$ *nezávislé a* $a_1, \ldots, a_n$ *jsou konstanty, pak*

$$Var\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 Var(X_i).$$

Důkaz.

1. $Var(X) = \mathbb{E}[(X-\mu)^2] = \sum_x (x-\mu)^2 f(x) = \sum_x (x^2 - 2\mu x + \mu^2) f(x) = \sum_x x^2 f(x) - 2\mu \sum_x x f(x) + \mu^2 \sum_x f(x) = \mathbb{E}[X^2] - 2\mu \mathbb{E}[X] + \mu^2 = \mathbb{E}[X^2] - \mu^2$.
2. $Var(aX + b) = \mathbb{E}[(aX + b - a\mu - b)^2] = \mathbb{E}[a^2(X-\mu)^2] = a^2 Var(X)$.
3. Později. □

Příklad 93.

Nechť $X \sim Binomial(n, p)$ a nechť $X = \sum_i X_i$, kde $X_i = 1$ pokud na i-tý hod padne orel, jinak $X_i = 0$. NV X_i jsou nezávislé a $\mathbb{P}(X_i = 1) = p$ a $\mathbb{P}(X_i = 0) = 1 - p$. Již jsme ukázali, že $\mathbb{E}(X_i) = p$. Proto

$$\mathbb{E}(X_i^2) = p \cdot 1^2 + (1 - p) \cdot 0^2 = p,$$

$$Var(X_i) = \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2 = p - p^2 = p(1 - p),$$

a tedy

$$Var(X) = Var\left(\sum_i X_i\right) = \sum_i Var(X_i) = np(1 - p).$$

Poznámka: Všimněme si, že $Var(X) = 0$, pokud $p = 1$ nebo $p = 0$.

Jestliže $X_1, \ldots, X_n$ jsou NV, definujeme výběrovou střední hodnotu jako

$$\overline{X_n} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

a výběrovou varianci jako

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X_n})^2$$

Věta 94.

Nechť $X_1, \ldots, X_n$ jsou IID[7] *a nechť $\mu = \mathbb{E}(X_i)$ a $\sigma^2 = Var(X_i)$. Pak*

$$\mathbb{E}(\overline{X_n}) = \mu, \qquad Var(\overline{X_n}) = \frac{\sigma^2}{n} \qquad a \qquad \mathbb{E}(S_n^2) = \sigma^2.$$

[7] Jestliže $X_1, \ldots, X_n$ jsou nezávislé NV a každá má stejné marginální rozdělení s distribuční funkcí F, pak říkáme, že $X_1, \ldots, X_n$ jsou IID (independent and identically distributed).

Kovariance a korelace

Pokud jsou X a Y náhodné veličiny, pak kovariance a korelace mezi X a Y určuje, jak silná je lineární závislost mezi X a Y.

Definice 95.

Nechť X a Y jsou náhodné veličiny se střední hodnotou μ_X a μ_Y a směrodatnými odchylkami σ_X a σ_Y. Definujme kovarianci mezi X a Y jako

$$Cov(X, Y) = \mathbb{E}((X - \mu_X)(Y - \mu_Y))$$

a korelaci jako

$$\rho = \rho_{X,Y} = \rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}.$$

Věta 96.

Kovariance splňuje

$$Cov(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

a korelace splňuje

$$-1 \leq \rho(X, Y) \leq 1.$$

Pro $Y = aX + b$, kde a, b jsou konstanty, je

$$\rho(X, Y) = \begin{cases} 1 & \text{pro } a > 0 \\ -1 & \text{pro } a < 0. \end{cases}$$

Pro X a Y nezávislé je $Cov(X, Y) = \rho = 0$; NV X a Y s $\rho(X, Y) = 0$ nazýváme nekorelované (srovnejte následující větu s bodem 3 Věty 92). Opak obecně neplatí.

Věta 97.

Nechť X a Y jsou náhodné veličiny, pak

1. $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$.
2. $Var(X-Y) = Var(X) + Var(Y) - 2Cov(X,Y)$.
3. *Obecně, pro náhodné veličiny $X_1, \ldots, X_n$,*

$$Var\left(\sum_i a_i X_i\right) = \sum_i a_i^2 Var(X_i) + 2\sum\sum_{i<j} a_i a_j Cov(X_i, X_j).$$

- $Cov(X, Y) > 0$: NV X a Y jsou závislé v pozitivním smyslu
  - vyšší hodnoty X jsou svázány s vyššími hodnotamy Y (a nižší s nižšími)
    - např. výška a váha člověka
- $Cov(X, Y) < 0$: NV X a Y jsou závislé v negativním smyslu
  - vyšší hodnoty X jsou svázány s nižšími hodnotamy Y (a nižší s vyššími)
    - např. hloubka dezénu pneu a brzdná dráha
- Korelace je kovariance normovaná na interval $[-1, 1]$
  - umožňuje lepší srovnání a vyjadřuje lineární závislost
- Velká absolutní hodnota $\rho(X, Y)$ vyjadřuje velkou míru lineární závislosti NV X a Y
  - Vysoká hodnota $\rho(X, Y)$: hodnoty obou veličin se vyvíjejí podobně, nemusí ale mezi nimi existovat příčinný vztah!
- Nízká absolutní hodnota $\rho(X, Y)$ vyjadřuje, že X a Y jsou téměř nekorelované, tj. jsou nezávislé, nebo jejich závislost není lineární.

Příklad 98.

Uvažme rodinu se třemi dětmi. Nechť X značí počet dcer a Y značí počet starších bratrů nejmladšího dítěte se sdruženou pravděpodobností

	$Y=0$	$Y=1$	$Y=2$	
$X=0$	0	0	1/8	1/8
$X=1$	0	1/4	1/8	3/8
$X=2$	1/8	1/4	0	3/8
$X=3$	1/8	0	0	1/8
	1/4	1/2	1/4	1

Pak

- $\mathbb{E}[X] = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = 1,5$ a $\mathbb{E}[Y] = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1.$
- $\mathbb{E}[XY] = 0 \cdot \mathbb{P}(X = 0 \vee Y = 0) + 1 \cdot \mathbb{P}(X = 1, Y = 1) + 2 \cdot (\mathbb{P}(X = 1, Y = 2) + 2 \cdot \mathbb{P}(X = 2, Y = 1)) = 1 \cdot \frac{1}{4} + 2 \cdot \frac{3}{8} = 1$
- $Cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 1 - 1,5 = -0,5,$

odkud X a Y nejsou nezávislé. Platí tak, že čím více je dcer, tím méně je starších bratrů.

Střední hodnota a variance důležitých NV

rozdělení	střední hodnota	variance (rozptyl)
Bodové v a	a	0
$Bernoulli(p)$	p	$p(1-p)$
$Binomial(n,p)$	np	$np(1-p)$
$Geometric(p)$	$1/p$	$(1-p)/p^2$
$Poisson(\lambda)$	λ	λ
$Uniform(a,b)$	$(a+b)/2$	$(b-a)^2/12$
$Normal(\mu,\sigma^2)$	μ	σ^2
$Exponencial(\beta)$	β	β^2
t_ν	0 pro $\nu > 1$	$\nu/(\nu-2)$ pro $\nu > 2$
χ^2_p	p	$2p$

Nechť $X = (X_1, \ldots, X_k)^T$ je náhodný vektor.

