Obsah

1 O čem to je? 3

1.1 Teorie pravděpodobnosti .......................................................... 4
1.2 Statistika ....................................................................................... 4

2 Základní pojmy teorie pravděpodobnosti 4

2.1 Náhodný pokus ............................................................................ 4
2.2 Laplaceova (klasická) definice pravděpodobnosti ....................... 5
   2.2.1 Základní pojmy ..................................................................... 5
   2.2.2 Náhodná veličina ................................................................. 6
2.3 Vlastnosti pravděpodobnosti ....................................................... 6
   2.3.1 Uplný systém jevů ................................................................. 6
2.4 Problémy Laplaceovy definice pravděpodobnosti ......................... 7
   2.4.1 Rozšíření Laplaceova modelu pravděpodobnosti .................. 7
2.5 Kombinatorické pojmy a vzorce ................................................. 7
2.6 Kolmogorovova definice pravděpodobnosti ................................. 9
   2.6.1 Borelova σ-algebra .............................................................. 9
   2.6.2 Pravděpodobnost (=pravděpodobnostní míra) ...................... 10

3 Nezávislost a podmínněná pravděpodobnost 10

3.1 Nezávislé jevy ............................................................................ 10
3.2 Podmínněná pravděpodobnost .................................................... 12
   3.2.1 Podmínněná nezávislost ..................................................... 15

4 Náhodné veličiny 16

4.1 Náhodná veličina ....................................................................... 16
4.2 Nezávislost náhodných veličin ................................................. 18
4.3 Směs náhodných veličin ............................................................ 19
4.4 Druhy náhodných veličin ........................................................... 20
   4.4.1 Diskrétní náhodné veličiny ................................................ 20
4.4.2 Spojité náhodné veličiny .......................................................... 20
4.4.3 Smíšené náhodné veličiny ...................................................... 21
4.4.4 Směs náhodných veličin stejného typu .................................... 23
4.5 Kvantilová funkce náhodné veličiny ........................................ 24
4.6 Jak reprezentovat náhodnou veličinu v počítači .......................... 25
4.7 Operace s náhodnými veličinami .............................................. 25
4.8 Jak realizovat náhodnou veličinu na počítači .............................. 28

5 Charakteristiky náhodných veličin ............................................. 29
5.1 Střední hodnota ................................................................. 29
5.2 Rozptyl (disperze) .................................................................. 30
5.3 Směrodatná odchylka ............................................................ 31
5.4 Obecné a centrální momenty ................................................. 31
5.5 Normovaná náhodná veličina ................................................. 32
5.6 Základní typy diskrétních rozdělení ....................................... 32
5.6.1 Diracovo ................................................................. 32
5.6.2 Rovnoměrné ............................................................. 32
5.6.3 Alternativní (Bernoulliovo) .............................................. 33
5.6.4 Binomické Bi(m, q) ...................................................... 33
5.6.5 Poissonovo Po(λ) ......................................................... 33
5.6.6 Geometrické .................................................................. 34
5.6.7 Hypergeometrické ........................................................ 34
5.7 Základní typy spojitých rozdělení ......................................... 35
5.7.1 Rovnoměrné R(a, b) ....................................................... 35
5.7.2 Normální (Gaussovo) N(μ, σ^2) ...................................... 35
5.7.3 Logaritmickonormální LN(μ, σ^2) = exp(N(μ, σ^2)) .......... 35
5.7.4 Exponenciální Ex(τ) ....................................................... 36
5.8 Cebýševova nerovnost ......................................................... 36

6 Náhodné vektory ................................................................. 38
6.1 Diskrétní náhodný vektor ...................................................... 40
6.2 Spojitý náhodný vektor .......................................................... 40
6.3 Obecnější náhodné veličiny .................................................. 40
6.4 Číselné charakteristiky náhodného vektoru ............................. 41
6.4.1 Víceozměrně normální rozdělení N(μ, Σ) ......................... 42
6.5 Reprezentace náhodných vektorů v počítači ......................... 43

7 Lineární prostor náhodných veličin ........................................... 43
7.1 Lineární podprostor N náhodných veličin s nulovými středními hodnotami 44
7.2 Lineární regrese .................................................................. 44

8 Základní pojmy statistiky ........................................................... 45
8.1 K čemu potřebujeme statistiku ............................................... 45
8.2 Náhodný výběr, odhad, empirické rozdělení ............................ 45
8.3 Odhad střední hodnoty .......................................................... 46
8.4 Odhad k-tého obecného momentu EX^k .................................. 48
8.5 Odhad rozptylu .................................................................... 49
8.5.1 Odhad rozptylu při známé střední hodnotě .......................... 49
8.5.2 Rozdělení χ^2 s n stupni volnosti, χ^2(n) .......................... 49
Odvad plný... 50
8.5.4 Eficience odhadů rozptylu pro normální rozdělení 54
8.6 Odhad smerodatné odchylky 54
8.7 Histogram a popis empirického rozdělení 55
8.8 Odhad mediánu 55
8.9 Intervalové odhady 56
8.10 Intervalové odhady parametrů normálního rozdělení N(\(\mu, \sigma^2\)) 56
8.10.1 Odhad střední hodnoty při známém rozptylu \(\sigma^2\) 56
8.10.2 Odhad střední hodnoty při neznámém rozptylu 57
8.10.3 Studentovo t-rozdělení 57
8.10.4 Odhad střední hodnoty při neznámém rozptylu 2 58
8.10.5 Odhad rozptylu 59
8.10.6 Intervalové odhady spojitých rozdělení, která nejsou normální 59
8.11 Obecné odhady parametrů 60
8.11.1 Metoda momentů 60
8.11.2 Metoda maximální věrohodnosti 62
8.11.3 Příklady na odhady parametrů 63
9 Testování hypotéz 69
9.1 Základní pojmy a principy testování hypotéz 69
9.2 Testy střední hodnoty normálního rozdělení 73
9.2.1 Při známém rozptylu \(\sigma^2\) 73
9.2.2 Při neznámém rozptylu 73
9.3 Testy rozptylu normálního rozdělení 74
9.4 Porovnání dvou normálních rozdělení 74
9.4.1 Testy rozptylu dvou normálních rozdělení [Fisher] 74
9.4.2 Testy středních hodnot dvou normálních rozdělení se stejným známým rozptylem \(\sigma^2\) 76
9.4.3 Testy středních hodnot dvou normálních rozdělení s různými známnými rozptyly \(\sigma_X^2, \sigma_Y^2\) 76
9.4.4 Testy středních hodnot dvou normálních rozdělení se stejným neznámým rozptylem \(\sigma^2\) 77
9.4.5 Testy středních hodnot dvou normálních rozdělení - párový test 78
9.5 Korelace, její odhad a testování 78
9.5.1 Test nekorelovanosti dvou normálních rozdělení 79
9.6 \(\chi^2\)-test dobré shody 79
9.6.1 Základní podoba testu 79
9.6.2 Modifikace 81
9.6.3 \(\chi^2\)-test nezávislosti dvou rozdělení 81
9.6.4 \(\chi^2\)-test dobré shody dvou rozdělení 82
9.7 Neparameetrické testy 83
9.7.1 Znaměnkový test 83
9.7.2 Wilcoxonův test (jednověký běhový) 83
1 O čem to je?
Motivační příklad (pojistka auta bez marže):
1A. Proti krádeži: je-li cena 1 000 000 Kč a riziko ukradení během pojistného období 0.001

\[ 1 000 000 \cdot 0.001 = 1 000 \text{ Kč} \]

1B. Pro případ havárie:

⇒ Pojem náhodné veličiny a TEORIE PRAVDĚPODOBNOSTI

2. Pravděpodobnost krádeže auta, střední škoda při havárii, přesnost odhadů?

⇒ STATISTIKA

1.1 Teorie pravděpodobnosti
je nástroj pro účelné rozhodování v systémech, kde budoucí pravdivost jevů závisí na okolnostech, které zcela neznáme.

Poskytuje model takových systémů a kvantifikaci výsledků.

Pravděpodobnostní popis ⇒ chování systému

1.2 Statistika
je nástroj pro hledání a ověřování pravděpodobnostního popisu reálných systémů na základě jejich pozorování.

Chování systému ⇒ pravděpodobnostní popis

Statistika poskytuje daleko víc: nástroj pro zkoumání světa, pro hledání a ověřování závislostí, které nejsou zjevné.

2 Základní pojmy teorie pravděpodobnosti

2.1 Náhodný pokus
„Takový, na který si můžeme vsadit.“

Tedy nikoli:

- Jak je pravděpodobné, že ve skriptech na str. 42 je chyba?

- Jak je pravděpodobné, že král ... je živ?

- Jak je pravděpodobné, že zítra bude v menze dobrý oběd?

Vhodná losovací zařízení:

- Kostka, čtyřštěn, dvanáctistěn...
Tužka, dlouhý hranol...

„Kolo štěstí.“

Urna s losy, které nelze před vylosováním rozlišit.

2.2 Laplaceova (klasická) definice pravděpodobnosti

Předpoklad: Náhodný pokus s \( m \in \mathbb{N} \) různými, po dvou neslučitelnými výsledky, které jsou stejně možné.

Jev, který nastává právě při \( k \) z těchto výsledků, má pravděpodobnost \( k/m \).

(Urna s \( m \) losy, z nichž \( k \) „vyhrává“.)

1. problém: Co to je „stejně možné“? „Stejně pravděpodobné“? (definice kruhem!)

Elementární jevy jsou všechny „stejně možné“ výsledky (losy).

Množina všech elementárních jevů: \( \Omega \) (urna)

Jev: \( A \subseteq \Omega \) (množina vyhrávajících losů)

Úmluva. Jevy budeme ztotožňovat s příslušnými množinami elementárních jevů a používat pro ně množinové operace (místo výrokových).

2.2.1 Základní pojmy

Jev jistý: \( \Omega, 1 \) (všechny losy vyhrávají)

Jev nemožný: \( \emptyset, 0 \) (žádný los nevyhrává)

Konjunkce jevů („and“): \( A \cap B \) (losy, které vyhrávají v obou tazích)

Disjunkce jevů („or“): \( A \cup B \) (losy, které vyhrávají v aspoň jednom tahu)

Jev opačný k \( A \): \( \bar{A} = \Omega \setminus A \)

\( A \Rightarrow B \): \( A \subseteq B \)

Jevy neslučitelné: \( A_1, \ldots, A_n : \bigcap_{i \leq n} A_i = \emptyset \)

Jevy po dvou neslučitelné: \( A_1, \ldots, A_n : \forall i, j \in \{1, \ldots, n\}, i \neq j : A_i \cap A_j = \emptyset \)

Jevové pole: všechny jevy pozorovatelné v náhodném pokusu, zde \( \exp \Omega \) (=množina všech podmnožin množiny \( \Omega \))

Pravděpodobnost jevu \( A \):

\[ P(A) = \frac{|A|}{|\Omega|}, \]

kde \(|\.\|\) značí počet prvků množiny
2.2.2 Náhodná veličina

je libovolná funkce \( X : \Omega \to \mathbb{R} \)

\[
\text{Střední hodnota:} \quad E_X = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} X(\omega)
\]

Příklad: Elementární jevy jsou možné výsledky hry, náhodná veličina je výše výhry. Střední hodnota je spravedlivá cena za účast ve hře.

2.3 Vlastnosti pravděpodobnosti

\( P(A) \in (0, 1) \)

\( P(0) = 0, \quad P(1) = 1 \)

\( P(\overline{A}) = 1 - P(A) \)

\( A \subseteq B \Rightarrow P(A) \leq P(B) \)

\( A \subseteq B \Rightarrow P(B \setminus A) = P(B) - P(A) \)

\( A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B) \) \text{(aditivita)}

\( P(A \cup B) = P(A) + P(B) - P(A \cap B) \)

2.3.1 Úplný systém jevů

tvoří jevy \( B_i, i \in I \), jestliže jsou po dvou neslučitelné a \( \bigcup_{i \in I} B_i = 1 \).

Speciální případ pro 2 jevy: \( \{C, \overline{C}\} \)

Je-li \( \{B_1, \ldots, B_n\} \) úplný systém jevů, pak

\[
\sum_{i=1}^{n} P(B_i) = 1
\]

a pro libovolný jev \( A \)

\[
P(A) = \sum_{i=1}^{n} P(A \cap B_i) .
\]

Speciálně:

\[
P(A) = P(A \cap C) + P(A \cap \overline{C}) .
\]

Motivační příklad (kolik je nekuřákov): Mužů je v populaci 48 %, kuřáků a kuřáček dohromady 30 %. Jakých hodnot může nabývat pravděpodobnost, že náhodně vybraný člověk je muž a nekuřák?
$M \ldots \muž, P(M) = 0.48$

$K \ldots kuřák, P(K) = 0.3, P(\overline{K}) = 0.7$

Hledaná pravděpodobnost $P(M \cap \overline{K})$ jevů opačného k $M \cup K$,

$max \{P(M), P(K)\} = 0.52 \leq P(M \cup K) \leq P(M) + P(K) = 0.52 + 0.3 = 0.82$

$1 - 0.82 = 0.18 \leq P(M \cap \overline{K}) \leq 1 - 0.52 = 0.48$

Možné jsou všechny hodnoty z intervalu $\langle 0.18, 0.48 \rangle$.

### 2.4 Problémy Laplaceovy definice pravděpodobnosti

**2. problém:** Nedovoluje nekonečné množiny jevů, geometrickou pravděpodobnost...

Nelze mít nekonečně mnoho stejně pravděpodobných výsledků.

**Příklad:** Podíl plochy pevniny k povrchu Země je pravděpodobnost, že náhodně vybraný bod na Zemi leží na pevnině (je-li výběr bodů prováděn „rovnoměrně“).

**Příklad:** „Kolo štěstí“ s nestejnými oblouky odpovídajícími různým výsledkům.

**Příklad (Buffonova úloha):** Na linkovaný papír hodíme jehlu, jejíž délka je rovna vzdálenosti mezi linkami. Jaká je pravděpodobnost, že jehla protne nějakou linku?

### 3. problém:

Nedovoluje iracionální hodnoty pravděpodobnosti.

### 2.4.1 Rozšíření Laplaceova modelu pravděpodobnosti

**Příklad:** Místo hrací kostky hážíme krabičkou od zápalek, jejíž strany jsou nestejně dlouhé. Jaká je pravděpodobnost možných výsledků?

Připustíme, že **elementární jevy nemusí být stejně pravděpodobné**. Ztrácíme návod, jak vybrat „správnou“ pravděpodobnost.

Je to funkce, která jevům přiřazuje čísla z intervalu $\langle 0, 1 \rangle$ a splňuje jisté podmínky. Nemáme návod, jak z nich vybrat tu pravou.

To je role statistiky, která k danému opakovatelnému pokusu hledá pravděpodobnostní model.

### 2.5 Kombinatorické pojmy a vzorce

*(Dle [Zvára, Štěpán].)*

V urně je $n$ losů, postupně vytáhneme $k$ z nich.

**Permutace (pořadí) bez opakování:** Vytáhneme všech $n$ losů bez vracení, záleží na pořadí. Počet permutací je $n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 1$, každá má pravděpodobnost $\frac{1}{n!}$.

<table>
<thead>
<tr>
<th>výběr</th>
<th>s vracením (opakováním)</th>
<th>bez vracením (opakování)</th>
</tr>
</thead>
<tbody>
<tr>
<td>uspořádaný</td>
<td>$n^k$</td>
<td>$\frac{n!}{(n-k)!}$</td>
</tr>
<tr>
<td>(variaci)</td>
<td>s p-stmi $\frac{1}{n^k}$</td>
<td>s p-stmi $\frac{1}{n^k}$</td>
</tr>
<tr>
<td>neuspořádaný</td>
<td>$(n+k-1)$</td>
<td>$\frac{n!}{k!(n-k)!}$</td>
</tr>
<tr>
<td>(kombinace)</td>
<td>s různými p-stmi</td>
<td>s p-stmi $\frac{1}{n^k}$</td>
</tr>
</tbody>
</table>
Z této tabulky pouze kombinace s opakováním nejsou všechny stejně pravděpodobné (odpovídají různému počtu variací s opakováním) a nedovolují proto použití Laplaceova modelu pravděpodobnosti.

Permutace (pořadí) bez opakování jsou speciální případ variací bez opakování pro \( n = k \).

**Permutace s opakováním**: Tvoříme posloupnost délky \( k \) z \( n \) hodnot, přičemž j-tá hodnota se opakuje \( k_j \)-krát, \( \sum_{j=1}^{n} k_j = k \). Počet různých posloupností je

\[
\frac{k!}{k_1! \cdot \ldots \cdot k_n!}.
\]

Speciálně pro \( n = 2 \) dostáváme

\[
\frac{k!}{k_1! \cdot k_2!} = \frac{k!}{k_1! \cdot (k-k_1)!} = \binom{k}{k_1},
\]

což je počet kombinací bez opakování (ovšem k1-prvkův z k prvků).

<table>
<thead>
<tr>
<th>( n )</th>
<th>4</th>
<th>10</th>
<th>100</th>
<th>1 000</th>
<th>10 000</th>
</tr>
</thead>
<tbody>
<tr>
<td>počet 4-prvkových variací z ( n ) prvků bez opakování, ( \frac{n!}{(n-4)!} )</td>
<td>24</td>
<td>5 040</td>
<td>94 109 400</td>
<td>0.994 \cdot 10^{12}</td>
<td>0.9994 \cdot 10^{16}</td>
</tr>
<tr>
<td>počet 4-prvkových variací z ( n ) prvků s opakováním, ( n^4 )</td>
<td>256</td>
<td>10 000</td>
<td>10^8</td>
<td>10^{12}</td>
<td>10^{16}</td>
</tr>
<tr>
<td>počet 4-prvkových kombinací z ( n ) prvků bez opakování, ( \binom{n}{4} )</td>
<td>1</td>
<td>210</td>
<td>3 921 225</td>
<td>41 417 124 750</td>
<td>4.164 \cdot 10^{14}</td>
</tr>
<tr>
<td>počet 4-prvkových kombinací z ( n ) prvků s opakováním, ( \binom{n+3}{4} )</td>
<td>35</td>
<td>715</td>
<td>4 421 275</td>
<td>41 917 125 250</td>
<td>4.169 \cdot 10^{14}</td>
</tr>
</tbody>
</table>

**Věta.** Pro dané \( k \in \mathbb{N} \) a pro \( n \to \infty \) se poměr počtů variací (resp. kombinací) bez opakování a s opakováním blíží jedné, tj.

\[
\lim_{n \to \infty} \frac{n!}{(n-k)! \cdot n^k} = 1, \quad \lim_{n \to \infty} \frac{\binom{n}{k}}{\binom{n+k-1}{k}} = 1.
\]

**Důkaz.**

\[
\frac{n!}{(n-k)! \cdot n^k} = \frac{n \cdot (n-1) \cdot \ldots \cdot (n-(k-1))}{n^k} = 1 \left(1 - \frac{1}{n}\right) \cdot \ldots \cdot \left(1 - \frac{k-1}{n}\right) \to 1,
\]

\[
\frac{\binom{n}{k}}{\binom{n+k-1}{k}} = \frac{n \cdot (n-1) \cdot \ldots \cdot (n-(k-1))}{(n+(k-1)) \cdot \ldots \cdot (n+1) \cdot n} = \frac{1 \left(1 - \frac{1}{n}\right) \cdot \ldots \cdot \left(1 - \frac{k-1}{n}\right)}{(1 + \frac{k-1}{n}) \cdot \ldots \cdot (1 + \frac{1}{n})} \to 1
\]

(počet činitelů \( k \) je konstantní).
Důsledek. Pro \( n \gg k \) je počet variací (resp. kombinací) s opakováním přibližně

\[
\frac{n!}{(n-k)!} \approx n^k, \quad \binom{n}{k} \approx \frac{n^k}{k!}.
\]

Jednodušší bývá uspořádaný výběr s vracením nebo neuspořádaný výběr bez vracení.

2.6 Kolmogorovova definice pravděpodobnosti
Elementárních jevů = všechny možné výsledky pokusu = prvky množiny \( \Omega \). Může jich být nekonečně mnoho, nemusí být stejně pravděpodobné.

Jevy jsou podmnožiny množiny \( \Omega \), ale ne nutně všechny; tvoří podmnožinu \( \mathcal{A} \subseteq \exp \Omega \), která splňuje následující podmínky:

\[
\begin{align*}
(A1) & \quad \emptyset \in \mathcal{A}. \\
(A2) & \quad A \in \mathcal{A} \Rightarrow \overline{A} \in \mathcal{A}. \\
(A3) & \quad (\forall n \in \mathbb{N} : A_n \in \mathcal{A}) \Rightarrow \bigcup_{n \in \mathbb{N}} A_n \in \mathcal{A}.
\end{align*}
\]

Systém \( \mathcal{A} \) podmnožin nějaké množiny \( \Omega \), který splňuje podmínky (A1-3), se nazývá \( \sigma \)-algebra.

Důsledky: \( \Omega = \emptyset \in \mathcal{A}, \)

\[
\left( \forall n \in \mathbb{N} : A_n \in \mathcal{A} \right) \Rightarrow \bigcap_{n \in \mathbb{N}} A_n = \bigcup_{n \in \mathbb{N}} \overline{A_n} \in \mathcal{A}.
\]

Ministerský úředník: Volme \( \mathcal{A} = \exp \Omega \).
Vede k nežadoucím problémům, např. Banachův-Tarského paradox.

(A3) je uzavřenost na spočetná sjednocení.

Ministerský úředník: Volme uzavřenost na jakákoliv sjednocení.
Ukazuje se jako příliš silný požadavek.

Inženýr: Volme uzavřenost na konečná sjednocení. Nedovoluje např. vyjádřit kruh jako sjednocení obdélníků.
\( \mathcal{A} \) nemusí ani obsahovat všechny jednobodové množiny, v tom případě elementární jevy nemusí být jevy!

2.6.1 Borelova \( \sigma \)-algebra
\( \mathcal{B}(\mathbb{R}) \) je nejmenší \( \sigma \)-algebra podmnožin \( \mathbb{R} \), která obsahuje všechny intervaly.

Obsahuje všechny intervely otevřené, uzavřené i polouzavřené, i jejich spočetná sjednocení, a některé další množiny (např. Cantorovo diskontinuum), ale je menší než \( \exp \mathbb{R} \). Její prvky nazýváme borelovské množiny.
2.6.2 Pravděpodobnost (=pravděpodobnostní míra)

je funkce $P : \mathcal{A} \rightarrow [0, 1)$, splňující podmínky

(P1) $P(1) = 1$,

(P2) $P\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \sum_{n \in \mathbb{N}} P(A_n)$, pokud jsou množiny (=jevy) $A_n$, $n \in \mathbb{N}$, po dvou neslučitelné.

(spočetná aditivita)

Pravděpodobnostní prostor je trojice $(\Omega, \mathcal{A}, P)$, kde $\Omega$ je neprázdná množina, $\mathcal{A}$ je $\sigma$-algebra podmnožin množiny $\Omega$ a $P : \mathcal{A} \rightarrow [0, 1)$ je pravděpodobnost.

Dříve uvedené vlastnosti pravděpodobnosti jsou důsledkem (P1), (P2).

Inženýr: Spokojme se s konečnou aditivitou.

Problémem je např. přechod od obsahu obdélníka k obsahu kruhu.

Příklad („nekonečná ruleta“): Výsledkem může být libovolné přirozené číslo, každé má pravděpodobnost 0.

Ministerský úředník: Požadujeme úplnou aditivitu (pro jakékoli soubory po dvou neslučitelných jevů).

Pak bychom nepřipouštěli ani rovnoměrné rozdělení na intervalu.

Pravděpodobnost zachovává limity monotónních posloupností jevů (množin):

Nechť $(A_n)_{n \in \mathbb{N}}$ je posloupnost jevů.

\[
A_1 \subseteq A_2 \subseteq \ldots \Rightarrow P\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \lim_{n \to \infty} P(A_n),
\]

\[
A_1 \supseteq A_2 \supseteq \ldots \Rightarrow P\left(\bigcap_{n \in \mathbb{N}} A_n\right) = \lim_{n \to \infty} P(A_n).
\]

Laplaceův model  |  Kolmogorovův model
---|---
konečně mnoho jevů | i nekonečně mnoho jevů
p-sti jen racionální | p-sti i iracionální
$P(A) = 0 \Rightarrow A = \emptyset$ | možné jevy s nulovou p-stí
p-sti určeny strukturou jevů | p-sti neurčeny strukturou jevů

Příklad (Buffonova úloha – řešení): Na linkovaný papír hodíme jehlu, jejíž délka je rovna vzdálenosti mezi linkami. Jaká je pravděpodobnost, že jehla protne nějakou linku?

\[
\frac{2}{\pi} \approx 0.63661977236758134307553505349005744.
\]

3 Nezávislost a podmíněná pravděpodobnost

3.1 Nezávislé jevy

Motivace: Dva jevy spolu „nesouvisí“. 
Definice: \[ P(A \cap B) = P(A) \cdot P(B). \]

To je ovšem jen náhražka, která říká mnohem méně, než jsme chtěli!
(Podobně \( P(A \cap B) = 0 \) neznamená, že jevy \( A, B \) jsou neslučitelné.)

Pro nezávislé jevy \( A, B \)
\[ P(A \cup B) = P(A) + P(B) - P(A) \cdot P(B). \]

Důkaz:
\[ P(A \cup B) = P(A) + P(B) - P(A) \cap P(B) = P(A) + P(B) - P(A) \cdot P(B). \]

Jsou-li jevy \( A, B \) nezávislé, pak jsou nezávislé také jevy \( A, \bar{B} \) (a též dvojice jevů \( \bar{A}, B \) a \( \bar{A}, \bar{B} \)).

Důkaz:
\[ P(A \cap \bar{B}) = P(A) - P(A \cap B) = P(A) - P(A) \cdot P(B) = \]
\[ = P(A) \cdot (1 - P(B)) = P(A) \cdot P(\bar{B}). \]

Jevy \( A_1, \ldots, A_n \) se nazývají po dvou nezávislé, jestliže každé dva z nich jsou nezávislé.

To je málo:

Příklad. Máme dva hody mincí a jevy
\( A \) ... při prvním hodu padne líč,
\( B \) ... při druhém hodu padne líč,
\( C \) ... při pravém jednom hodu padne líč.

\[ P(A) = P(B) = P(C) = \frac{1}{2}, \quad P(A \cap B) = P(A \cap C) = P(B \cap C) = \frac{1}{4} = P(A) \cdot P(B) = P(A) \cdot P(C) = P(B) \cdot P(C), \]
\[ P(A \cap B \cap C) = 0 \neq \frac{1}{8} = P(A) \cdot P(B) \cdot P(C), \]

takže jevy \( A, B, C \) jsou po dvou nezávislé, ale nejsou nezávislé.

Množina jevů \( \mathcal{M} \) se nazývá nezávislá, jestliže
\[ P\left( \bigcap_{A \in \mathcal{K}} A \right) = \prod_{A \in \mathcal{K}} P(A) \]
pro všechny konečné podmnožiny \( \mathcal{K} \subseteq \mathcal{M} \).

Příklad. Spínače zapojené dle obrázku jsou nezávisle sepnuty s pravděpodobností 0.9. S jakou pravděpodobností celá soustava povede proud?
Dvě sériově spojené části, každá vede s pravděpodobností

\[ 0.9 + 0.9 - 0.9^2 = 0.99 , \]

celek vede s pravděpodobností

\[ 0.99^2 = 0.9801 . \]

### 3.2 Podmíněná pravděpodobnost

**Motivační příklad (alkohol za volantem):**

90 % všech nehod způsobili střídaví řidiči.

**Alkoholik:** Když se napiju, budu mít 9× menší riziko havárie.

**Statistik:** To by byla pravda, kdyby opilých bylo stejně jako střídavých.

Ve skutečnosti 99 % řidičů bylo střídavých. Kolikrát se pozitivu alkoholu zvyšuje riziko nehody?

**Příklad:** Pravděpodobnosti výsledků tenisového zápasu se podstatně změní po odehrání prvního setu.

Máme pravděpodobnostní popis systému. Dostaneme-li dodatečnou informaci, že nastal jev \( B \), můžeme aktualizovat naší znalost o pravděpodobnosti libovolného jevu \( A \). Ten lze vyjádřit jako disjunktní sjednocení \( (A \cap B) \cup (A \cap \bar{B}) \), takže

\[ P(A) = P(A \cap B) + P(A \cap \bar{B}) . \]

Je-li \( P(B) \neq 0 \neq P(\bar{B}) \), můžeme roznásobit:

\[ P(A) = P(B) \frac{P(A \cap B)}{P(B)} + P(\bar{B}) \frac{P(A \cap \bar{B})}{P(\bar{B})} . \]

Funkce \( P(.|B), P(.|\bar{B}) : \mathcal{A} \to (0,1) \),

\[ P(A|B) = \frac{P(A \cap B)}{P(B)} , \quad P(A|\bar{B}) = \frac{P(A \cap \bar{B})}{P(\bar{B})} , \]

jsou pravděpodobnosti na \( \mathcal{A} \), neboť splňují

(P1) \( P(1|B) = \frac{P(1 \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1 \)

a pro \( A_n, n \in \mathbb{N} \), po dvou neslučitelné

(P2) \( P\left( \bigcup_{n \in \mathbb{N}} A_n \big| B \right) = \frac{P\left( \left( \bigcup_{n \in \mathbb{N}} A_n \right) \cap B \right)}{P(B)} = \frac{P\left( \bigcup_{n \in \mathbb{N}} (A_n \cap B) \right)}{P(B)} = \sum_{n \in \mathbb{N}} P(A_n \cap B) = \sum_{n \in \mathbb{N}} P(A_n|B) . \)
Obdobně pro $P(.|\overline{B})$. Nazývají se **podmíněné pravděpodobnosti**.