Střední hodnota náhodného vektoru X je definována jako

$$\mu = (\mu_1 \ \ldots, \mu_k)^T = \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_k) \end{pmatrix}.$$

Kovarianční matice (též variančně-kovarianční matice) Σ je definována jako

$$\Sigma(X) = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_k) \\ Cov(X_2, X_1) & Var(X_2) & \cdots & Cov(X_2, X_k) \\ \vdots & \vdots & \vdots & \vdots \\ Cov(X_k, X_1) & Cov(X_k, X_2) & \cdots & Var(X_k) \end{pmatrix},$$

přičemž je tato matice symetrická.

Nerovnosti

Nerovnosti jsou užitečné k ohraničení hodnot, které je obtížné spočítat.

Věta 99 (Markovova nerovnost).

Nechť X je nezáporná náhodná veličina a nechť $\mathbb{E}(X)$ existuje. Pak pro libovolné $t > 0$ platí

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}(X)}{t}.$$

Důkaz.

Protože $X \geq 0$, máme

$$\begin{aligned}\mathbb{E}(X) &= \int_0^\infty x f(x)\,dx = \int_0^t x f(x)\,dx + \int_t^\infty x f(x)\,dx \\ &\geq \int_t^\infty x f(x)\,dx \geq t \int_t^\infty f(x)\,dx = t\mathbb{P}(X \geq t).\end{aligned}$$

□

Věta 100 (Čebyševova nerovnost).

Nechť $\mu = \mathbb{E}(X)$ a $\sigma^2 = Var(X)$. Pak pro $t > 0$ platí

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2} \qquad a \qquad \mathbb{P}(|Z| \geq k) \leq \frac{1}{k^2},$$

kde $Z = (X - \mu)/\sigma$.
Například tedy platí:

- $\mathbb{P}(|Z| > 2) \leq 1/4$
- $\mathbb{P}(|Z| > 3) \leq 1/9$.

Důkaz.

Použijeme Markovovu nerovnost a dostaneme

$$\mathbb{P}(|X - \mu| \geq t) = \mathbb{P}(|X - \mu|^2 \geq t^2) \leq \frac{\mathbb{E}(X - \mu)^2}{t^2} = \frac{\sigma^2}{t^2}.$$

Druhá část plyne z toho, že položíme $t = k\sigma$. □

Příklad 101.

- Testujeme predikční metodu, například neuronovou síť, na n případech.
  - $X_i = \begin{cases} 1 & \text{pokud se prediktor mýlí} \\ 0 & \text{pokud se nemýlí.} \end{cases}$
- Pak $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$ je pozorovaná míra chybovosti.
  - Každé X_i lze považovat za Bernoulliho rozdělení s neznámou střední hodnotou p.
  - Rádi bychom znali správnou, avšak neznámou míru chybovosti p.
  - Intuitivně očekáváme, že $\overline{X_n}$ by mělo být blízko p.
- Jak je pravděpodobné, že $\overline{X_n}$ není v ϵ okolí p?
  - Máme $Var(\overline{X_n}) = Var(X_i)/n = p(1-p)/n$ a

$$\mathbb{P}(|\overline{X_n} - p| \geq \epsilon) \leq \frac{Var(\overline{X_n})}{\epsilon^2} = \frac{p(1-p)}{n\epsilon^2} \leq \frac{1}{4n\epsilon^2},$$

    neboť $p(1-p) \leq 1/4$ pro všechna p.
  - Pro $\epsilon = 0,2$ a $n = 100$ je hranice $0,0625$.

Věta 102 (Hoeffdingova nerovnost).

Nechť $Y_1, \ldots, Y_n$ jsou nezávislá pozorování taková, že $\mathbb{E}(Y_i) = 0$, $a_i \leq Y_i \leq b_i$, nechť $\epsilon > 0$. Pak pro libovolné $t > 0$ je

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i \geq \epsilon\right) \leq e^{-t\epsilon} \prod_{i=1}^{n} e^{t^2(b_i - a_i)^2/8}.$$

Příklad 103.

Nechť $X_1, \ldots, X_n \sim Bernoulli(p)$. Nechť $n = 100$ a $\epsilon = 0,2$.
Čebyševova nerovnost dává

$$\mathbb{P}(|\overline{X_n} - p| \geq 0,2) \leq 0,0625.$$

Hoeffdingova nerovnost dává

$$\mathbb{P}(|\overline{X_n} - p| \geq 0,2) \leq 2e^{-2\cdot(100)\cdot(0,2)^2} = 0,00067,$$

což je mnohem menší než 0,0625.

Hoeffdingova nerovnost dává způsob, jak vytvořit interval spolehlivosti pro binomický parametr p (více o tom později, ve statistice). Fixujme $\alpha > 0$ a nechť

$$\epsilon_n = \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}.$$

Hoeffdingova nerovnost říká, že $\mathbb{P}(|\overline{X_n} - p| \geq \epsilon_n) \leq 2e^{-2n\epsilon_n^2} = \alpha$. Nechť

$$C = (\overline{X_n} - \epsilon_n, \overline{X_n} + \epsilon_n).$$

Pak

$$\mathbb{P}(p \notin C) = \mathbb{P}(|\overline{X_n} - p| \geq \epsilon_n) \leq \alpha, \quad \text{tedy} \quad \mathbb{P}(p \in C) \geq 1 - \alpha.$$

Tedy náhodně zvolený interval C obsahuje hodnotu parametru p s pravděpodobností $1 - \alpha$. Dodejme, že C se nazývá interval spolehlivosti s koeficientem spolehlivosti α.

Nerovnosti pro střední hodnoty

Věta 104 (Cauchyho-Schwarzova nerovnost).

Jestliže X a Y mají konečné variance, pak

$$\mathbb{E}(|XY|) \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}.$$

- Pokud je $g''(x) \geq 0$ pro všechna x, pak g je konvexní.
- Pokud je g konvexní, pak g leží nad tečnou.
- Funkce g je konkávní, jestliže $-g$ je konvexní.
  - Příklady konvexních funkcí: $g_1(x) = x^2$ a $g_2(x) = e^x$.
  - Příklady konkávních funkcí: $g_3(x) = -x^2$ a $g_4(x) = \ln x$.

Věta 105 (Jensenova nerovnost).

- *Jestliže g je konvexní, pak* $\mathbb{E}(g(X)) \geq g(\mathbb{E}(X))$.
- *Jestliže g je konkávní, pak* $\mathbb{E}(g(X)) \leq g(\mathbb{E}(X))$.

Důsledek 106.

- *Platí, že* $\mathbb{E}(X^2) \geq (\mathbb{E}(X))^2$.
- *Pokud je X pozitivní, pak* $\mathbb{E}(1/X) \geq 1/\mathbb{E}(X)$.
- *Je-li* log *konkávní, je* $\mathbb{E}(\log X) \leq \log \mathbb{E}(X)$.

Konvergence náhodných veličin

- Důležitá část teorie pravděpodobnosti se věnuje chování sekvence náhodných veličin.
- Říká se jí large sample theory (limitní teorie, asymptotická teorie).
  - Otázkou je, co můžeme říct o limitním chování posloupnosti náhodných veličin $X_1, X_2, X_3, \ldots$?

- Statistika a data mining úzce souvisí se získáváním dat.
- Co se děje, když máme více a více dat?

- V analýze posloupnost reálných čísel x_n konverguje k limitě x, jestliže pro každé $\epsilon > 0$ je $|x_n - x| < \epsilon$ pro všechna dostatečně velká n.
  - Je-li například $x_n = x$ pro všechna n, pak zřejmě $\lim_{n\to+\infty} x_n = x$.