Podmíněné pravděpodobnosti navíc splňují $B \subseteq A \Rightarrow P(A|B) = 1, \quad P(A \cap B) = 0 \Rightarrow P(A|B) = 0$, speciálně $P(B|B) = 1, \quad P(\overline{B}|B) = 0$.

Původní pravděpodobnost $P(.)$ jsme vyjádřili jako konvexní kombinaci pravděpodobností $P(.|B)$, $P(.|\overline{B})$, odpovídajících situacím, kdy jev $B$ nastal, resp. nenastal:

$$P(A) = P(B) \, P(A|B) + P(\overline{B}) \, P(A|\overline{B}).$$

Tato podmínka spolu s $P(B|B) = 1 = P(\overline{B}|\overline{B})$ určuje pravděpodobnosti $P(.|B)$, $P(.|\overline{B})$ jednoznačně. (Pokud není jedna z pravděpodobností $P(B)$, $P(\overline{B})$ nulová.)

Obecněji:

**Věta o úplné pravděpodobnosti:** Nechť $B_i, \, i \in I$, je (spočetný) úplný systém jevů a $\forall i \in I: P(B_i) \neq 0$. Pak pro každý jev $A$ platí

$$P(A) = \sum_{i \in I} P(B_i) \, P(A|B_i).$$

**Důkaz:**

$$P(A) = P\left(\left(\bigcup_{j \in I} B_j\right) \cap A\right) = P\left(\bigcup_{j \in I} (B_j \cap A)\right) = \sum_{i \in I} P(B_i \cap A) = \sum_{i \in I} P(B_i) \, P(A|B_i).$$

**Bayesova věta:** Nechť $B_i, \, i \in I$, je spočetný úplný systém jevů a $\forall i \in I: P(B_i) \neq 0$. Pak pro každý jev $A$ splňující $P(A) \neq 0$ platí

$$P(B_i|A) = \frac{P(B_i) \, P(A|B_i)}{\sum_{j \in I} P(B_j) \, P(A|B_j)}.$$

**Důkaz** (s využitím věty o úplné pravděpodobnosti):

$$P(B_i|A) = \frac{P(B_i \cap A)}{P(A)} = \frac{P(B_i) \, P(A|B_i)}{\sum_{j \in I} P(B_j) \, P(A|B_j)}.$$

**Motivační příklad (test nemoci):** Test nemoci je u 1 % zdravých falešně pozitivní a u 10 % nemocných falešně negativní. Nemocných je v populaci 0.001. Jaká je pravděpodobnost, že pacient s pozitivním testem je nemocný?

$N \ldots$ nemocný, $P \ldots$ pozitivní test
P(N) = 0.001, \ P(\overline{N}) = 0.999 \\
P(P|N) = 0.01, \ P(P|\overline{N}) = 1 - P(\overline{P}|N) = 0.9 \\

P(P) = P(P|N) \cdot P(N) + P(P|\overline{N}) \cdot P(\overline{N}) \\
      = 0.9 \cdot 0.001 + 0.1 \cdot 0.999 = 0.0009 + 0.00999 = 0.01089 \\

P(N|P) = \frac{P(P|N) \cdot P(N)}{P(P|N) \cdot P(N) + P(P|\overline{N}) \cdot P(\overline{N})} \\
      = \frac{P(P|N) \cdot P(N)}{P(P)} = \frac{0.0009}{0.01089} = 0.08264 \\

Význam: Pravděpodobnost \( P(A|B_i) \) odhadneme z pokusů nebo z modelu, pomocí nich určíme pravděpodobnost \( P(B_i|A) \), které slouží k „optimálnímu“ odhadu, který z jevů \( B_i \) nastal.

Problém: Ke stanovení aposteriorní pravděpodobnosti \( P(B_i|A) \) potřebujeme znát i apriorní pravděpodobnost \( P(B_i) \).

Příklad: Informační kanál

\( B_j \) ... vyslán \( j \)-tý vstupní znak, \( j \in \{1, \ldots, m\} \)
\( A_i \) ... přijat \( i \)-tý výstupní znak, \( i \in \{1, \ldots, k\} \) (může být \( k \neq m \))
Lze odhadnout podmíněné pravděpodobnosti \( P(A_i|B_j) \), že znak \( j \) bude přijat jako \( i \).
Z apriorních pravděpodobností (vyslání znaku \( j \)) \( P(B_j) \) můžeme maticovým násobením určit pravděpodobnosti přijatých znaků:

\[
\begin{bmatrix}
    P(A_1) & P(A_2) & \cdots & P(A_k)
\end{bmatrix} =
\begin{bmatrix}
    P(B_1) & P(B_2) & \cdots & P(B_1) \\
    P(B_1) & P(B_2) & \cdots & P(B_2) \\
    \vdots & \vdots & \ddots & \vdots \\
    P(B_1) & P(B_2) & \cdots & P(B_1)
\end{bmatrix} \begin{bmatrix}
    P(A_1|B_1) & P(A_2|B_1) & \cdots & P(A_k|B_1) \\
    P(A_1|B_2) & P(A_2|B_2) & \cdots & P(A_k|B_2) \\
    \vdots & \vdots & \ddots & \vdots \\
    P(A_1|B_m) & P(A_2|B_m) & \cdots & P(A_k|B_m)
\end{bmatrix}
\]

Všechny matice v tomto vzorci mají jednotkové součety řádků (takové matice nazýváme stochastické). Pokud byl přijat znak \( i \), je podmíněné rozdělení pravděpodobnosti vstupních znaků

\[
P(B_j|A_i) = \frac{P(A_i|B_j) \cdot P(B_j)}{P(A_i)}.
\]

Rozdělení pravděpodobností vyslaných znaků je

\[
\begin{bmatrix}
    P(B_1) & P(B_2) & \cdots & P(B_m)
\end{bmatrix} =
\begin{bmatrix}
    P(A_1) & P(A_2) & \cdots & P(A_k)
\end{bmatrix} \begin{bmatrix}
    P(A_1|B_1) & P(A_2|B_1) & \cdots & P(A_k|B_1) \\
    P(A_1|B_2) & P(A_2|B_2) & \cdots & P(A_k|B_2) \\
    \vdots & \vdots & \ddots & \vdots \\
    P(A_1|B_m) & P(A_2|B_m) & \cdots & P(A_k|B_m)
\end{bmatrix}^{-1},
\]

pokud \( k = m \) a příslušná inverzní matice existuje.

Motivační příklad (alkohol za volantem – řešení):
90\% všech nehod způsobili střízliví řidiči.
99% řidičů bylo střízlivých.
Označme jevy
A ... požil alkohol, \( P(A) = 0.01 \),
H ... způsobil nehodu, \( P(A|H) = 0.1 \).

\[
0.1 = P(A|H) = \frac{P(H|A) \cdot P(A)}{P(H|A) \cdot P(A) + P(H|\bar{A}) \cdot P(\bar{A})} = \frac{P(H|A) \cdot 0.01}{P(H|A) \cdot 0.01 + P(H|\bar{A}) \cdot 0.99} = \frac{1}{1 + \frac{P(H|\bar{A})}{P(H|A)} \cdot 0.99}.
\]

Požitím alkoholu se zvýšuje riziko nehody
\[
\frac{P(H|A)}{P(H|\bar{A})} = 11 \times .
\]

Kdyby bylo 50% řidičů opilých, \( P(A) = 0.5 \), jejich podíl na haváriích by byl

\[
P(A|H) = \frac{P(H|A) \cdot P(A)}{P(H|A) \cdot P(A) + P(H|\bar{A}) \cdot P(\bar{A})} = \frac{P(H|A) \cdot 0.5}{P(H|A) \cdot 0.5 + P(H|\bar{A}) \cdot 0.5} = \frac{1}{1 + \frac{P(H|\bar{A})}{P(H|A)} \cdot 0.5} = \frac{1}{1 + \frac{1}{11}} = \frac{11}{12}.
\]

(Neuvažovali jsme, že účastníků nehody bývá víc a přítomnost alkoholu u jejích účastníků nemusí být nezávislá.)

3.2.1 Podmíněná nezávislost

Příklad. A. Kolik reálných parametrů (stupňů volnosti) je třeba pro určení pravděpodobností všech jevů, které lze popsat logickými výrazy z několika jevů?
B. Jak se toto číslo změní, předpokládáme-li, že výchozí jevy jsou nezávislé?

Řešení. A. Každý z uvažovaných jevů lze vyjádřit v úplné disjunktivní normální formě jako disjunkci výrazů, které jsou konjunkcemi n výchozích jevů nebo jejich negací; těch je \( 2^n \) a tvoří úplný systém jevů, popsáných \( 2^n - 1 \) parametry (součet jejich pravděpodobností 1).
B. Stačí n pravděpodobností výchozích jevů.

<table>
<thead>
<tr>
<th>( n )</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>10</th>
<th>16</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 2^n - 1 )</td>
<td>3</td>
<td>7</td>
<td>15</td>
<td>31</td>
<td>1023</td>
<td>65535</td>
<td>1073741823</td>
</tr>
</tbody>
</table>

Je vidět, že předpoklad nezávislosti může vést na mnohem jednodušší model. Často je však neopodstatněný.

Motivační příklad (popis systému se závislými jevy) : Chceme popsat pravděpodobnosti jevů, složených z následujících 4 charakteristik lidí:
- rád nosí růžové oblečení,
- používá make-up,
• pracuje jako zdravotní sestra,
• prodělal rakovinu prsu.

Byla zjištěna velmi podstatná závislost; potřebujeme $2^4 - 1 = 15$ parametrů.

**Definice.** Náhodné jevy $A, B$ jsou podmíněně nezávislé za podmínky $C$, jestliže

$$P(A \cap B|C) = P(A|C) \cdot P(B|C).$$

Obdobně definujeme podmíněnou nezávislost více jevů.

**Motivační příklad (popis systému se závislými jevy) – řešení:** Přidáme ještě pátý parametr:

• je žena,

a shledáme, že při jeho znalosti lze ostatní 4 parametry považovat za podmíněně nezávislé. Stačí nám 4 parametry pro ženy, 4 pro muže a 1 pro pravděpodobnost, že osoba je žena, tj. 9 parametrů místo 15, resp. 31.

### 4 Náhodné veličiny

**Příklad:** Auto v ceně 10 000 $ bude do roka ukradeno s pravděpodobností $1 : 1000$. Adekvální cena ročního pojišťovného (bez zisku pojišťovny) je $10 000 / 1 000 = 10$ $.$

Někdy tento jednoduchý postup selhává:

**Příklad:** Pro stanovení havarijního pojištění potřebujeme znát nejen pravděpodobnost havárie (resp. počtu havárií za pojišťovné období), ale i „průměrnou“ škodu při jedné havárii, lépe pravděpodobnostní rozdělení výše škody.

⇒ Musíme studovat i náhodné pokusy, jejichž výsledky nejsou jen dva (jev nastal/nenastal), ale více hodnot, vyjádřených reálnými čísly.

#### 4.1 Náhodná veličina

Na pravděpodobnostním prostoru $(\Omega, A, P)$ je měřitelná funkce $X : \Omega \to \mathbb{R}$, tj. taková, že pro každý interval $I$ platí

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

Je popsaná pravděpodobnostmi

$$P_X(I) = P(X \in I) = P(\{ \omega \in \Omega \mid X(\omega) \in I \}),$$

definovanými pro libovolný interval $I$ (a tedy i pro libovolné sjednocení spočetně mnoha intervalů a pro libovolnou borelovskou množinu). $P_X$ je pravděpodobnostní míra na Borelově $\sigma$-algebře určující rozdělení náhodné veličiny $X$. 
K tomu, aby stačila znalost $P_X$ na intervalech, se potřebujeme omezit na tzv. **perfektní míry**; s jinými se v praxi nesetkáme.

Pravděpodobnostní míra $P_X$ splňuje podmínky:

$$P_X(\mathbb{R}) = 1,$$

$$P_X(\bigcup_{n \in \mathbb{N}} I_n) = \sum_{n \in \mathbb{N}} P_X(I_n),$$

pokud jsou množiny $I_n, n \in \mathbb{N}$, po dvou disjunktní,

$$P_X(\emptyset) = 0, \quad P_X(\mathbb{R} \setminus I) = 1 - P_X(I),$$

$I \subseteq J \Rightarrow P_X(I) \leq P_X(J), P_X(J \setminus I) = P_X(J) - P_X(I)$.

Popisy náhodné veličiny

<table>
<thead>
<tr>
<th>prostor elementárních jevů</th>
<th>$\Omega$</th>
<th>$\mathbb{R}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma$-algebra jevů</td>
<td>$\mathcal{A}$</td>
<td>$\mathcal{B}((\mathbb{R}))$</td>
</tr>
<tr>
<td>pravděpodobnostní míra</td>
<td>$P$</td>
<td>$P_X((\mathbb{R}, \mathcal{B}((\mathbb{R})), P_X)$</td>
</tr>
<tr>
<td>pravděpodobnostní prostor</td>
<td>$\mathcal{B}(\mathbb{R})$</td>
<td>$P_X(\Omega)$</td>
</tr>
<tr>
<td>náhodná proměnná</td>
<td>$X: \Omega \rightarrow \mathbb{R}, \ \omega \mapsto X(\omega)$</td>
<td>$X: x \mapsto x$</td>
</tr>
<tr>
<td>$P(X \in I)$</td>
<td>$P{\omega \in \Omega \mid X(\omega) \in I}$</td>
<td></td>
</tr>
</tbody>
</table>

**Příklad:** Počet figurek Člověče nezlob se!, které vstupují do hry po jednom hodu kostkou.

<table>
<thead>
<tr>
<th>prostor elementárních jevů</th>
<th>$\Omega = {1, 2, 3, 4, 5, 6}$</th>
<th>$\mathbb{R}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma$-algebra jevů</td>
<td>$\mathcal{A} = \exp \Omega$</td>
<td>$\mathcal{B}(\mathbb{R})$</td>
</tr>
<tr>
<td>pravděpodobnostní míra</td>
<td>$P(A) = \frac{</td>
<td>A</td>
</tr>
<tr>
<td>náhodná proměnná</td>
<td>$X(\omega) = \begin{cases} 1, &amp; \omega = 6 \ 0, &amp; \text{jinak} \end{cases}$</td>
<td>$X : x \mapsto x$</td>
</tr>
</tbody>
</table>

**Úspornější reprezentace:** omezíme se na intervaly tvaru $I = (-\infty, t), t \in \mathbb{R}$,

$$P(X \in (-\infty, t)) = P(X \leq t) = P_X((-\infty, t)) = F_X(t).$$

$F_X : \mathbb{R} \rightarrow (0, 1)$ je **distribuční funkce** náhodné veličiny $X$. Ta stačí, neboť

$$(a, b) = (-\infty, b) \setminus (-\infty, a), \quad P_X((a, b)) = P(a < X \leq b) = F_X(b) - F_X(a),$$

$$(a, \infty) = \mathbb{R} \setminus (-\infty, a), \quad P_X((b, \infty)) = 1 - F_X(b),$$

$$\{b\} = \bigcap_{n \rightarrow \infty} (b - \frac{1}{n}, b) \quad P_X(\{a\}) = P(X = a) = \lim_{n \rightarrow \infty} (F_X(b) - F_X(b - \frac{1}{n})) = F_X(b) - \lim_{a \rightarrow b-} F_X(a),$$

$$\ldots$$

Vlastnosti distribuční funkce:

- neklesající,
- zprava spojitá,
- $\lim_{t \rightarrow -\infty} F_X(t) = 0, \quad \lim_{t \rightarrow \infty} F_X(t) = 1$.

**Věta:** Tyto podmínky jsou nejen **nutné**, ale i **postačující**.
Příklad: Reálnému číslu \( r \) odpovídá náhodná veličina (značená též \( r \)) s Diracovým rozdělením v \( r \):

\[
P_r(I) = \begin{cases} 
0 & \text{pro } r \notin I, \\
1 & \text{pro } r \in I,
\end{cases}
\quad F_r(t) = \begin{cases} 
0 & \text{pro } t < r, \\
1 & \text{pro } t \geq r.
\end{cases}
\]

\( F_r \) je posunutá Heavisideova funkce.

Tvrzení: \( X \leq Y \Rightarrow F_X \geq F_Y \).

4.2 Nezávislost náhodných veličin

Náhodné veličiny \( X_1, X_2 \) jsou nezávislé, pokud pro všechny intervaly \( I_1, I_2 \) jsou jevy \( X_1 \in I_1, X_2 \in I_2 \) nezávislé, tj.

\[
P(X_1 \in I_1, X_2 \in I_2) = P(X_1 \in I_1) \cdot P(X_2 \in I_2).
\]

Stačí se omezit na intervaly tvaru \((-\infty, t)\), tj.

\[
P(X_1 \leq t_1, X_2 \leq t_2) = P(X_1 \leq t_1) \cdot P(X_2 \leq t_2) = F_{X_1}(t_1) \cdot F_{X_2}(t_2)
\]

pro všechna \( t_1, t_2 \in \mathbb{R} \).

Náhodné veličiny \( X_1, \ldots, X_n \) jsou nezávislé, pokud pro libovolné intervaly \( I_1, \ldots, I_n \) platí

\[
P(X_1 \in I_1, \ldots, X_n \in I_n) = P(X_1 \in I_1) \cdot \ldots \cdot P(X_n \in I_n) = \prod_{i=1}^{n} P(X_i \in I_i).
\]

Ekvivalentně stačí požadovat

\[
P(X_1 \leq t_1, \ldots, X_n \leq t_n) = \prod_{i=1}^{n} P(X_i \leq t_i) = \prod_{i=1}^{n} F_{X_i}(t_i)
\]

pro všechna \( t_1, \ldots, t_n \in \mathbb{R} \).

Na rozdíl od definice nezávislosti více než 2 jevů, zde není třeba požadovat nezávislost pro libovolnou podmnožinu náhodných veličin \( X_1, \ldots, X_n \). Ta vyplývá z toho, že libovolnou náhodnou veličinu \( X_i \) lze „vynechat“ tak, že zvolíme příslušný interval \( I_i = \mathbb{R} \), resp. \( t_i = \infty \). Pak \( P(X_i \in I_i) = 1 \) a v součinu se tento činitel neprojeví.

Spočetná nekonečná množina náhodných veličin je nezávislá, je-li každá její konečná podmnožina nezávislá.

Náhodné veličiny \( X_1, \ldots, X_n \) jsou po dvou nezávislé, pokud každé dvě (různé) z nich jsou nezávislé. To je slabší podmínka než nezávislost veličin \( X_1, \ldots, X_n \).
4.3 Směs náhodných veličin

Příklad: Náhodné veličiny \( V, U \) jsou výsledky studenta při odpovědích na dvě zkouškové otázky. Učitel náhodně vybere s pravděpodobností \( c \) první otázku, s pravděpodobností \( 1 - c \) druhou; podle odpovědi na vybranou otázku udělí známku. Jaké rozdělení má výsledná známka \( X \)?

Matematický model vyžaduje vytvoření odpovídajícího pravděpodobnostního prostoru pro tento pokus.

Nechť \( V \), resp. \( U \) je náhodná veličina na pravděpodobnostním prostoru \((\Omega_1, A_1, P_1)\), resp. \((\Omega_2, A_2, P_2)\), přičemž \( \Omega_1 \cap \Omega_2 = \emptyset \).

Nechť \( c \in (0, 1) \).

Definujeme nový pravděpodobnostní prostor \((\Omega, A, P)\), kde
\[
\Omega = \Omega_1 \cup \Omega_2, \quad A = \{A_1 \cup A_2 \mid A_1 \in A_1, \ A_2 \in A_2\},
\]
\[
P(A_1 \cup A_2) = c \, P_1(A_1) + (1 - c) \, P_2(A_2) \text{ pro } A_1 \in A_1, \ A_2 \in A_2.
\]

Definujeme funkci \( X : \Omega \to \mathbb{R} \):
\[
X(\omega) = \begin{cases} 
V(\omega) & \text{pro } \omega \in \Omega_1, \\
U(\omega) & \text{pro } \omega \in \Omega_2.
\end{cases}
\]

\( X \) je náhodná veličina na \((\Omega, A, P)\).

\( X \) nazýváme směs náhodných veličin \( V, U \) s koeficientem \( c \) (angl. mixture), značíme \( \text{Mix}_c(V, U) \). Má pravděpodobnostní míru
\[
P_X = c \, P_V + (1 - c) \, P_U
\]

a distribuční funkci
\[
F_X = c \, F_V + (1 - c) \, F_U.
\]

Podobně definujeme obecnější směs náhodných veličin \( V_1, \ldots, V_n \) s koeficienty \( c_1, \ldots, c_n \in (0, 1) \), \( \sum_{i=1}^{n} c_i = 1 \), značíme \( \text{Mix}_{(c_1,\ldots,c_n)}(V_1, \ldots, V_n) = \text{Mix}_c(V_1, \ldots, V_n) \), kde \( c = (c_1, \ldots, c_n) \).

Má pravděpodobnostní míru \( \sum_{i=1}^{n} c_i \, P_{V_i} \) a distribuční funkci \( \sum_{i=1}^{n} c_i \, F_{V_i} \). (Lze zobecnit i na spočetně mnoho náhodných veličin.)

Podíl jednotlivých složek je určen vektorem koeficientů \( c = (c_1, \ldots, c_n) \). Jejich počet je stejný jako počet náhodných veličin ve směsi. Jelikož \( c_n = 1 - \sum_{i=1}^{n-1} c_i \), poslední koeficient někdy vynecháváme.

Speciálně pro dvě náhodné veličiny \( \text{Mix}_{(c,1-c)}(V, U) = \text{Mix}_c(V, U) \) (kde \( c \) je číslo, nikoli vektor).

Příklad: Směsí reálných čísel \( r_1, \ldots, r_n \) s koeficienty \( c_1, \ldots, c_n \) je náhodná veličina \( X = \text{Mix}_{(c_1,\ldots,c_n)}(r_1, \ldots, r_n) \),
\[
P_X(I) = P(X \in I) = \sum_{i:r_i \in I} c_i, \quad F_X(t) = \sum_{i:r_i \leq t} c_i.
\]
Lze ji popsat též pravděpodobnostní funkcí \( p_X : \mathbb{R} \to (0, 1) \),
\[
p_X(t) = P_X(\{t\}) = P(X = t) = \begin{cases} 
c_i & \text{pro } t = r_i, \\
0 & \text{jinak}
\end{cases}
\]
(pokud jsou \( r_1, \ldots, r_n \) navzájem různá). Možno zobecnit i na spočetně mnoho reálných čísel.
4.4 Druhy náhodných veličin

4.4.1 Diskrétní náhodné veličiny
(z předchozího příkladu)

Existuje spočetná množina $O_X$, pro kterou $P_X(\mathbb{R} \setminus O_X) = P(X \notin O_X) = 0$. Nejmenší taková množina (pokud existuje) je $\Omega_X = \{t \in \mathbb{R} : P_X(\{t\}) \neq 0\} = \{t \in \mathbb{R} : P(X = t) \neq 0\}$.

Diskrétní náhodnou veličinu popisuje pravděpodobnostní funkce $p_X(t) = P_X(\{t\}) = P(X = t)$. Splňuje

$$\sum_{t \in \mathbb{R}} p_X(t) = 1.$$ 

4.4.2 Spojité náhodné veličiny

Mají spojitou distribuční funkci.

Náhodná veličina $X$ je absolutně spojitá, jestliže existuje nezáporná funkce $f_X : \mathbb{R} \rightarrow (0, \infty)$ (hustota náhodné veličiny $X$) taková, že

$$F_X(t) = \int_{-\infty}^{t} f_X(u) \, du.$$ 

Hustota splňuje $\int_{-\infty}^{\infty} f_X(u) \, du = 1$.

Není určena jednoznačně, ale dvě hustoty $f_X, g_X$ téže náhodné veličiny splňují

$$\int_{I} (f_X(x) - g_X(x)) \, dx = 0$$ pro všechny intervyly $I$.

Lze volit $f_X(t) = \frac{dF_X(t)}{dt}$, pokud derivace existuje.
\( P_X(\{t\}) = 0 \) pro všechna \( t \).
Některé spojité náhodné veličiny nejsou absolutně spojité; mají spojitou distribuční funkci, kterou nelze vyjádřit jako integrál. Tyto případy dále neuvážujeme.

### 4.4.3 Smíšené náhodné veličiny

**Motivační příklad (dešťové srážky):**
Srážkový úhrn v mm za 24 hodin má rozdělení s distribuční funkcí

\[
F_X(t) = \begin{cases} 
1 - \frac{1}{3} \exp \left( \frac{t}{100} \right), & t \geq 0, \\
0 & \text{jinak.}
\end{cases}
\]

(Po \( 2/3 \) dní neprší.)

Směs předchozích dvou případů;
\( \Omega_X \neq \emptyset, P_X(\mathbb{R} \setminus \Omega_X) = P(X \notin \Omega_X) \neq 0. \)

Nelze je popsát ani pravděpodobnostní funkcí (existuje, ale neurčuje celé rozdělení) ani hustotou (neexistuje, nevychází konečná).
Každou náhodnou veličinu se smíšeným rozdělením lze jednoznačně vyjádřit ve tvaru $X = \text{Mix}_c(V, U)$, kde $V$ je spojitá, $U$ je diskrétní a $c \in (0, 1)$: Nespojitostí je spočetné mnoho, lze je očíslovat; $n$-tá je v bodě $r_n$ a má velikost

$$c_n := F_X(r_n) - \lim_{t \to r_n-} F_X(t).$$

Odpovídá jí složka směsi $r_n$ s Diracovým rozdělením a váhou $c_n$.

$$F_X(t) = \sum_n c_n F_{r_n}(t) + G(t),$$

$$G := F_X - \sum_n c_n F_{r_n}$$

je spojitá neklesající funkce,

$$\lim_{t \to -\infty} G(t) = 0,$$

$$\lim_{t \to \infty} G(t) = 1 - \sum_n c_n =: c,$$

$$F_V := \frac{G}{c}$$

je distribuční funkce spojité náhodné veličiny $V$,

$$X = \text{Mix}_{(c,c_1,c_2,...)}(V, r_1, r_2, \ldots),$$

$$U = \frac{1}{\sum_n c_n} \text{Mix}_{(c_1,c_2,...)}(r_1, r_2, \ldots),$$

je diskrétní složka náhodné veličiny $X$,

$$X = \text{Mix}_{(c,1-c)}(V, U).$$
Motivační příklad (dešťové srážky – pokračování):
Srážkový úhrn v mm za 24 hodin má rozdělení s distribuční funkcí
\[ F_X(t) = \begin{cases} 1 - \frac{1}{3} \exp\left(-\frac{t}{100}\right), & t \geq 0, \\ 0 & \text{jinak.} \end{cases} \]

Po 2/3 dní neprší, diskrétní složka je \( U = 0 \) s váhou \( c_1 := 2/3 \).
Spojitá složka \( V \) má váhu \( c := 1/3 \), distribuční funkci
\[ F_V(t) = \begin{cases} 1 - \exp\left(-\frac{t}{100}\right), & t \geq 0, \\ 0 & \text{jinak} \end{cases} \]
a hustotu
\[ f_V(t) = \begin{cases} \frac{1}{100} \exp\left(-\frac{t}{100}\right), & t \geq 0, \\ 0 & \text{jinak} \end{cases} \]
(exponenciální rozdělení).

4.4.4 Směs náhodných veličin stejného typu
\[ X = \text{Mix}_{(c,1-c)}(V,U). \]
Jsou-li \( V,U \) diskrétní, má \( X \) pravděpodobnostní funkci
\[ p_X = c p_V + (1 - c) p_U. \]
Jsou-li \( V,U \) absolutně spojité, má \( X \) hustotu
\[ f_X = c f_V + (1 - c) f_U. \]
Obdobně pro směs více náhodných veličin.
4.5 Kvantilová funkce náhodné veličiny

Příklad. Pokud absolvent školy říká, že patří mezi 5 % nejlepších, pak tvrdí, že distribuční funkce prospěchu (náhodně vybraného absolventa) má u jeho prospěchu hodnotu nejvýše 0.05. (Předpokládejme, že lepšímu prospěchu odpovídá nižší průměr známek.)

Neosoučasně, školní znamená, že hodnota distribuční funkce udává podíl těch absolventů, kteří měli lepší nebo stejný prospěch.

Obráceně se lze ptát, jaký prospěch je potřeba k tomu, aby se absolvent dostal mezi 5 % nejlepších.