- V pravděpodobnosti je situace trochu komplikovanější.
- Nechť $X_1, X_2, \ldots$ jsou nezávislé NV každá s rozdělením $N(0,1)$.
  - Jelikož všechny NV mají stejné rozdělení, zdálo by se, že X_n „konverguje“ k $X \sim N(0,1)$.
- To ale není v pořádku, neboť $\mathbb{P}(X_n = X) = 0$ pro všechna n.
  - Dvě spojité NV jsou si rovny s pravděpodobností nula.
  - $\mathbb{P}(X = Y) = \mathbb{P}(X - Y = 0) = 0$.

- Jiný příklad.
  - Nechť $X_1, X_2, \ldots$, kde $X_i \sim N(0, 1/n)$
  - X_n je koncentrováno kolem 0 pro velká n
  - Chtěli bychom proto říct, že X_n konverguje k 0.
    - Ale $\mathbb{P}(X_n = 0) = 0$ pro všechna n.

- Potřebujeme nějaký nástroje k definici konvergence.

Zde se podíváme na následující dva:

- Zákon velkých čísel

  říká, že výběrový průměr $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ konverguje v pravděpodobnosti ke střední hodnotě $\mu = \mathbb{E}(X_i)$, tedy, že $\overline{X}_n$ je blízko μ s velkou pravděpodobností.

- Centrální limitní věta

  říká, že $\sqrt{n}(\overline{X}_n - \mu)$ konverguje v rozdělení k normálnímu rozdělení, tedy, že výběrový průměr má přibližně normální rozdělení pro velká n.

Typy konvergence

Definice 107.

Nechť $X_1, X_2, \ldots$ je posloupnost NV a X je další NV. Nechť F_n značí distribuční funkci X_n a F distribuční funkci X.

1. X_n konverguje k X v pravděpodobnosti, $X_n \xrightarrow{P} X$, jestliže pro každé $\epsilon > 0$ platí

$$\mathbb{P}(|X_n - X| \geq \epsilon) \to 0$$

   pro $n \to +\infty$.

2. X_n konverguje k X v rozdělení, $X_n \rightsquigarrow X$, jestliže

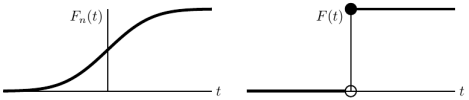
$$\lim_{n \to +\infty} F_n(t) = F(t)$$

   pro všechna t, kde je F spojitá.

Pokud je limitní NV bodová, $\mathbb{P}(X = c) = 1$ a $X_n \xrightarrow{P} X$, píšeme $X_n \xrightarrow{P} c$.
Podobně pro $X_n \rightsquigarrow X$ píšeme $X_n \rightsquigarrow c$.

Příklad 108 (Konvergence v rozdělení).

- Nechť $X_n \sim N(0, 1/n)$. Konverguje X_n k 0? Platí $\sqrt{n}X_n \sim N(0,1)$?
- Nechť F je distribuční funkce s bodovým rozdělením v 0.
- Nechť Z je standardní normální NV.
  - Pro $t < 0$ je $F_n(t) = \mathbb{P}(X_n < t) = \mathbb{P}(\sqrt{n}X_n < \sqrt{n}t) = \mathbb{P}(Z < \sqrt{n}t) \to 0$, neboť $\sqrt{n}t \to -\infty$.
  - Pro $t > 0$ je $F_n(t) = \mathbb{P}(X_n < t) = \mathbb{P}(\sqrt{n}X_n < \sqrt{n}t) = \mathbb{P}(Z < \sqrt{n}t) \to 1$, neboť $\sqrt{n}t \to +\infty$.
  - Tedy $F_n(t) \to F(t)$ pro všechna $t \neq 0$, tj. $X_n \rightsquigarrow 0$.
- Avšak $F_n(0) = 1/2 \neq F(0) = 1$, a proto konvergence neplatí v $t = 0$. Na tom ale nezáleží, protože v $t = 0$ není F spojitá a definice konvergence v rozdělení vyžaduje konvergenci v bodech, kde je funkce spojitá.



Příklad 109 (Konvergence v pravděpodobnosti).

- Nechť $X_n \sim N(0, 1/n)$. Konverguje X_n k 0?
- Opět je F distribuční funkce s bodovým rozdělením v 0.
- Jak je to s konvergencí v pravděpodobnosti?
- Pro libovolné $\epsilon > 0$ dává Markovova nerovnost

$$\mathbb{P}(|X_n| \geq \epsilon) = \mathbb{P}(|X_n|^2 \geq \epsilon^2) \leq \frac{\mathbb{E}(X_n^2)}{\epsilon^2} = \frac{\frac{1}{n}}{\epsilon^2} \to 0$$

  pro $n \to +\infty$.
- Tedy $X_n \xrightarrow{P} 0$.

Následující věta popisuje vztah mezi typy konvergence.

Věta 110.

Platí, že:

$$X_n \xrightarrow{P} X \text{ implikuje, že } X_n \rightsquigarrow X.$$

Opačná implikace neplatí, až na následující speciální případ:

Jestliže $X_n \rightsquigarrow X$ *a* $\mathbb{P}(X = c) = 1$ *pro nějaké reálné* c*, pak* $X_n \xrightarrow{P} X$.

Věta 111.

Nechť X_n, X, Y_n, Y jsou náhodné veličiny. Nechť g je spojitá funkce.

1. *Jestliže $X_n \xrightarrow{P} X$ a $Y_n \xrightarrow{P} Y$, pak $X_n + Y_n \xrightarrow{P} X + Y$.*
2. *Jestliže $X_n \rightsquigarrow X$ a $Y_n \rightsquigarrow c$, pak $X_n + Y_n \rightsquigarrow X + c$.*
3. *Jestliže $X_n \xrightarrow{P} X$ a $Y_n \xrightarrow{P} Y$, pak $X_n Y_n \xrightarrow{P} XY$.*
4. *Jestliže $X_n \rightsquigarrow X$ a $Y_n \rightsquigarrow c$, pak $X_n Y_n \rightsquigarrow cX$.*
5. *Jestliže $X_n \xrightarrow{P} X$, pak $g(X_n) \xrightarrow{P} g(X)$.*
6. *Jestliže $X_n \rightsquigarrow X$, pak $g(X_n) \rightsquigarrow g(X)$.*

Obecně $X_n \rightsquigarrow X$ a $Y_n \rightsquigarrow Y$ neimplikuje $X_n + Y_n \rightsquigarrow X + Y$.

Zákon velkých čísel

Zákon velkých čísel je jeden z hlavních výsledků teorie pravděpodobnosti. Říká, že střední hodnota velkého výběru se blíží střední hodnotě rozdělení. Například při velkém počtu hodů padne orel kolem poloviny případů.
Nechť $X_1, X_2, \ldots$ jsou IID NV. Nechť $\mu = \mathbb{E}(X_i)$ a $\sigma^2 = Var(X_i)$. Zopakujme, že výběrový průměr $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ a že $\mathbb{E}(\overline{X}_n) = \mu$ a $Var(\overline{X}_n) = \sigma^2/n$.

Věta 112 (Slabý zákon velkých čísel).

Jestliže $X_1, \ldots, X_n$ jsou IID, pak $\overline{X}_n \xrightarrow{P} \mu$.

Tedy rozdělení $\overline{X}_n$ se začíná koncentrovat kolem μ s rostoucím n.

Důkaz.

Předpokládejme, že $\sigma < +\infty$. Tento předpoklad není nutný, ale zjednodušuje důkaz.
Z Čebyševovy nerovnosti máme

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \epsilon) \leq \frac{Var(X_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2},$$

což jde k 0 pro $n \to +\infty$. □

Příklad 113.

Uvažujme hody mincí, kde orel padá s pravděpodobností rovnou p. Nechť X_i je výsledek jednoho hodu (0 nebo 1), tedy

$$p = \mathbb{P}(X_i = 1) = \mathbb{E}(X_i).$$

Poměr orlů po n hodech je $\overline{X}_n$. Zákon velkých čísel říká, že $\overline{X}_n$ konverguje v pravděpodobnosti k p. To však neznamená, že $\overline{X}_n$ se bude rovnat p, ale že pro velká n bude rozdělení $\overline{X}_n$ koncentrované těsně kolem p.

Nechť $p = 1/2$. Jak velké musí být n, aby $\mathbb{P}(0,4 \leq \overline{X}_n \leq 0,6) \geq 0,7$?

Máme $\mathbb{E}(\overline{X}_n) = p = 1/2$ a $Var(\overline{X}_n) = \sigma^2/n = p(1-p)/n = 1/(4n)$. Čebyševova nerovnost pak dává

$$\begin{aligned}\mathbb{P}(0,4 \leq \overline{X}_n \leq 0,6) &= \mathbb{P}(|\overline{X}_n - \mu| \leq 0,1) = 1 - \mathbb{P}(|\overline{X}_n - \mu| > 0,1) \\ &\geq 1 - \frac{1}{4n(0,1)^2} = 1 - \frac{25}{n},\end{aligned}$$

což je větší než $0,7$ pro $n = 84$.

Silný zákon velkých čísel

Zatímco slabý zákon velkých čísel říká, že $\overline{X}_n$ konverguje v pravděpodobnosti ke střední hodnotě $\mathbb{E}(X_i)$, silný zákon velkých čísel říká, že skoro jistě konverguje ke střední hodnotě.

Věta 114 (Silný zákon velkých čísel).

Nechť $X_1, X_2, \ldots$ jsou IID. Jesliže $\mu = \mathbb{E}(|X_i|) < +\infty$, pak

$$\mathbb{P}\left(\lim_{n\to+\infty} \frac{X_1 + \cdots + X_n}{n} = \mu\right) = 1.$$

Slabý vs. silný zákon velkých čísel

Slabý zákon velkých čísel říká, že pro libovolně specifikovanou velkou hodnotu n^* je $(X_1 + \cdots + X_{n^*})/n^*$ blízko μ. Neříká však, že $(X_1 + \cdots + X_n)/n$ musí být blízko μ pro všechny hodnoty $n > n^*$, tj. připouští možnost, že se velké hodnoty

$$\left| \frac{X_1 + \cdots + X_n}{n} - \mu \right|$$

mohou vyskytnout nekonečně často.

Silný zákon říká, že toto nemůže nastat, tedy, že s pravděpodobností 1 bude

$$\left| \frac{X_1 + \cdots + X_n}{n} - \mu \right|$$

větší než jakékoliv $\epsilon > 0$ pouze konečně mnohokrát.

Centrální limitní věta

- Zákon velkých čísel říká, že rozdělení $\overline{X}_n$ jde k μ.
- To nám nepomůže aproximovat tvrzení o $\overline{X}_n$.
- K tomu potřebujeme centrální limitní větu.

- Nechť $X_1, \ldots, X_n$ jsou IID se střední hodnotou μ a variancí σ^2.
- Centrální limitní věta říká, že $\overline{X}_n = \frac{1}{n}\sum_i X_i$ má rozdělení, které je přibližně rovno normálnímu rozdělení se střední hodnotou μ a variancí σ^2/n.
- Toto je pozoruhodné, neboť o rozdělení X_i nic nepředpokládáme, mimo toho, že střední hodnoty a variance existují.

Věta 115 (Centrální limitní věta).

Nechť $X_1, \ldots, X_n$ jsou IID se střední hodnotou μ a variancí σ^2. Nechť $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Pak tzv. normalizovaná či standardizovaná veličina

$$Z_n \equiv \frac{\overline{X}_n - \mu}{\sqrt{Var(\overline{X}_n)}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \rightsquigarrow Z,$$

kde $Z \sim N(0,1)$. Jinak řečeno,

$$\lim_{n\to+\infty} \mathbb{P}(Z_n \leq z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2}\, dx.$$

Poznámka. Pravděpodobnostní tvrzení o $\overline{X}_n$ lze aproximovat pomocí normálního rozdělení. Aproximujeme pravděpodobnostní tvrzení, nikoliv samotnou náhodnou veličinu.

Příklad 116.

- Předpokládejme, že počet chyb na jeden počítačový program má Poissonovo rozdělení se střední hodnotou 5.
- Dostaneme 125 programů.
- Nechť $X_1, \ldots, X_{125}$ jsou počty chyb v programech.
- Budeme aproximovat $\mathbb{P}(\overline{X}_n < 5,5)$.
- Nechť $\mu = \mathbb{E}(X_i) = \lambda = 5$ a $\sigma^2 = Var(X_i) = \lambda = 5$.
- Pak $\mathbb{P}(\overline{X}_n < 5,5) = \mathbb{P}\left(\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} < \frac{\sqrt{n}(5,5 - \mu)}{\sigma}\right) \approx \mathbb{P}(Z < 2,5) \approx 0,9938.$

- Centrální limitní věta říká, že
$$Z_n = \sqrt{n}(\overline{X}_n - \mu)/\sigma$$
je přibližně $N(0,1)$.
- My však neznáme σ.
- Později uvidíme, že σ^2 lze odhadnout z $X_1, \ldots, X_n$ pomocí výběrové variance
$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Platí centrální limitní věta, pokud nahradíme σ za S_n? Odpověď je ano.

Věta 117.

Předpokládejme stejné podmínky jako u centrální limitní věty. Pak

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \rightsquigarrow N(0,1).$$

Popisná statistika

Statistika

- (Popisná) statistika = odvození (číselných) charakteristik o datech a jejich vizualizace.
  - Například roční příjmy občanů podle dat finančních úřadů.

- Matematická statistika = použití matematických metod pro odvozování závěrů platných pro celý soubor objektů na základě malého vzorku (společně s kvalitativním odhadem věrohodnosti výsledného sdělení).
  - Například choroby populace z dat získaných u několika nahodile vybraných osob.
  - Pro daná data zjišťujeme, jaké vlastnosti popisované objekty mají a jak věrohodné jsou odvozené výsledky.

Popisná statistika: Příklad

Příklad 118.

- Soubor objektů mohou být studenti nějakého základního kurzu, jako číselné údaje pak můžeme zkoumat:
  - průměrný počet bodů získaný z předmětu v minulém semestru a rozptyl těchto hodnot
  - průměrné dosažené známky z tohoto a jiných předmětů a korelace mezi výsledky
  - korelace dat vypovídajících o historii dřívějšího studia u konkrétních studentů
  - korelace neúspěchů ve studiu a počtu hodin týdně odpracovaných studentem mimo fakultu.

Poznámky ke statistikám posuzovaných veličin

- Aritmetický průměr bodů říká málo o kvalitě přednášky. Zajímavější hodnotou je počet bodů, pro které je stejně studentů pod ní i nad ní (či první a poslední čtvrtina, desetina, atp.). Tyto hodnoty nazýváme statistiky.
- Rozumné hodnocení by mělo mít normální rozdělení.
- Z číselných hodnot statistik pro konkrétní výběr lze kvalitativně popsat věrohodnost závěrů.
  - Například pokud výsledky hodnocení nevykazují dostatečnou variabilitu, jde o náznak, že s předmětem není něco v pořádku.
- Jak je to s věrohodností zpracovávaných dat?
  - Data mohou být nepřesná v důsledku nevhodné konstrukce experimentu a samotného sběru dat.
- V mnoha případech nic nevíme o charakteru rozdělení dat.