Pro $\alpha \in (0, 1)$ hledáme $t \in \mathbb{R}$ takové, že $F_X(t) = \alpha$. To nemusí existovat, ale vždy existuje $t$, pro které

$$P(X < t) \leq \alpha \leq P(X \leq t),$$

tj.

$$\lim_{u \to t^-} F_X(u) \leq \alpha \leq F_X(t),$$

Všechna taková $t$ tvoří omezený interval, z něhož bereme (obvykle) střed,

$$q_X(\alpha) = \frac{1}{2} \left( \sup \{ t \in \mathbb{R} \mid P(X < t) \leq \alpha \} + \inf \{ t \in \mathbb{R} \mid \alpha \leq P(X \leq t) \} \right).$$

Číslo $q_X(\alpha)$ se nazývá $\alpha$-kvantil náhodné veličiny $X$ a funkce $q_X : (0, 1) \to \mathbb{R}$ je kvantilová funkce náhodné veličiny $X$. Speciálně $q_X(\frac{1}{2})$ je medián, další kvartily mají také svá jména – tercil, kvartil (dolní $q_X(\frac{1}{4})$, horní $q_X(\frac{3}{4})$) ... decil ... centil neboli percentil ....

Vlastnosti kvantilové funkce:

- neklesající,
- $q_X(\alpha) = \frac{1}{2} \left( \lim_{\beta \to \alpha^-} q_X(\beta) + \lim_{\beta \to \alpha^+} q_X(\beta) \right)$.

Věta: Tyto podmínky jsou nutné i postačující.

Obrácený převod:

$$F_X(t) = \inf \{ \alpha \in (0, 1) \mid q_X(\alpha) > t \} = \sup \{ \alpha \in (0, 1) \mid q_X(\alpha) \leq t \}.$$

Funkce $F_X, q_X$ jsou navzájem inverzní tam, kde jsou spojité a rostoucí (tyto podmínky stačí ověřit pro jednu z nich).
4.6 Jak reprezentovat náhodnou veličinu v počítači

1. **Diskrétní**: Nabývá-li pouze konečného počtu hodnot $t_k$, $k = 1, \ldots, n$, stačí k reprezentaci tyto hodnoty a jejich pravděpodobnosti $p_X(t_k) = P_X(\{t_k\}) = P(X = t_k)$, čímž je plně popsána pravděpodobnostní funkce $2n$ čísly (až na nepřesnost zobrazení reálných čísel v počítači).

Pokud diskrétní náhodná veličina nabývá (spojeně) nekonečně mnoho hodnot, musíme některé vynechat, zejména ty, které jsou málo pravděpodobné. Pro každé $\varepsilon > 0$ lze vybrat konečně mnoho hodnot $t_k$, $k = 1, \ldots, n$, tak, že $P_X(\mathbb{R}\setminus\{t_1, \ldots, t_n\}) = P(X \notin \{t_1, \ldots, t_n\}) \leq \varepsilon$. Zbývá však problém, jakou hodnotu přiřadit zbývajícím (byť málo pravděpodobným) případům.

2. **(Absolutně) spojitá**: Hustotu můžeme přibližně popsat hodnotami $f(t_k)$ v „dostatečně mnoha“ bodech $t_k$, $k = 1, \ldots, n$, ale jen za předpokladu, že je „dostatečně hladká“. Zajímají nás z ní spíše integrály typu

$$F_X(t_{k+1}) - F_X(t_k) = \int_{t_k}^{t_{k+1}} f_X(u) \, du,$$

z nichž lze přibližně zkonstruovat distribuční funkci. Můžeme pro reprezentaci použít přímo hodnoty distribuční funkce $F_X(t_k)$. Tam, kde je hustota velká, potřebujeme volit body husté.

Můžeme volit body $t_k$, $k = 1, \ldots, n$, tak, aby přírůstky $F_X(t_{k+1}) - F_X(t_k)$ měly zvolenou velikost. Zvolíme tedy $\alpha_k \in (0, 1)$, $k = 1, \ldots, n$, a k nim najdeme čísla $t_k = q_X(\alpha_k)$.

Paměťová náročnost je velká, závisí na jemnosti škály hodnot náhodné veličiny, resp. její distribuční funkce.

Často je rozdělení známého typu a stačí doplnit několik parametrů, aby bylo plně určeno.

Mnohé obecnější případy se snažíme vyjádřit alespoň jako směs náhodných veličin s rozděleními známého typu, abychom vystačili s konečně mnoha parametry.

3. **Smíšená**: Jako u spojité náhodné veličiny. Tento popis je však pro diskrétní část zbytečně nepřesný.

Můžeme použít rozklad na diskrétní a spojitou část.

4.7 Operace s náhodnými veličinami

Zde $I, J \subseteq \mathbb{R}$ jsou intervaly nebo spočetná sjednocení intervalů.

**Přičtení konstanty** $r$ odpovídá posunutí ve směru vodorovné osy:

\[
P_{X+r}(I + r) = P_X(I), \quad P_{X+r}(J) = P_X(J - r), \\
F_{X+r}(t + r) = F_X(t), \quad F_{X+r}(u) = F_X(u - r), \\
q_{X+r}(\alpha) = q_X(\alpha) + r.
\]
Vynásobení nenulovou konstantou \( r \) odpovídá podobnost ve směru vodorovné osy.

\[
P_{rX}(rI) = P_X(I), \quad P_{rX}(J) = P_X \left( \frac{J}{r} \right).
\]

Pro distribuční funkci musíme rozlišit případy:
\[ r > 0: \quad F_{rX}(rt) = F_X(t), \quad F_{rX}(u) = F_X\left(\frac{u}{r}\right), \quad q_{rX}(\alpha) = r q_X(\alpha), \]

\[ r = -1: \quad F_{-X}(-t) = P(-X \leq -t) = P(X \geq t) = 1 - P(X < t), \quad \text{v bodech spojitosti distribuční funkce} \]

\[ F_{-X}(-t) = 1 - P(X < t) = 1 - P(X \leq t) = 1 - F_X(t), \]
\[ F_X(u) = 1 - F_X(-u), \quad \text{v bodech nespojitosti limita zprava} \]
(středová symetrie grafu podle bodu \((0, \frac{1}{2})\) s opravou na spojitost zprava),
\[ q_X(\alpha) = -q_X(1 - \alpha). \]

\[ r < 0: \quad \text{kombinace předchozích případů.} \]

**Zobrazení spojitou rostoucí funkci** \( h \):
\[
P_{h(X)}(h(I)) = P_X(I), \quad F_{h(X)}(h(t)) = F_X(t), \quad F_{h(X)}(u) = F_X(h^{-1}(u)), \quad q_{h(X)}(\alpha) = h(q_X(\alpha)) \quad \text{v bodech spojitosti kvantilové funkce.} \]

**Zobrazení neklesající funkce** \( h \):
\[ F_{h(X)}(u) = \sup \{ F_X(t) \mid h(t) \leq u \}. \]

**Zobrazení nerostoucí funkce** \( h \) lze řešit jako zobrazení náhodné veličiny \( -X \) neklesající funkci \( g(t) = h(-t) \).

**Součet náhodných veličin** není jednoznačně určen, jedině za předpokladu nezávislosti. Ani pak není vztah jednoduchý.

**Směs náhodných veličin** viz výše. Na rozdíl od součtu je plně určena (marginálními) rozděleními vstupních náhodných veličin a koeficienty směsi.

\[ h(\text{Mix}_c(U, V)) = \text{Mix}_c(h(U), h(V)) \]

(je jedno, jestli jakoukoli funkci \( h \) aplikujeme před, nebo po vytvoření směsi)

### 4.8 Jak realizovat náhodnou veličinu na počítači

1. Vytvoříme náhodný (nebo pseudonáhodný) generátor náhodné veličiny \( X \) s rovnoměrným rozdělením na \((0, 1)\).

2. Náhodná veličina \( q_Y(X) \) má stejné rozdělení jako \( Y \). (Stačí tedy na každou realizaci náhodné veličiny \( X \) aplikovat funkci \( q_Y \).)
Všechna rozdělení spojitých náhodných veličin jsou stejná až na (nelineární) změnu měřítka.

5 Charakteristiky náhodných veličin

5.1 Střední hodnota

Motivační příklad (omezení kouření):
Chceme vyhodnotit, zda zákaz kouření v restauracích vedl k celkovému omezení kouření. Značení: E. nebo μ. Může být definována zvláště pro

- **diskrétní** náhodnou veličinu $U$: $EU = \sum_{t \in \mathbb{R}} t \cdot p_U(t) = \sum_{t \in \Omega_U} t \cdot p_U(t)$,

- **spojitou** náhodnou veličinu $V$: $EV = \int_{-\infty}^{\infty} t \cdot f_V(t) \, dt$,

- **směs** náhodných veličin $X = \text{Mix}_c(V, U)$: $EX = c \cdot EV + (1 - c) \cdot EU$.

(může být $V$ diskrétní, $U$ spojitá; toto není linearita střední hodnoty!)

Lze vyjít z definice pro diskrétní náhodnou veličinu a ostatní případy dostat jako limitu pro aproximaci jiných rozdělení diskrétním (nebo naopak).

Všechny tři případy pokrývá univerzální vzorec s použitím kvantilové funkce

$$EX = \frac{1}{0} q_X(\alpha) \, d\alpha.$$ 

Ten lze navíc jednoduše zobecnit na střední hodnotu jakékoli funkce $h$ náhodné veličiny:

$$E(h(X)) = \int_{0}^{1} h(q_X(\alpha)) \, d\alpha.$$ 

Speciálně pro **diskrétní** náhodnou veličinu

$$E(h(U)) = \sum_{t \in \Omega_U} h(t) \cdot p_U(t),$$

pro **spojitou** náhodnou veličinu

$$E(h(V)) = \int_{-\infty}^{\infty} h(t) \cdot f_V(t) \, dt.$$ 

(Ale funkce spojité náhodné veličiny nemusí být spojitá náhodná veličina.)

Střední hodnota je vodorovnou souřadnicí těžiště grafu distribuční funkce, jsou-li jeho elementy váženy přírůstkem distribuční funkce:
Pokud pracujeme se střední hodnotou, automaticky předpokládáme, že existuje a je konečná (což není vždy splněno).

**Vlastnosti střední hodnoty**

\[
\begin{align*}
E(r) &= r, \\
E(X + Y) &= EX + EY, \\
E(X - Y) &= EX - EY, \\
E(rX) &= rEX, \\
E(X^2) &= (EX)^2 + DX,
\end{align*}
\]

speciálně \( E(EX) = EX \), speciálně \( E(X + r) = EX + r \), obecněji \( E(rX + sY) = rEX + sEY \).

*(To je lineáritá střední hodnoty.)*

\[
E(Mix_c(V, U)) = cEV + (1 - c)EU.
\]

*(To není lineáritá střední hodnoty.)*

Pouze pro nezávislé náhodné veličiny

\[
E(X \cdot Y) = EX \cdot EY.
\]

### 5.2 Rozptyl (disperze)

**Motivační příklad (stejné podnebí):**

Chceme najít místo s podobným podnebím. Průměrná teplota nestačí. Důležité je i kolísání teplot.

Značení: \( \sigma^2 \), D., var.

\[
DX = E\left((X - EX)^2\right) = E\left(X^2\right) - (EX)^2,
\]

\[
E\left(X^2\right) = (EX)^2 + DX.
\]

*(1)*

**Vlastnosti:**

\[
DX = \int_0^1 (q_X(\alpha) - EX)^2 \, d\alpha.
\]

\[
\begin{align*}
DX &\geq 0, \\
Dr &= 0, \\
D(X + r) &= DX, \\
D(rX) &= r^2 DX.
\end{align*}
\]
\[
D (\text{Mix}_c(V, U)) = E (\text{Mix}_c(V, U)^2) - (E (\text{Mix}_c(V, U)))^2 \\
= c E (V^2) + (1 - c) E (U^2) - (c E V + (1 - c) E U)^2 \\
= c \left( D V + (E V)^2 \right) + (1 - c) \left( D U + (E U)^2 \right) \\
- \left( c^2 (E V)^2 + 2 c (1 - c) E V E U + (1 - c)^2 (E U)^2 \right) \\
= c D V + (1 - c) D U + c (1 - c) (E V)^2 \\
- 2 c (1 - c) E V E U + c (1 - c) (E U)^2 \\
= c D V + (1 - c) D U + c (1 - c) (E V - E U)^2 .
\]

Pouze pro nezávislé náhodné veličiny

\[
D (X + Y) = DX + DY, \quad D (X - Y) = DX + DY .
\]

### 5.3 Směrodatná odchylka

Značení: \( \sigma \).

Má stejný fyzikální rozměr jako původní náhodná veličina (rozptyl nikoli).

\[
\sigma_X = \sqrt{D X} = \sqrt{E \left( (X - E X)^2 \right)} 
\]

Vlastnosti:

\[
\sigma_X = \sqrt{\int_0^1 \left( q_X(\alpha) - E X \right)^2 d \alpha}.
\]

\[
\sigma_X \geq 0 , \quad \sigma_r = 0 , \quad \sigma_{X+r} = \sigma_X , \quad \sigma_{rX} = |r| \sigma_X .
\]

Pouze pro nezávislé náhodné veličiny

\[
\sigma_{X+Y} = \sigma_{X-Y} = \sqrt{DX + DY} = \sqrt{\sigma_X^2 + \sigma_Y^2} .
\]

### 5.4 Obecné a centrální momenty

\( k \in \mathbb{N} \)

k-tý obecný moment (značení nezavádíme): \( E (X^k) \), speciálně:

pro \( k = 1 \): \( EX \),
pro \( k = 2 \): \( E (X^2) = (EX)^2 + DX \).

Alternativní značení: \( m_k, \mu'_k \).

k-tý centrální moment (značení nezavádíme): \( E \left( (X - EX)^k \right) \), speciálně:

pro \( k = 1 \): \( 0 \),
pro $k = 2$: $DX$.

Alternativní značení: $\mu_k$.

Pomocí kvantilové funkce:

$$E(X^k) = \int_0^1 (q_X(\alpha))^k d\alpha.$$  
$$E((X - EX)^k) = \int_0^1 (q_X(\alpha) - EX)^k d\alpha.$$

### 5.5 Normovaná náhodná veličina

je taková, která má nulovou střední hodnotu a jednotkový rozptyl:

$$\text{norm } X = \frac{X - EX}{\sigma_X}$$

(pokud má vzorec smysl). Zpětná transformace je

$$X = EX + \sigma_X \text{ norm } X.$$  \hfill (2)

**Motivační příklad (biochemická vyšetření):**

Laboratorní výsledky vydají mnoho čísel; abychom poznali, která jsou obvyklá a která znepokojivá, museli bychom znát alespoň jejich střední hodnoty a směrodatné odchylky. Po znormování hned vidíme, které údaje zasluhují pozornost, aniž bychom museli studovat jejich typické hodnoty.