Popisná statistika

- V popisné statistice máme k dispozici nástroje, které umožňují dobře porozumět struktuře a povaze i velmi rozsáhlých dat.
- V matematice pracujeme s abstraktním matematickým popisem pravděpodobnosti, který je použitelný pro analýzu daných dat, zejména když máme k dispozici teoretický model, kterému mají odpovídat.
- Závěry statických šetření na vzorcích konkrétních souborů dat může dát matematická statistika.
- Do jaké míry je takový popis adekvátní pro konkrétní výběr dat je možné vyjádřit pomocí metod matematické statistiky.

Terminologie

- Statistický soubor je přesně definovaná množina základních statistických jednotek, která je dána výčtem nebo nějakými pravidly.
  - Statistickým souborem jsou například všichni obyvatelé Olomouce, kde každý obyvatel zvlášť je statistickou jednotkou.
- Na každé statistické jednotce měříme jeden nebo více statistických znaků, například číselné hodnoty jako výška, váha, věk atd.
- Základním objektem pro zkoumání jednotlivých znaků je soubor hodnot, zpravidla ve formě uspořádaných hodnot, přičemž k porovnávání a poměřování jednotlivých hodnot potřebujeme měřítko.

Typy měřítek znaků

Hodnoty mohou být následujících typů:

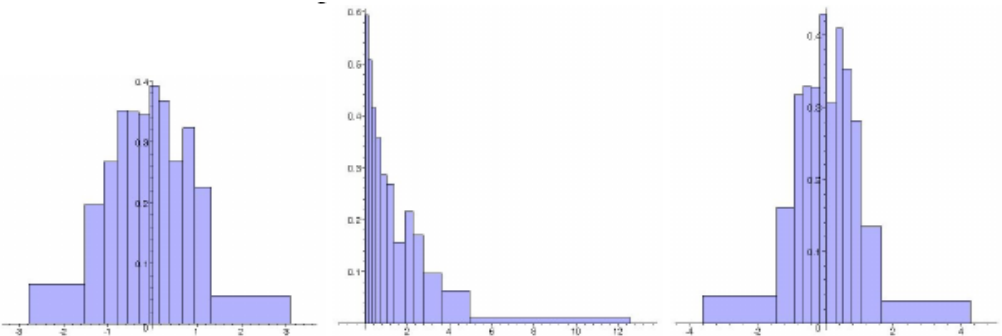
- nominální – mezi hodnotami není žádný vztah, jde pouze o označení možných hodnot
  - například názvy politických stran
  - jsme schopni interpretovat pouze rovnost $x = y$.
- ordinální – hodnoty s uspořádáním
  - výška, váha, počet hvězdiček u hotelů atd.
  - jsme schopni interpretovat rovnost a nerovnost $x < y$, případně $x > y$.
- intervalové – číselné hodnoty, kde jde o porovnání velikostí, nikoliv o absolutní hodnotu
  - umíme posoudit i rozdíl $x - y$.
- poměrové – pevně stanovené měřítko a nula
  - většina fyzikálních nebo ekonomických veličin
  - máme k dispozici rovnost, nerovnost, rozdíl i podíl x/y.

Uspořádání hodnot

- Mějme soubor hodnot $x_1, x_2, \ldots, x_n$, které lze uspořádat, a které vznikly měřením n statistických jednotek.
  - Uspořádáme je do uspořádaného souboru hodnot $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.
  - Číslo n nazýváme rozsah souboru.
- Pokud pracujeme s rozsáhlými soubory znaků, které připouští málo hodnot, uvádíme pouze četnosti výskytu.
  - U průzkumu preferencí politických stran uvádíme u každé možné hodnoty počet jejích výskytů.
- Pokud je naopak možných hodnot mnoho, dělíme možný rozsah hodnot na vhodný počet intervalů a o statistickém znaku uvádíme četnost hodnot v daných intervalech.
  - Například mzda mezi 1500 až 2000 EUR, výška 180 až 190 cm, atd.
  - Intervalům se říká třídy a počtu znaků ve třídě třídní četnosti.
    - Používáme také kumulativní četnosti a kumulativní třídní četnosti, které pro danou třídu vznikají součtem třídních četností s hodnotami nejvýše jako má ta daná třída.
  - Nejčastěji uvažujeme střed a_i dané třídy za hodnotu, která ji reprezentuje.
    - Hodnota $a_i n_i$, kde n_i je četnost výskytu této třídy, představuje celkový příspěvek této třídy.
  - Místo četností často zobrazujeme relativní četnosti $\frac{a_i}{n}$, resp. relativní kumulativní četnosti.

Vizualizace

- Graf, který na jedné ose vynáší intervaly jednotlivých tříd a nad nimi obdélníky s výškou rovnou četnosti se nazývá histogram.
  - Obdobně se znázorňuje kumulativní četnost.
- Na obrázku jsou histogramy souborů o rozsahu $n = 500$, které vznikly náhodným generováním dat s různými standardními rozděleními (normální, χ^2 a studentovo).



Míry polohy statistických znaků – průměry

Mějme (nesetříděný) soubor $(x_1, \ldots, x_n)$ hodnot měřeného znaku pro všechny zpracovávané statistické jednotky a nechť $n_1, \ldots, n_m$ jsou třídní četnosti m možných hodnot $a_1, \ldots, a_m$.

Definice 119.

Aritmetický průměr (často jen průměr): $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{j=1}^{m} n_j a_j$.

Geometrický průměr: $\bar{x}^G = \sqrt[n]{x_1 x_2 \cdots x_n}$ a má smysl pouze u kladných hodnot znaků.

Harmonický průměr: $\bar{x}^H = \left(\frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i}\right)^{-1}$ a je definován jen pro kladné hodnoty znaků.

- Platí $\bar{x}^H \leq \bar{x}^G \leq \bar{x}$.
- Aritmetický průměr je invariantní vůči afinním transformacím:
  - pro lib. skaláry a, b platí $\overline{(a + b \cdot x)} = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) = a + \frac{b}{n}\sum_{i=1}^{n} x_i = a + b \cdot \bar{x}$, je tedy vhodný pro intervalové typy měřítek.
- Logaritmus geometrického průměru znaků je aritmetický průměr logaritmů znaků.
  - Je vhodný pro znaky, které se kumulují multiplikativně, jako například úrokové míry.
    - Je-li úroková míra v jednotlivých časových jednotkách x_i %, bude za celé období výsledek takový, jakoby byla po celou dobu konstantní úroková míra $\bar{x}^G$ %.

Příklad na průměrnou rychlost

Příklad 120.

Auto jelo z Brna do Prahy rychlostí 160 km/h a z Prahy do Brna rychlostí 120 km/h. Jakou jelo průměrnou rychlostí?

Řešení. Pro průměrnou rychlost musí platit, že auto jedoucí touto rychlostí stráví na trase stejnou dobu. Označíme-li d vzdálenost obou měst v kilometrech a v_p průměrnou rychlost, tak

$$\frac{d}{160} + \frac{d}{120} = \frac{2d}{v_p},$$

odkud

$$v_p = \frac{2}{\frac{1}{160} + \frac{1}{120}} = 137,14\,.$$

Průměrná rychlost je tedy harmonický průměr jednotlivých průměrných rychlostí.

Medián, kvartil, decil, percentil,...

- Další způsob vyjádření míry hodnot nabývaných znaky je pro parametr $\alpha \in (0,1)$ nalezení hodnoty x_α tak, aby 100α % hodnot znaku bylo nejvýše x_α a zbylé hodnoty byly větší než x_α. Číslu x_α říkáme α-kvantil
  - Pokud takový znak není určen jednoznačně, volíme průměr mezi krajními hodnotami.
- Nejobvyklejší hodnoty x_α jsou:
  - medián (výběrový medián) definovaný vztahem $\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & \text{pro liché } n \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) & \text{pro sudé } n, \end{cases}$

    kde $x_{(k)}$ představuje hodnotu v uspořádaném souboru hodnot.
  - dolní a horní kvartil $Q_1 = x_{0,25}$ a $Q_3 = x_{0,75}$.
  - p-tý kvantil (výběrový kvantil či percentil) x_p, pro $0 < p < 1$.
  - modus definovaný jako hodnota $\hat{x}$ znaku s největší četností v souboru x.
- Aritmetický průměr, medián a modus představují očekávatelné hodnoty znaků.
  - Průměr u znaku podílového typu, medián u poměrového typu a modus u typu ordinálního nebo nominálního.

- Rozptyl souboru znaků x je definován vztahem

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

  Směrodatná odchylka s_x je odmocnina z výběrového rozptylu.
- Někdy se místo s_x^2 používá tzv. výběrový rozptyl, který se liší tím, že se ve jmenovateli zlomku používá $(n-1)$.
- V případě třídních četností n_j hodnot a_j pro m tříd dává stejný výraz hodnotu rozptylu

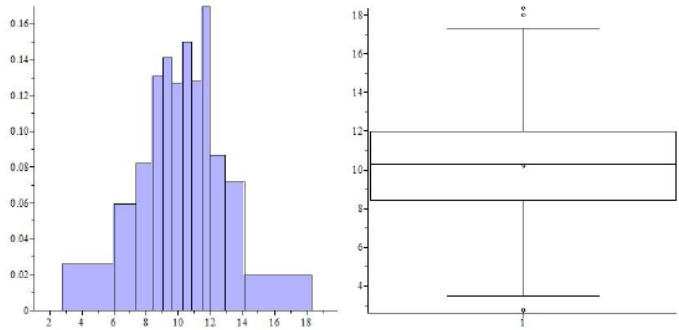
$$s_x^2 = \frac{1}{n}\sum_{i=1}^{m} n_j(a_j - \bar{x})^2.$$

- Dále se můžeme setkat s tzv. rozpětím výběru $R = x_{(n)} - x_{(1)}$ a kvartilovým rozpětím výběru $Q = Q_3 - Q_1$.
- Používá se i průměrná odchylka, která je dána průměrnou vzdáleností hodnot od mediánu

$$D_x = \frac{1}{n}\sum_{i=1}^{n}|x_i - \tilde{x}|.$$

Diagramy

Pro zobrazení statistiky jednotlivých znaků nebo jejich korelací se používá mnoho standardizovaných nástrojů. Jedním z nich jsou tzv. krabicové diagramy.



Na obrázku je zobrazen histogram a krabicový diagram stejného souboru hodnot.
Střední linka je medián, kraje boxu jsou kvartily, packy ukazují 1,5 kvartilového rozsahu, ne však víc než kraje rozsahu výběru, případné hodnoty mimo jsou přímo naznačeny body.

Příklad 121.

V rybníku se vylovilo 425 kaprů a u všech byly zjištěny jejich hmotnosti. Pak se vhodně zvolily hmotnostní intervaly a sestavila se následující tabulka četností:

Hmotnost (kg)	0–1	1–2	2–3	3-4	4–5	5–6	6–7
Střed třídy	0,5	1,5	2,5	3,5	4,5	5,5	6,5
Četnost	75	90	97	63	48	42	10

Načrtněte histogram, určete aritmetický, geometrický a harmonický průměr hmotnosti kaprů. Dále určete medián, horní a dolní kvartil, modus, rozptyl, směrodatnou odchylku, variační koeficient a načrtněte příslušný krabicový diagram.

Matematická statistika

Matematická statistika

- Pracuje s výběrem základního souboru a snaží se popsat míru relevantnosti zjištěných statistik, případně zjistit či upřesnit vhodný teoretický model pro chování celého souboru a z něj pak třeba odhadovat pravděpodobnost nějakého budoucího jevu.
  - Pokud například padne k orlů v n hodech mincí, tak můžeme chtít vyvodit s jakou pravděpodobností padne v následujících dvou hodech orel.
- Existují dva základní přístupy:
  - klasická (frekvenční) statistika
  - Bayesovská statistika.

Klasická (frekvenční) statistika

- Vychází z toho, že pravděpodobnosti jsou dány četnostmi výskytů jevů ve velkých vzorcích dat – lze je tedy aproximovat nekonečnými modely – a využít pro odhady spolehlivosti centrální limitní větu.
  - Na pravděpodobnost pohlíží jako na idealizaci relativní četnosti případů, v nichž se vyskytne určitý výsledek při opakovaných pokusech.
  - Tato zdánlivá výhoda/rigoróznost se může ale rychle stát nevýhodou, jakmile se začneme zabývat spolehlivostí samotných dat a vhodností zvoleného experimentu.
  - Stejně tak je obtížné frekvenční statistiku dobře použít pro odhad pravděpodobnosti výskytu jednorázového děje.
- U hodu mincí vychází z předpokladu, že jednotlivé hody jsou nezávislé a u všech je stejná pravděpodobnost orla dána parametrem $\theta = p$ (který však neznáme).

Bayesovská statistika

- Je příkladem matematizace „selského rozumu“, kdy chceme naše původní přesvědčení postupně pozměňovat ve světle nových dat.
  - Historicky byl první Bayesovský přístup (např. Laplace a další již v 18. století), který byl prakticky zcela vystřídán frekvenční statistikou ve 20. století.
  - V posledních desetiletích se však Bayesovská statistika vrátila, společně s dalšími novými přístupy.
- U hodu mincí považuje Bayesovská statistika parametr θ za náhodnou proměnnou, data získaná experimentem za konstanty a pokouší se z nich vydedukovat informace o rozložení pravděpodobnosti náhodné veličiny θ.

Klasická (frekvenční) statistika

Náhodný výběr z populace

- Mějme (velký) základní statistický soubor s N jednotkami, tzv. populaci a nějaký číselný znak pro každou z jednotek, tedy soubor hodnot

$$(x_1, \ldots, x_N).$$

  - Z něj máme k dispozici výběrový soubor s hodnotami $(X_1, \ldots, X_n)$.
  - Abychom se vyhnuli diskusi skutečné velikosti základního statistického souboru s N jednotkami, budeme předpokládat, že vybíráme položky výběrového souboru jednu po druhé a každou vybranou jednotku poté do populace vracíme.
  - Zároveň předpokládáme, že každá položka má stejnou pravděpodobnost výběru $1/N$.
  - Hovoříme pak o náhodném výběru.

Náhodný výběr z populace

- Způsob realizace náhodného výběru nyní interpretujeme tak, že pracujeme s vektorem $(X_1, \ldots, X_n)$ nezávislých náhodných veličin a že všechny tyto veličiny mají stejné rozdělení pravděpodobnosti.
  - Zejména tedy sdílí distribuční funkci $F_X(x)$ a momenty $E(X_i) = \mu$ a $var(X_i) = \sigma^2$.
- Dalším krokem je odvození charakteristik výběrového průměru $\overline{X_n}$ a výběrového rozptylu

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X_n})^2.$$

Podle následující věty volíme koeficient $\frac{1}{n-1}$ místo $\frac{1}{n}$ proto, aby $\mathbb{E}(S_n^2) = \sigma^2$.

Věta 122.