### 5.6 Základní typy diskrétních rozdělení

#### 5.6.1 Diracovo

Jediný možný výsledek $r \in \mathbb{R}$.

$$p_X(r) = 1, \quad EX = r, \quad DX = 0.$$  

Všechna diskrétní rozdělení jsou směsi Diracových rozdělení.

#### 5.6.2 Rovnoměrné

Je $m$ možných výsledků stejně pravděpodobných.

Speciálně pro obor hodnot $\{1, 2, \ldots, m\}$ dostáváme

$$p_X(k) = \frac{1}{m}, \quad k \in \{1, 2, \ldots, m\},$$  
$$EX = \frac{m + 1}{2}, \quad DX = \frac{1}{12} (m + 1) (m - 1).$$
5.6.3 Alternativní (Bernoulliovo)
Jsou 2 možné výsledky. (Směs dvou Diracových rozdělení.)
Pokud výsledky jsou 0,1, kde 1 má pravděpodobnost \( q \in (0,1) \), dostáváme

\[
p_X(1) = q, \quad p_X(0) = 1 - q, \quad EX = q, \quad DX = q(1-q).
\]

5.6.4 Binomické Bi\((m,q)\)
Počet úspěchů z \( m \) nezávislých pokusů, je-li v každém stejná pravděpodobnost úspěchu \( q \in (0,1) \). (Součet nezávislých alternativních rozdělení.)

\[
p_X(k) = \binom{m}{k} q^k (1-q)^{m-k}, \quad k \in \{0,1,2,\ldots,m\}, \quad EX = mq, \quad DX = mq(1-q).
\]

Součet nezávislých náhodných veličin s rozděleními Bi\((m,q)\), Bi\((n,q)\) má binomické rozdělení Bi\((m+n,q)\).
Výpočetní složitost výpočtu \( p_X(k) \) je \( O(k) \), celého rozdělení \( O(m^2) \).

5.6.5 Poissonovo Po\((\lambda)\)
Limitní případ binomického rozdělení pro \( m \to \infty \) při konstantním \( mq = \lambda > 0 \) (tedy \( q \to 0 \)).

\[
p_X(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k \in \{0,1,2,\ldots\}.
\]

<table>
<thead>
<tr>
<th>hodnota</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bi(30,0.1)</td>
<td>0.042</td>
<td>0.141</td>
<td>0.228</td>
<td>0.236</td>
<td>0.177</td>
<td>0.102</td>
<td>0.047</td>
<td>0.018</td>
</tr>
<tr>
<td>Bi(100,0.03)</td>
<td>0.047</td>
<td>0.147</td>
<td>0.225</td>
<td>0.227</td>
<td>0.171</td>
<td>0.101</td>
<td>0.050</td>
<td>0.021</td>
</tr>
<tr>
<td>Po(3)</td>
<td>0.050</td>
<td>0.149</td>
<td>0.224</td>
<td>0.224</td>
<td>0.168</td>
<td>0.101</td>
<td>0.050</td>
<td>0.022</td>
</tr>
</tbody>
</table>

Pravděpodobnostní funkce Poissonova rozdělení a binomických rozdělení se stejnou střední hodnotou 3

Jednotlivé pravděpodobnosti se počítají snáze než u binomického rozdělení (ovšem všechny nevypočítáme, protože jich je nekonečně mnoho).

\[
EX = \lambda, \quad DX = \lambda.
\]

„Střední hodnota se rovná rozptylu;“ jedná se vždy o bezrozměrné celočíselné náhodné veličiny (počet výskytů).
Poissonovo rozdělení jako limitní případ binomického \( m q_m = \lambda \), tj. \( q_m = \frac{\lambda}{m} \):

\[
p_X(k) = \binom{m}{k} q_m^k (1 - q_m)^{m-k} = \frac{m (m-1) \ldots (m-(k-1))}{k!} \left( \frac{\lambda}{m} \right)^k \left( 1 - \frac{\lambda}{m} \right)^{m-k} = \frac{\lambda^k}{k!} \prod_{i=1}^{k} \left( 1 - \frac{1}{m} \right) \to \frac{\lambda^k}{k!} e^{-\lambda}
\]

Součet nezávislých náhodných veličin s Poissonovým rozdělením \( \text{Po}(\lambda) \), \( \text{Po}(\mu) \) má Poissonovo rozdělení \( \text{Po}(\lambda + \mu) \).

5.6.6 Geometrické

Počet úspěchů do prvního neúspěchu, je-li v každém pokusu stejná pravděpodobnost úspěchu \( q \in (0, 1) \).

\[
p_X(k) = q^k (1 - q), \quad k \in \{0, 1, 2, \ldots\},
\]

\[
\mathbb{E} X = \frac{q}{1-q}, \quad \text{DX} = \frac{q}{(1-q)^2}.
\]

5.6.7 Hypergeometrické

Počet výskytů v \( m \) vzorcích, vybraných z \( M \) losů, v nichž je \( K \) výskytů \( (1 \leq m \leq K \leq M) \).

\[
p_X(k) = \frac{\binom{K}{k} \binom{M-K}{m-k}}{\binom{M}{m}}, \quad k \in \{0, 1, 2, \ldots, m\},
\]

\[
\mathbb{E} X = \frac{m K}{M}, \quad \text{DX} = \frac{m K (M-K) (M-m)}{M^2 (M-1)}.
\]

Výpočetná složitost výpočtu \( p_X(k) \) je \( O(m) \), celého rozdělení \( O(m^2) \).

Binomické rozdělení jako limitní případ hypergeometrického \( m \to \infty \) při konstantním \( \frac{K_M}{M} = q \), tj. \( \frac{M-K_M}{M} = 1-q \):

\[
p_X(k) = \frac{\binom{K_M}{k} \binom{M-K_M}{m-k}}{\binom{M}{m}} \to \frac{K_M^k}{m^k} \cdot \frac{(1-q)^{m-k}}{m^k} = \frac{m!}{k! (m-k)!} \cdot \frac{K_M^k}{M^k} \cdot \frac{(M-K_M)^{m-k}}{M^{m-k}} = \binom{m}{k} q^k (1-q)^{m-k}.
\]
5.7 Základní typy spojitých rozdělení

5.7.1 Rovnoměrné \( R(a, b) \)

\[
f_X(t) = \begin{cases} \frac{1}{b-a} & \text{pro } t \in (a, b), \\ 0 & \text{jinak,} \end{cases}
\]

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

\[
q_X(\alpha) = a + (b - a) \alpha,
\]

\[
EX = \frac{a + b}{2}, \quad DX = \frac{1}{12} (b - a)^2, \quad \sigma_X = \frac{1}{2} \sqrt{3} (b - a).
\]

5.7.2 Normální (Gaussovo) \( N(\mu, \sigma^2) \)

A. Normované \( N(0, 1) \):

\[
\varphi(t) = f_{N(0,1)}(t) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{t^2}{2} \right)
\]

Distribuční funkce je transcendentní (Gaussův integrál) \( \Phi \),

\[
\Phi(u) = F_{N(0,1)}(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{t^2}{2} \right) \, dt,
\]

kvantilová funkce \( \Phi^{-1} \) je inverzní k \( \Phi \).

B. Obecné \( N(\mu, \sigma^2) = \mu + \sigma \cdot N(0, 1) \):

\[
f_{N(\mu, \sigma^2)}(t) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{(t - \mu)^2}{2 \sigma^2} \right), \quad EX = \mu, \quad DX = \sigma^2.
\]

Součet dvou nezávislých veličin s normálním rozdělením \( N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2) \) má normální rozdělení \( N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \).

5.7.3 Logaritmickonormální \( LN(\mu, \sigma^2) = \exp(N(\mu, \sigma^2)) \)

je rozdělení náhodné veličiny \( X = \exp(Y) \), kde \( Y \) má \( N(\mu, \sigma^2) \)

\[
f_X(u) = \begin{cases} \frac{1}{u \sigma \sqrt{2\pi}} \exp \left( -\frac{(\ln u - \mu)^2}{2 \sigma^2} \right) = \frac{f_{N(\mu, \sigma^2)}(\ln u)}{u} & \text{pro } u > 0, \\ 0 & \text{jinak,} \end{cases}
\]

\[
F_X(u) = \begin{cases} F_{N(\mu, \sigma^2)}(\ln u) & \text{pro } u > 0, \\ 0 & \text{jinak,} \end{cases}
\]

\[
EX = \exp \left( \mu + \frac{\sigma^2}{2} \right),
\]

\[
DX = (\exp (2 \mu + \sigma^2)) (\exp (\sigma^2) - 1),
\]

\[
\sigma_X = \left( \exp \left( \mu + \frac{\sigma^2}{2} \right) \right) \sqrt{\exp (\sigma^2) - 1}.
\]
Součin dvou nezávislých veličin s logaritmickonormálním rozdělením LN($\mu_1, \sigma^2_1$), LN($\mu_2, \sigma^2_2$) má logaritmickonormální rozdělení LN($\mu_1 + \mu_2, \sigma^2_1 + \sigma^2_2$).

5.7.4 Exponenciální $\text{Ex}(\tau)$

Např. rozdělení času do první poruchy, jestliže (podmíněná) pravděpodobnost poruchy za časový interval $[t, t + \delta]$ závisí jen na $\delta$, nikoli na $t$:

$$f_X(t) = \begin{cases} \frac{1}{\tau} \exp\left(-\frac{t}{\tau}\right) & \text{pro } t > 0, \\ 0 & \text{jinak,} \end{cases}$$

$$F_X(u) = \begin{cases} 1 - \exp\left(-\frac{u}{\tau}\right) & \text{pro } u > 0, \\ 0 & \text{jinak,} \end{cases}$$

$$q_X(\alpha) = -\tau \ln (1 - \alpha),$$

$$\text{EX} = \tau,$$

$$\text{DX} = \tau^2,$$

$$\sigma_X = \tau,$$

$$\text{Ex}(\tau) = \tau \text{Ex}(1).$$

5.8 Čebyševova nerovnost

Motivační příklad (10 000 hodů mincí):

Při 10 000 hodech mincí má počet líčů $X$ binomické rozdělení $\text{Bi}(10 000, \frac{1}{2})$ se střední hodnotou $\text{EX} = 5 000$, směrodatnou odchylkou $\sigma_X = \sqrt{10 000 \cdot \frac{1}{4}} = 50$.

Jak malá je pravděpodobnost, že se výsledek bude lišit od střední hodnoty o nejméně 200 = $4\sigma_X$?

$$\sum_{k=0}^{4 800} p_{\text{Bi}(10 000,0.5)}(k) + \sum_{k=5 200}^{10 000} p_{\text{Bi}(10 000,0.5)}(k) =$$

$$= 1 - \sum_{k=4 801}^{5 199} p_{\text{Bi}(10 000,0.5)}(k) = 1 - \sum_{k=4 801}^{5 199} \binom{10 000}{k} \frac{1}{2^{10 000}} \approx 0.0000659.$$

Buffonova úloha – přesnost odhadu
Při $n$ hodech ježlou má počet protnutí linky $X$ binomické rozdělení $\text{Bi}(n, 2/\pi)$ se
střední hodnotou $\text{EX} = \frac{2}{\pi} n \doteq 0.6366 n$.

směrodatnou odchylkou $\sigma_X = \sqrt{\frac{2}{\pi} \left(1 - \frac{2}{\pi}\right) n} \doteq 0.48 \sqrt{n}$.

Kolik hodů $n$ potřebujeme, abychom $\pi$ odhadli na 99 % s přesností 1 %?

Uf! Šetřme elektřinu, ne hlavu!

Věta:

$$\forall \delta > 0 : P(|\text{norm} X| \geq \delta) \leq \frac{1}{\delta^2},$$

kde $\text{norm} X = \frac{X - \text{EX}}{\sigma_X}$ (pokud má výraz smysl).

Důkaz pomocí kvantilové funkce:

$$1 = D(\text{norm} X) = E\left((\text{norm} X)^2\right) - \left(E(\text{norm} X)\right)^2 = \int_0^1 (q_{\text{norm} X}(\alpha))^2 d\alpha$$

$$\geq \int_I (q_{\text{norm} X}(\alpha))^2 d\alpha,$$

kde $I = \{\alpha \in (0,1) : |q_{\text{norm} X}(\alpha)| \geq \delta\}$ jsou 2 intervylo blový delce $P(|\text{norm} X| \geq \delta)$,

$$1 \geq \int_I (q_{\text{norm} X}(\alpha))^2 d\alpha \geq \int_I \delta^2 d\alpha = \delta^2 P(|\text{norm} X| \geq \delta).$$

Ekvivalentní tvary ($\varepsilon = \delta \sigma_X$):

$$\forall \delta > 0 : P(|\text{norm} X| < \delta) \geq 1 - \frac{1}{\delta^2},$$

$$\forall \delta > 0 : P\left(\left|\frac{X - \text{EX}}{\sigma_X}\right| \geq \delta\right) \leq \frac{1}{\delta^2},$$

$$\forall \varepsilon > 0 : P(|X - \text{EX}| \geq \varepsilon) \leq \frac{\sigma_X^2}{\varepsilon^2} = \frac{\text{DX}}{\varepsilon^2},$$

$$\forall \varepsilon > 0 : P(|X - \text{EX}| < \varepsilon) \geq 1 - \frac{\sigma_X^2}{\varepsilon^2} = 1 - \frac{\text{DX}}{\varepsilon^2}.$$

Motivační příklad (10 000 hodů mincí – pokračování):
Jak malá je pravděpodobnost odchylky od střední hodnoty o více než $4 \sigma_X$?

Buffonova úloha – přesnost odhadu
Při $n$ hodech ježlou má počet protnutí linky $X$ binomické rozdělení $\text{Bi}(n, 2/\pi)$. Kolik hodů $n$ potřebujeme, abychom $\pi$ odhadli na 99 % s přesností 1 %?
\[
0.01 \geq \frac{\sigma_X^2}{\varepsilon^2} = \frac{\sigma_X^2}{(0.01 \text{EX})^2} = \frac{2}{\pi} \left(1 - \frac{2}{\pi}\right) n = \frac{10000 \left(\frac{\pi}{2} - 1\right)}{n},
\]

\[
n \geq 1000000 \left(\frac{\pi}{2} - 1\right) \approx 570796.
\]

Uf! Šetřme ruce, ne hlavu!

Distribuční funkce absolutní hodnoty normovaného normálního rozdělení \( |N(0, 1)| \) (červeně) ve srovnání s mezí dle Čebyševovy nerovnosti (modré)

Distribuční funkce absolutní hodnoty normovaného spojitého rovnoměrného rozdělení \( |\mathbb{R} (-\sqrt{3}, \sqrt{3})| \) (červeně) ve srovnání s mezí dle Čebyševovy nerovnosti (modré)

Distribuční funkce absolutní hodnoty normovaného binomického rozdělení \( |\text{norm Bi}(2, 0.5)| \) (červeně) ve srovnání s mezí dle Čebyševovy nerovnosti (modré)

6 Náhodné vektory

Náhodný vektor \((n\)-rozměrná náhodná veličina) na pravděpodobnostním prostoru \((\Omega, \mathcal{A}, P)\) je měřitelná funkce \(X: \Omega \to \mathbb{R}^n\), tj. taková, že pro každý \(n\)-rozměrný interval \(I\) platí

\[
X^{-1}(I) = \{\omega \in \Omega \mid X(\omega) \in I\} \in \mathcal{A}.
\]

Lze psát

\[
X(\omega) = (X_1(\omega), \ldots, X_n(\omega)),
\]
kde zobrazení \( X_k : \Omega \rightarrow \mathbb{R}, k = 1, \ldots, n \), jsou náhodné veličiny. 
Náhodný vektor lze považovat za vektor náhodných veličin \( \mathbf{X} = (X_1, \ldots, X_n) \). 
Je popsaný pravděpodobnostmi 

\[
P_X(I_1 \times \ldots \times I_n) = P(X_1 \in I_1, \ldots, X_n \in I_n) = P(\{ \omega \in \Omega \mid X_1(\omega) \in I_1, \ldots, X_n(\omega) \in I_n \}) ,
\]

kde \( I_1, \ldots, I_n \) jsou intervaly v \( \mathbb{R} \). 
Z těch vyplývají pravděpodobnosti 

\[
P_X(I) = P(\mathbf{X} \in I) = P(\{ \omega \in \Omega \mid \mathbf{X}(\omega) \in I \}) ,
\]
definované pro libovolnou borelovskou množinu \( I \) v \( \mathbb{R}^n \) (speciálně pro libovolné sjednocení spočetně mnoha \( n \)-rozměrných intervalů) a určující rozdělení náhodného vektoru \( \mathbf{X} \). 

**Úspornější reprezentace:** Stačí intervaly tvaru \( I_k = (-\infty, t_k) \), \( t_k \in \mathbb{R} \), 

\[
P(X_1 \in (-\infty, t_1), \ldots, X_n \in (-\infty, t_n)) = P(X_1 \leq t_1, \ldots, X_n \leq t_n) = P_X((-\infty, t_1) \times \ldots \times (-\infty, t_n)) = F_X(t_1, \ldots, t_n) .
\]

\( F_X : \mathbb{R}^n \rightarrow (0, 1) \) je distribuční funkce náhodného vektoru \( \mathbf{X} \). Je 

- neklesající (ve všech proměnných),
- zprava spojitá (ve všech proměnných),
- \( \lim_{t_1 \to \infty, \ldots, t_n \to \infty} F_X(t_1, \ldots, t_n) = 1 \),
- \( \forall k \in \{1, \ldots, n\} \forall t_1, \ldots, t_{k-1}, t_{k+1}, \ldots, t_n : \lim_{t_k \to -\infty} F_X(t_1, \ldots, t_n) = 0 \).

**Věta:** Tyto podmínky jsou nutné, nikoli postačující. 

Postačující podmínky bychom dostali, kdybychom např. pro dvě dimenze přidali podmínky 

\[
0 \leq P_X((a, b) \times (c, d)) = F_X(b, d) - F_X(a, d) - F_X(b, c) + F_X(a, b) ,
\]
pro všechna \( a, b, c, d \in \mathbb{R}, a < b, c < d \).

Nestačí znát marginální rozdělení náhodných veličin \( X_1, \ldots, X_n \), neboť ta neobsahují informace o závislosti.

Podmínky nezávislosti pro složky náhodného vektoru jsou 

\[
F_{X_1, X_2}(t_1, t_2) = F_{X_1}(t_1) \cdot F_{X_2}(t_2) ,
\]
obecněji 

\[
F_X(t_1, \ldots, t_n) = \prod_{k=1}^{n} F_{X_k}(t_k) .
\]
6.1 Diskrét néhodný vektor

má všechny složky diskrét. Lze jej popsat též sdruženou pravděpodobností funkcí\[ p_X : \mathbb{R}^n \to [0, 1) \]
\[ p_X(t_1, \ldots, t_n) = P(X_1 = t_1, \ldots, X_n = t_n), \]
 která je nemulová jen ve spočetně mnoha bodech. Diskrétní náhodné veličiny \( X_1, \ldots, X_n \) jsou nezávislé, právě když
\[ P(X_1 = t_1, \ldots, X_n = t_n) = \prod_{i=1}^{n} P(X_i = t_i) \]
pro všechna \( t_1, \ldots, t_n \in \mathbb{R} \). Ekvivalentní formulace:
\[ p_X(t_1, \ldots, t_n) = \prod_{i=1}^{n} p_{X_i}(t_i). \]

6.2 Spojit néhodný vektor

má všechny složky spojité. Lze jej popsat též sdruženou hustotou pravděpodobnosti což je (každá) nezáporná funkce \( f_X : \mathbb{R}^n \to [0, \infty) \) taková, že
\[ F_X(t_1, \ldots, t_n) = \int_{-\infty}^{t_1} \cdots \int_{-\infty}^{t_n} f_X(u_1, \ldots, u_n) \, du_1 \cdots du_n, \]
pro všechna \( t_1, \ldots, t_n \in \mathbb{R} \). Pokud to jde, volíme
\[ f_X(t_1, \ldots, t_n) = \frac{\partial}{\partial t_1} \cdots \frac{\partial}{\partial t_n} F_X(t_1, \ldots, t_n) = D_1 D_2 \cdots D_n F_X(t_1, \ldots, t_n). \]
Speciálně pro intervaly \( \langle a_i, b_i \rangle \) dostáváme
\[ P(X_1 \in \langle a_1, b_1 \rangle, \ldots, X_n \in \langle a_n, b_n \rangle) = P_X(\langle a_1, b_1 \rangle \times \cdots \times \langle a_n, b_n \rangle) \]
\[ = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f_X(u_1, \ldots, u_n) \, du_1 \cdots du_n \]
Spojité náhodné veličiny \( X_1, \ldots, X_n \) jsou nezávislé, právě když
\[ f_X(t_1, \ldots, t_n) = \prod_{i=1}^{n} f_{X_i}(t_i). \]
pro skoro všechna \( t_1, \ldots, t_n \in \mathbb{R} \).

6.3 Obecnější náhodné veličiny

Komplexní náhodná veličina je náhodný vektor se dvěma složkami interpretovanými jako reálná a imaginární část.
Na těchto hodnotách nemusí být definována žádná aritmetika ani uspořádání. Mohli bychom všechny hodnoty očíslovat, ale není žádný důvod, proč bychom to měli udělat právě určitým způsobem (ktéby by ovlivnil následné numerické výpočty).
(Příklad: Číslování politických stran ve volbách.)
6.4 Číselné charakteristiky náhodného vektoru

Střední hodnota
- náhodného vektoru \( \mathbf{X} = (X_1, \ldots, X_n) \): \( \mathbb{E}\mathbf{X} := (\mathbb{E}X_1, \ldots, \mathbb{E}X_n) \)
- komplexní náhodné veličiny: \( X = \Re(X) + i\Im(X) \): \( \mathbb{E}X := \mathbb{E}\Re(X) + i\mathbb{E}\Im(X) \)
- nenumerické náhodné veličiny: nemá smysl

Rozptyl náhodného vektoru \( \mathbf{X} = (X_1, \ldots, X_n) \): \( \mathbf{D}\mathbf{X} := (D_X, \ldots, D_X) \)

Je-li \( U \) náhodná veličina, \( a, b \in \mathbb{R} \), pak \( aU + b \) má charakteristiky

\[
\mathbb{E}(aU + b) = a\mathbb{E}U + b, \quad \mathbf{D}(aU + b) = a^2\mathbf{D}U.
\]

Na rozdíl od jednorozměrné náhodné veličiny, střední hodnota a rozptyl náhodného vektoru nedávají dostatečnou informaci pro výpočet rozptylu jeho lineárních funkcí. Proto zavádíme další charakteristiky. Např.

\[
\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y,
\]

\[
\mathbf{D}(X + Y) = \mathbf{D}(X) + \mathbf{D}(Y) + 2\text{cov}(X,Y),
\]

kde \( \text{cov}(X,Y) := \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y \) je Kovariance náhodných veličin \( X,Y \), též

\[
\text{cov}(X,Y) = \mathbb{E}((X - \mathbb{E}X)(Y - \mathbb{E}Y)),
\]

neboť

\[
\mathbb{E}((X - \mathbb{E}X)(Y - \mathbb{E}Y)) = \mathbb{E}(XY - X\mathbb{E}Y - Y\mathbb{E}X + \mathbb{E}X\mathbb{E}Y) = \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y - \mathbb{E}X\mathbb{E}Y + \mathbb{E}X\mathbb{E}Y = 0.
\]

Pro existenci kovariance je postačující existence rozptylů \( DX, DY \).

Vlastnosti kovariance:
- \( \text{cov}(X,X) = DX \), \( \text{cov}(Y,Y) = DY \)
- \( \text{cov}(Y,X) = \text{cov}(X,Y) \)
- \( \text{cov}(aX + b, cY + d) = ac \text{cov}(X,Y) \) \( (a,b,c,d \in \mathbb{R}) \)

(srovnejte s vlastnostmi rozptylu jako speciálního případu), speciálně \( \text{cov}(X,-X) = -DX \).

Pro nezávislé náhodné veličiny \( X,Y \) je \( \text{cov}(X,Y) = 0 \).

Použitím kovariance pro normované náhodné veličiny vyjde korelace:

\[
\varrho(X,Y) = \text{cov}(\text{norm}X,\text{norm}Y) = \frac{\text{cov}(X,Y)}{\sigma_X\sigma_Y} = \mathbb{E}(\text{norm}X \cdot \text{norm}Y)
\]
Speciálně \( \varrho(X, X) = 1 \).

**Vlastnosti korelace**
\[
\begin{align*}
\varrho(X, X) &= 1, & \varrho(X, -X) &= -1, & \varrho(X, Y) &\in (-1, 1), \\
\varrho(Y, X) &= \varrho(X, Y), \\
\varrho(aX + b, cY + d) &= \text{sign}(ac) \varrho(X, Y) & (a, b, c, d \in \mathbb{R}, & a \neq 0 \neq c)
\end{align*}
\]
(až na znaménko nezáleží na prosté lineární transformaci).

Důsledek: \( \varrho(aX + b, X) = \text{sign}(a) \).

Jsou-li náhodné veličiny \( X, Y \) nezávislé, je \( \varrho(X, Y) = 0 \). Obrácená implikace však neplatí (není to postačující podmínka pro nezávislost). Náhodné veličiny \( X, Y \) splňující \( \varrho(X, Y) = 0 \) nazýváme nekorelované.

Pro náhodný vektor \( \mathbf{X} = (X_1, \ldots, X_n) \) je definována **kovarianční matice**
\[
\Sigma_{\mathbf{X}} = \\
\begin{bmatrix} 
\text{cov}(X_1, X_1) & \text{cov}(X_1, X_2) & \cdots & \text{cov}(X_1, X_n) \\
\text{cov}(X_2, X_1) & \text{cov}(X_2, X_2) & \cdots & \text{cov}(X_2, X_n) \\
\vdots & \vdots & \ddots & \vdots \\
\text{cov}(X_n, X_1) & \text{cov}(X_n, X_2) & \cdots & \text{cov}(X_n, X_n) \\
\text{cov}(X_1, X_n) & \text{cov}(X_2, X_n) & \cdots & \text{cov}(X_n, X_n) \\
\end{bmatrix}
\]

Je symetrická pozitivně semidefinitní, na diagonále má rozptyly.

Podobně je definována **korelační matice**
\[
\varrho_{\mathbf{X}} = \\
\begin{bmatrix} 
1 & \varrho(X_1, X_2) & \cdots & \varrho(X_1, X_n) \\
\varrho(X_1, X_2) & 1 & \cdots & \varrho(X_2, X_n) \\
\vdots & \vdots & \ddots & \vdots \\
\varrho(X_1, X_n) & \varrho(X_2, X_n) & \cdots & 1 \\
\end{bmatrix}
\]

Je symetrická pozitivně semidefinitní.

### 6.4.1 Vícerozměrné normální rozdělení \( N(\mu, \Sigma) \)

Popisuje speciální případ náhodného vektoru, jehož složky mají normální rozdělení a mohou být korelované. Má hustotu
\[
f_{N(\mu, \Sigma)}(\mathbf{t}) := \frac{1}{\sqrt{(2\pi)^n \det \mathbf{T}^{-1}}} \exp \left( -\frac{1}{2} (\mathbf{t} - \mu) \mathbf{T} (\mathbf{t} - \mu)^T \right),
\]

kde \( \mathbf{t} = (t_1, \ldots, t_n) \in \mathbb{R}^n \),
\( \mu = (\mu_1, \ldots, \mu_n) \in \mathbb{R}^n \),
\( \mathbf{T} \in \mathbb{R}^{n \times n} \) je matice, BÚNO symetrická.

Parametry rozdělení:
\( \mu = (\mu_1, \ldots, \mu_n) \in \mathbb{R}^n \) je střední hodnota náhodného vektoru,
\( \Sigma := \mathbf{T}^{-1} \) je kovarianční matice, speciálně její hlavní diagonála
\[(\Sigma_{11}, \Sigma_{22}, \ldots, \Sigma_{nn}) \in \mathbb{R}^n\] je rozptyl náhodného vektoru, marginální rozdělení i-té složky je \(N(\mu_i, \Sigma_{ii})\); pomocí těchto parametrů píšeme
\[
f_{\text{N}(\mu, \Sigma)}(t) := \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp \left( -\frac{1}{2} (t - \mu)^T \Sigma^{-1} (t - \mu) \right).
\]

6.5 Reprezentace náhodných vektorů v počítači

Obdobná jako u náhodných veličin, avšak s rostoucí dimenzí rychle roste paměťová náročnost. To by se nestalo, kdyby náhodné veličiny byly nezávislé; pak by stačilo znát marginální rozdělení.

Proto velkou úsporu může přínést i podmíněná nezávislost.

Pokud najdeme úplný systém jevů, které zajišťují podmíněnou nezávislost dvou náhodných veličin, pak můžeme jejich rozdělení popsat jako směs rozdělení nezávislých náhodných veličin (a tedy úsporněji).