Pro výběrový průměr $\overline{X_n}$ spočítaný z náhodného výběru rozsahu n z rozdělení s konečnou střední hodnotou μ a konečným rozptylem σ^2 platí $\mathbb{E}(\overline{X_n}) = \mu$, $Var(\overline{X_n}) = \sigma^2/n$, a pro výběrový rozptyl S_n^2 platí $\mathbb{E}(S_n^2) = \sigma^2$.

Náhodný výběr z normálního rozdělení

- V praktických úlohách je třeba znát nejen číselné charakteristiky výběrového průměru a rozptylu, ale jejich úplné rozdělení pravděpodobnosti.
  - To můžeme odvodit pouze pokud známe rozdělení pravděpodobnosti X_i.
  - Jako užitečnou ilustraci si ukažme výsledek pro náhodný výběr z normálního rozdělení.

Věta 123.

Je-li $(X_1, \ldots, X_n)$ *náhodný výběr z rozdělení* $N(\mu, \sigma^2)$*, pak jsou* $\overline{X_n}$ *a* S_n^2 *nezávislé veličiny a platí*

$$\overline{X_n} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad a \qquad \frac{n-1}{\sigma^2} S_n^2 \sim \chi^2_{n-1}\,.$$

Bodové a intervalové odhady

Bodové a intervalové odhady

- Uvažme anketu mezi 500 studenty o spokojenosti s kurzem ve formě bodů od 1 do 10.
  - Spokojenost jednotlivých studentů X_i je aproximována náhodnou veličinou s rozdělením $N(\mu, \sigma^2)$, přičemž zjištěné hodnoty z celé populace minulého semestru jsou $\mu = 6$ a $\sigma = 2$.
- V běžícím semestru je provedeno namátkové šetření u $n = 15$ studentů.
  - Výsledkem je hodnocení, kde se vyskytují dvě 3, tři 4, tři 5, pět 6 a dvě 7.
  - Výběrový průměr je tedy $\overline{X_{15}} \doteq 5,133$ a výběrový rozptyl $S^2_{15} \doteq 1,695$.
- Z předpokladů víme, že $\overline{X_n} \sim N(\mu, \sigma^2/n)$, a tedy $Z = \sqrt{n}\frac{\overline{X_n}-\mu}{\sigma} \sim N(0,1)$.
- Pro vyjádření spolehlivosti našeho odhadu určíme interval, který bude odhadovaný parametr obsahovat s předem zvolenou pravděpodobností $100(1-\alpha)$ %.
  - Hovoříme o hladině spolehlivosti $0 < \alpha < 1$.

Bodové a intervalové odhady

- Nejprve považujme za neznámý nový parametr μ, zatímco o rozptylu budeme předpokládat, že zůstal stejný. Pak

$$1-\alpha = \mathbb{P}(|Z| < z(1-\alpha/2)) = \mathbb{P}(|\sqrt{n}\frac{\overline{X_n}-\mu}{\sigma}| < z(1-\alpha/2))$$
$$= \mathbb{P}\left(\overline{X_n} - \frac{\sigma}{\sqrt{n}}z(1-\alpha/2) < \mu < \overline{X_n} + \frac{\sigma}{\sqrt{n}}z(1-\alpha/2)\right)$$

  a máme interval, jehož hranice jsou náhodné veličiny, a který s předem danou pravděpodobnostní bude obsahovat odhadovaný parametr μ.
  - $z(\beta)$ je β-kvantil standardního normálního rozdělení $N(0,1)$, který najdeme v tabulkách.
- Střed intervalu nazýváme bodovým odhadem pro parametr μ, celý interval intervalovým odhadem.
- Výsledek můžeme interpretovat i tak, že na hladině spolehlivosti α je nebo není odhadovaný parametr μ odlišný od jiné hodnoty μ_0.

Bodové a intervalové odhady

- V případě našich dat vyjde pro $\alpha = 0,05$, že $\mu \in (4,121; 6,145)$.
  - Na hladině spolehlivosti 5 % nemůžeme potvrdit, že se názor studentů na kurz zhoršil, protože uvedený interval obsahuje i hodnotu $\mu_0 = 6$.
- Pro $\alpha = 0,1$ vyjde, že $\mu \in (4,284; 5,983)$.
  - Na úrovni 10 % už takový úsudek uděláme, protože hodnota $\mu_0 = 6$ do intervalu nepadne.

Bodové a intervalové odhady

- Pokud bychom předpokládali, že spokojenost s letošním kurzem bude mít rozptyl odpovědí jiný než loni, museli bychom postupovat odlišně.
- Místo normalizované veličiny Z uvedené výše budeme stejně postupovat s veličinou

$$T = \sqrt{n}\frac{\overline{X_n} - \mu}{S_n}.$$

  - Tato NV má rozdělení $T \sim t_{n-1}$; v našem případě je $n = 15$.
- Vyjde tak intervalový odhad

$$\overline{X_n} - \frac{S_n}{\sqrt{n}} t_{n-1}(1-\alpha/2) < \mu < \overline{X_n} + \frac{S_n}{\sqrt{n}} t_{n-1}(1-\alpha/2).$$

  - Pro $\alpha = 0,05$ máme $\mu \in (4,412; 5,854)$.
  - Pro $\alpha = 0,03$ máme $\mu \in (4,321; 5,945)$.
    - Už na úrovni 3 % spolehlivosti máme za to, že je názor na kurz horší.
  - To odpovídá intuici, že by výrazně menší výběrová směrodatná odchylka $S_n = 1,302$ (než odchylka $\sigma = 2$ z minulého šetření) měla být podstatná pro naše úvahy.

Příklad 124.

Při 600 hodech kostkou padla šestka celkem 45 krát. Je možné tvrdit, že jde o ideální kostku na hladině $\alpha = 0,01$?

Řešení. Pro ideální kostku je pravděpodobnost hodu šestky v každém hodu rovna $p = 1/6$. Počet šestek v 600 hodech je dán NV $\overline{X_n}$, která má binomické rozdělení $\overline{X_n} \sim Bi(600, 1/6)$, a tedy $\mu = 100$ a $var(\overline{X_n}) = 250/3$. Toto rozdělení lze podle centrální limitní věty aproximovat rozdělením $N(\mu, \sigma^2/n) = N(100, 250/3)$. Naměřenou hodnotu $\overline{X_n} = 45$ můžeme považovat za náhodný výběr o jednom členu. Považujeme-li rozptyl za známý, pak je 99% interval spolehlivosti pro střední hodnotu μ roven

$$(45 - \sqrt{250/3}z(0,995), 45 + \sqrt{250/3}z(0,995)) .$$

Z tabulek zjistíme, že kvantil $z(0,995) \approx 2,58$, což dává interval $(21, 69)$. Na ideální kostce je ale $\mu = 100$, a proto nejde v tomto smyslu o ideální kostku na hladině $\alpha = 0,01$.

Příklad 125.

NV X má normální rozdělení $N(\mu, \sigma^2)$, kde μ a σ^2 nejsou známy. V následující tabulce jsou uvedeny četnosti jednotlivých realizací této NV.

X_i	8	11	12	14	15	16	17	18	20	21
n_i	1	2	3	4	7	5	4	3	2	1

Vypočtěte výběrový průměr, výběrový rozptyl, výběrovou směrodatnou odchylku a určete 99% interval spolehlivosti pro střední hodnotu μ.

Řešení.

Výběrový průměr $\overline{X} = \sum n_i X_i / \sum n_i = 490/32 \approx 15,3$.

Výběrový rozptyl $S^2 = \sum n_i (X_i - \overline{X})^2 / (\sum n_i - 1) = 1943/256 \approx 7,6$, a tedy výběrová směrodatná odchylka je $S \approx 2,8$.

$100(1-\alpha)\%$ interval spolehlivosti pro střední hodnotu μ při neznámém rozptylu je $\mu \in (\overline{X} - \frac{S}{\sqrt{n}} t_{n-1}(1-\alpha/2), \overline{X} + \frac{S}{\sqrt{n}} t_{n-1}(1-\alpha/2))$. Dosazením $\overline{X} = 15,3$, $n = 32$, $S \approx 2,8$, $\alpha = 0,01$ a z tabulek $t_{31}(0,995) \approx 2,75$ máme 99% interval spolehlivosti $\mu \in (14,0; 16,7)$.

Horní a dolní odhady

Někdy nás zajímá pouze horní nebo dolní odhad, tedy statistiky U a L pro něž $\mathbb{P}(\mu < U)$ a $\mathbb{P}(L < \mu)$. Jde o tzv. jednostranné intervaly spolehlivosti $(-\infty, U)$ a $(L, +\infty)$.

Pro náhodnou veličinu $Z = \sqrt{n}\frac{\overline{X}-\mu}{\sigma} \sim N(0,1)$ máme

$$1 - \alpha = \phi(z(1-\alpha)) = P(Z < z(1-\alpha)),$$

odkud

$$1 - \alpha = \mathbb{P}(\overline{X} - \frac{\sigma}{\sqrt{n}} z(1-\alpha) < \mu),$$

a tedy $L = \overline{X} - \frac{\sigma}{\sqrt{n}} z(1-\alpha)$. Obdobně $U = \overline{X} + \frac{\sigma}{\sqrt{n}} z(1-\alpha)$ a pro rozdělení s neznámým rozptylem

$$\mu \geq \overline{X} - \frac{S}{\sqrt{n}} t_{n-1}(1-\alpha) \qquad \text{a} \qquad \mu \leq \overline{X} + \frac{S}{\sqrt{n}} t_{n-1}(1-\alpha).$$

Příklad 126.

Předpokládejme, že výška desetiletých chlapců má normální rozdělení $N(\mu, \sigma^2)$ s neznámou střední hodnotou μ a rozptylem $\sigma^2 = 39,112$. Změřením výšky 15 chlapců jsme určili výběrový průměr $\overline{X} = 139,13$. Určete

a) 99% oboustranný interval spolehlivosti pro parametr μ.

b) dolní odhad μ na hladině spolehlivosti 95 %.

Řešení.

a) $100(1-\alpha)\%$ interval spolehlivosti pro neznámou střední hodnotu μ normálního rozložení je $\mu \in (\overline{X} - \frac{\sigma}{\sqrt{n}} z(1-\alpha/2), \overline{X} + \frac{\sigma}{\sqrt{n}} z(1-\alpha/2))$, kde $\overline{X}$ je výběrový průměr z n hodnot, σ je známý rozptyl a $z(1-\alpha/2)$ je příslušný kvantil. Dosazením ze zadání $n = 15$, $\sigma \approx 6,254$ a z tabulek $z(0,995) \approx 2,576$ dostaneme $\frac{\sigma}{\sqrt{n}} z(1-\alpha/2) \approx 4,16$, tedy $\mu \in (134,97; 143,29)$.

b) Dolní odhad L parametru μ na hladině spolehlivosti 95 % je $L = \overline{X} - \frac{\sigma}{\sqrt{n}} z(0,95)$. Z tabulek $z(0,95) \approx 1,645$, a proto $\mu \in (136,474; +\infty)$.

Testování hypotéz

Testování hypotéz

- Mějme dán náhodný vektor $X = (X_1, \ldots, X_n)$ (vzniklý z náhodného výběru) se sdruženou distribuční funkcí $F_X(x)$.
- Hypotéza je tvrzení o rozdělení určeném touto distribuční funkcí.
  - Zpravidla formulujeme dvě hypotézy:
    - nulovou hypotézu H_0 a
    - alternativní hypotézu H_A.
  - Výsledkem testu je rozhodnutí založené na konkrétní realizaci NV X, zda hypotézu H_0 zamítnout ve prospěch hypotézy H_A.

Testování hypotéz

- Vznikají chyby dvou typů:
  1. zamítneme H_0, přestože je platná
  2. nezamítneme H_0, ačkoliv není platná.
- Rozhodování probíhá tak, že vybereme tzv. kritický obor W, tedy množinu výsledků realizace testu, při kterých hypotézu zamítáme
  - Velikost kritického oboru volíme tak, aby platnou hypotézu zamítal s pravděpodobností nejvýše α.
  - Předem požadujeme dané ohraničení pravděpodobnostní chyby prvního typu tzv. hladinu testu α.
  - Často se volí hladiny testů $\alpha = 0,05$ nebo $\alpha = 0,01$.
- Prakticky užitečný je také postup, kdy určíme nejnižší možnou hladinu p testu, při které ještě hypotézu zamítáme a mluvíme o dosažené hladině testu, resp. p-hodnotě testu.

Lineární regrese

- Standardním příkladem užití lineární regrese je proložení přímky danými daty.
- Máme posloupnost měření, ve kterých zaznamenáváme hodnoty dvou veličin u nichž předpokládáme lineární závislost.
  - Klasickým příkladem je závislost výšky syna na výšce otce.

Regresní přímka

- Předpokládáme, že veličiny $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, kde x_i jsou dané konstanty, $i = 1, \ldots, n$.
- Hledáme nejlepší aproximaci $Y_i = b_0 + b_1 x_i$.
- Matice X příslušného lineárního modelu je

$$X^T = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix}.$$

- Odtud

$$\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} n & \overline{x} \\ n\overline{x} & \sum_{i=1}^n x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} n\overline{Y} \\ \sum_{i=1}^n x_i Y_i \end{pmatrix},$$

a proto

$$b_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^n (x_i - \overline{x})^2} \quad \text{a} \quad b_0 = \overline{Y} - b_1 \overline{x}.$$

Regresní přímka

- Lze odvodit

$$Var(b_1) = \sigma^2 / \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

- Pro testování hypotézy, zda střední hodnota veličiny Y nezávisí na x, tj. H_0 je tvaru $\beta_1 = 0$, můžeme použít statistiku

$$T = \frac{b_1}{S} \left(\sum_{i=1}^{n} (x_i - \overline{x})^2 \right)^{1/2} \sim t_{n-2}.$$

Regresní přímka

- Obdobně vypadá statistická analýza vícenásobné regrese, kde máme několik sad hodnot x_{ij} a vyhodnocujeme statistickou relevanci aproximace

$$Y_i = b_0 + b_1 x_{1i} + \cdots + b_k x_{ki}.$$

- Jednotlivé statistiky T_j umožňují t-test závislosti regrese na jednotlivých parametrech.
- Softwarové balíčky zpravidla uvádí také parametr vyjadřující, jak dobře jsou celkově hodnoty Y_i aproximovány.
  - Tento parametr se nazývá koeficient determinace $R^2 = 1 - \frac{RSS}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$.

Příklady

Příklad 127 (lineární regrese).

Určete lineární regresní model pro závislost veličiny Y na veličině X na základě naměřených seznamů dat: $X = [1, 4, 5, 7, 10]$, $Y = [3, 7, 8, 12, 18]$.

Příklad 128 (kvadratická regrese).

Orbitální stanice naměřila v pěti po sobě jdoucích dnech, ve stejnou hodinu následující rychlosti neznámého vesmírného tělesa (v km/s): 10, 11,4, 13,1, 15,8 a 18,7. Odhadněte rychlost tělesa desátého dne.