7 Lineární prostor náhodných veličin

\((\Omega, A, P)\) pravděpodobnostní prostor,
\(\mathcal{L}\) lineární prostor všech náhodných veličin na \((\Omega, A, P)\), tj. \(A\)-měřitelných funkcí \(\Omega \rightarrow \mathbb{R}\), sčítání náhodných veličin a jejich násobení reálným číslem = operace s funkcemi (bod po bodu),
\(\mathcal{L}_2\) lineární podprostor všech náhodných veličin z \(\mathcal{L}\), které mají rozptyl,
\[
\bullet : \mathcal{L}_2 \times \mathcal{L}_2 \rightarrow \mathbb{R}, \quad X \bullet Y := E(XY),
\]
je bilineární (= lineární v obou argumentech) a komutativní operace, skalární součin (pokud ztotožníme náhodné veličiny \(X, Y\), pro které \(P(X \neq Y) = 0\); za prvky prostoru pak považujeme třídy ekvivalence místo jednotlivých náhodných veličin),
\[
||X|| := \sqrt{X \bullet X} = \sqrt{E(X^2)}
\]
je norma,
\[
d(X, Y) := ||X - Y|| = \sqrt{E((X - Y)^2)}
\]
je metrika (vzdálenost) (bez předchozího ztotožnění pouze pseudometrika, mohla by být nulová i pro \(X \neq Y\)).
\(\mathcal{L}_2\) lze rozložit na 2 ortogonální podprostory:
\(\mathcal{R}\) = jednodimenzionální prostor všech konstatních náhodných veličin (tj. s Diracovým rozdělením)
\(\mathcal{N}\) = prostor všech náhodných veličin s nulovou střední hodnotou.
\(\mathcal{E}X\) je kolmý průmět \(X\) do \(\mathcal{R}\)(pokud ztotožníme toto reálné číslo s příslušnou konstantní náhodnou veličinou, jinak souřadnice ve směru \(\mathcal{R}\)),
\[
X - \mathcal{E}X = \text{průmět} X \text{ do } \mathcal{N},
\]
\[
\sigma_X = ||X - \mathcal{E}X|| = \text{vzdálenost } X \text{ od } \mathcal{R}.
\]
Z kolnosti vektorů \(X - \mathcal{E}X \in \mathcal{N}, \mathcal{E}X \in \mathcal{R}\) a Pythagorovy věty plyne (1)
\[
X \bullet X = ||X||^2 = ||X - \mathcal{E}X||^2 + ||\mathcal{E}X||^2,
\]
\[
E(X^2) = D_X + (\mathcal{E}X)^2.
\]
7.1 Lineární podprostor \( \mathcal{N} \) náhodných veličin s nulovými středními hodnotami

Speciálně pro náhodné veličiny z \( \mathcal{N} \):

\[
\begin{align*}
\sigma^2_X &= X \cdot X, \\
\sigma_X &= ||X||, \\
\text{cov}(X,Y) &= X \cdot Y, \\
g(X,Y) &= \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{X \cdot Y}{||X|| \ ||Y||} = \cos \angle(X,Y).
\end{align*}
\]

Důsledek: Náhodné veličiny \( X,Y \) s nulovými středními hodnotami jsou ortogonální, právě když jsou nekorelované.

Obecně v \( \mathcal{L}_2 \)

\( g(X,Y) \) je kosinus úhlu průmětů \( X,Y \) do \( \mathcal{N} \),

\( \text{cov}(X,Y) = X \cdot Y - EXEY \) je skalární součin průmětů \( X,Y \) do \( \mathcal{N} \).

POZOR! Nepleťte nezávislost náhodných veličin s lineární nezávislostí v lineárním prostoru, který tvoří!

7.2 Lineární regrese

Úloha: Je dán náhodný vektor \( X = (X_1, \ldots, X_n) \) a náhodná veličina \( Y \).

(Předpokládáme, že všechny náhodné veličiny jsou z \( \mathcal{L}_2 \)). Máme najít takové koeficienty \( c_1, \ldots, c_n \) aby lineární kombinace \( \sum_i c_i X_i \) byla co nejlepší approximací náhodné veličiny \( Y \) ve smyslu kritéria

\[
||\sum_k c_k X_k - Y||.
\]

Řešení: K vektoru \( Y \) hledáme nejblížší bod v lineárním podprostoru, který je lineárním obalem vektorů \( X_1, \ldots, X_n \); řešením je kolmý průmět. Ten je charakterizován tím, že vektor \( \sum_i c_i X_i - Y \) je kolmý na \( X_j, j = 1, \ldots, n \),

\[
\left( \sum_k c_k X_k - Y \right) \cdot X_j = 0,
\]

\[
\sum_i c_i (X_i \cdot X_j) = Y \cdot X_j.
\]

To je soustava lineárních rovnic pro neznámé koeficienty \( c_1, \ldots, c_n \) (soustava normálních rovnic).

Speciálně pro náhodné veličiny s nulovými středními hodnotami:

\[
\sum_i c_i \text{cov}(X_i, X_j) = \text{cov}(Y, X_j),
\]

takže matice soustavy je kovarjanční matice \( \Sigma_X \).
8 Základní pojmy statistiky

8.1 K čemu potřebujieme statistiku

Zkoumání společných vlastností velkého počtu obdobných jevů. Přitom nezkoumáme všechny, ale jen vybraný vzorek (kvůli ceně testů, jejich destruktivnosti apod.).

- Odhady parametrů pravděpodobnostního modelu
- Testování hypotéz

Potíže statistického výzkumu – viz [Rogalewicz].

8.2 Náhodný výběr, odhad, empirické rozdělení

**Soubor**

- základní (=populace)
- výběrový

Náhodný výběr jednoho prvku základního souboru (s rovnoměrným rozdělením) a změření zkoumané veličiny na tomto prvku určuje rozdělení náhodné veličiny.

Opakovaným výběrem dostaneme náhodný vektor, jehož složky mají stejné rozdělení a jsou nezávislé.

Takto vytvoříme výběrový soubor rozsahu \( n \), obvykle však vylučíme víceasobný výběr stejného prvku (výběr bez vracení). Jeho rozdělení se může poněkud lišit od původního. Tento rozdíl se obvykle zanedbává, neboť

1. pro velký rozsah základního souboru to není podstatné,
2. rozsah základního souboru někdy není znám,
3. výpočty se značně zjednoduší.

**Přesnost odhadu je dány velikostí výběrového souboru, nikoli populace.**

Náhodný výběr \( X := (X_1, \ldots, X_n) \) je vektor náhodných veličin, které jsou nezávislé a mají stejné rozdělení.

(Vynecháváme indexy, např. \( F_X \) místo \( F_{X_k} \).)

Provedením pokusu dostaneme realizaci náhodného výběru, \( x := (x_1, \ldots, x_n) \in \mathbb{R}^n \), kde \( n \) je rozsah výběru.

<table>
<thead>
<tr>
<th>funkce</th>
<th>funkční hodnota</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f : D \to \mathbb{R} )</td>
<td>( f(x) \in \mathbb{R}, \quad x \in \mathbb{R} )</td>
</tr>
<tr>
<td>náhodná veličina</td>
<td>realizace náhodné veličiny</td>
</tr>
<tr>
<td>( X : \Omega \to \mathbb{R} )</td>
<td>( x := X(\omega) \in \mathbb{R}, \quad \omega \in \Omega )</td>
</tr>
</tbody>
</table>

Náhodný vektor/výběr \( X = (X_1, \ldots, X_n) : \Omega \to \mathbb{R}^n \) realizace náhodného vektoru/výběru \( x = (x_1, \ldots, x_n) := X(\omega) \in \mathbb{R}^n, \quad \omega \in \Omega \)

Realizace náhodného výběru může mít význam trénovací množiny; neznámé parametry odhadujeme tak, aby na trénovací množině byly optimální.

Popisuje ji empirické rozdělení: Vybereme \( j \in \{1, \ldots, n\} \) s rovnoměrným rozdělením, výsledku
je $x_j$.

Je to diskrétní rozdělení, směs Diracových: $\text{Mix}(1/n,\ldots,1/n)(x_1,\ldots,x_n)$.

**Statistika** je (každá) měřitelná funkce $G$, definovaná na náhodném výběru libovolného (dostatečného) rozsahu. (Počítá se z náhodných veličin výběru, nikoli z parametrů rozdělení.)

"Měřitelná" znamená, že pro každé $t \in \mathbb{R}$ je definována pravděpodobnost

$$P(G(X_1,\ldots,X_n) \leq t) = F_G(x_1,\ldots,x_n)(t).$$

Statistika jako funkce náhodných veličin je rovněž náhodná veličina. Obvykle se používá jako **odhad parametrů rozdělení** (které nám zůstávají skryté).

Značení:

- $\vartheta$ ... jakákoli hodnota parametru (reálné číslo),
- $\vartheta^*$ ... skutečná (správná) hodnota parametru (reálné číslo),
- $\hat{\Theta}, \hat{\Theta}_n$ ... odhad parametru založený na náhodném výběru rozsahu $n$ (náhodná veličina)
- $\hat{\theta}, \hat{\theta}_n$ ... realizace odhadu (reálné číslo)

**Zádoucí vlastnosti odhadů:**

- $E\hat{\Theta}_n = \vartheta^*$, tj. $E(\hat{\Theta}_n - \vartheta^*) = 0$ nestranný (opak: vychýlený)
- $\lim_{n \to \infty} E\hat{\Theta}_n = \vartheta^*$, tj. $\lim_{n \to \infty} E(\hat{\Theta}_n - \vartheta^*) = 0$ asymptoticky nestranný
- eficientní = s malým rozptylem, což posuzujeme podle $E((\hat{\Theta}_n - \vartheta^*)^2) = D\hat{\Theta}_n + (E\hat{\Theta}_n - \vartheta^*)^2$, pro nestranný odhad se redukuje na $D\hat{\Theta}_n$
- nejlepší nestranný odhad je ze všech nestranných ten, který je nejvíce eficientní (mohou však existovat více eficientní vychýlené odhady)
- $\lim_{n \to \infty} E(\hat{\Theta}_n - \vartheta^*) = 0$, $\lim_{n \to \infty} \sigma\hat{\Theta}_n = 0$ konzistentní
- robustní, tj. odolný vůči šumu ("i při zašuměných datech dostáváme dobrý výsledek") – přesné kritérium chybí, ale je to velmi praktická vlastnost

### 8.3 Odhad střední hodnoty

Pomocí střední hodnoty empirického rozdělení (aritmetického průměru realizace náhodného výběru):

$$\bar{x} := E\text{Emp}(x) = \frac{1}{n} \sum_{j=1}^{n} x_j .$$

Když totéž provedeme s náhodnými veličinami výběru, dostaneme náhodnou veličinu

$$\bar{X} = \frac{1}{n} \sum_{j=1}^{n} X_j .$$

$\bar{X}$ = výběrový průměr, $\bar{x}$ = realizace výběrového průměru.

Alternativní značení: $\bar{X}_n, \bar{x}_n$ (pokud potřebujeme zdůraznit rozsah výběru)
Věta:

\[
E \bar{X}_n = \frac{1}{n} \sum_{j=1}^{n} EX = EX, \\
D \bar{X}_n = \frac{1}{n^2} \sum_{j=1}^{n} DX = \frac{1}{n} DX, \\
\sigma_{\bar{X}_n} = \sqrt{\frac{1}{n} DX} = \frac{1}{\sqrt{n}} \sigma_X,
\]
pokud existují. (Zde \(EX = EX_j\) atd.) Výběrový průměr minimalizuje kritérium nejmenších čtverců

\[
\ell_2(c) = E(c - \text{Emp}(x))^2 = \frac{1}{n} \sum_{i=1}^{n} (c - x_i)^2.
\]

Důsledek: Výběrový průměr je nestranný konzistentní odhad střední hodnoty.
(Nezávisle na typu rozdělení.)

Čeby sevova nerovnost pro \(X_n\) dává

\[
P(\left| \bar{X}_n - EX \right| \geq \varepsilon) \leq \frac{D \bar{X}_n}{\varepsilon^2} = \frac{DX}{n \varepsilon^2} \to 0 \quad \text{pro} \ n \to \infty.
\]
To platí i za obecnějších předpokladů (\(X_j\) nemusí mít stejné rozdělení) – slabý zákon velkých čísel.

Lidově se hovoří o „přesném součtu nepřesných čísel“, což je chyba, neboť součet \(\sum_{j=1}^{n} X_j\) má rozptyl \(nDX \to \infty\). **Relativní** chyba součtu **klesá, absolutní roste.**

Rozdělení výběrového průměru může být podstatně složitější než původní, jen ve speciálních případech je jednoduchá odpověď.

Věta: Výběrový průměr z **normálního** rozdělení \(N(\mu, \sigma^2)\) má normální rozdělení \(N (\mu, \frac{1}{n} \sigma^2)\) a je nejlepším nestranným odhadem střední hodnoty.

Podobná věta platí i pro jiná rozdělení alespoň asymptoticky:

**Centrální limitní věta:** Nechť \(X_j, \ j \in \mathbb{N},\) jsou nezávislé stejně rozdělené náhodné veličiny se střední hodnotou \(EX) a směrodatnou odchylkou \(\sigma_X \neq 0\). Pak normované náhodné veličiny

\[
Y_n = \text{norm} \bar{X}_n = \frac{\sqrt{n}}{\sigma_X} (\bar{X}_n - EX)
\]
konvergují k normovanému normálnímu rozdělení v následujícím smyslu:

\[
\forall t \in \mathbb{R} : \lim_{n \to \infty} F_{Y_n}(t) = \lim_{n \to \infty} F_{\text{norm} \bar{X}_n}(t) = \Phi(t),
\]
neboli

\[
\forall t \in \mathbb{R} : \lim_{n \to \infty} |F_{Y_n}(t) - \Phi(t)| = 0,
\]
Pokus má původní rozdělení 3. centrální moment, je konvergence dokonce **stejnoměrná**, tj.

\[
\lim_{n \to \infty} \sup_{t \in \mathbb{R}} |F_{Y_n}(t) - \Phi(t)| = 0,
\]

**Motivační příklad (10 000 hodů mincí):**
Jak malá je pravděpodobnost, že se výsledek bude lišit od střední hodnoty o nejméně 200 = 4 $\sigma_{X}$?

S pravděpodobností $1 - \Phi(4) \overset{?}{=} 1 - 0.9999683 = 0.0000317$ bude aspoň o $4 \sigma_{X}$ větší, se stejnou pravděpodobností o $4 \sigma_{X}$ menší, celkem

$$2 \left(1 - \Phi(4)\right) \overset{?}{=} 2 \cdot 0.0000317 = 0.0000633,$$

v dobré shodě s přesnějším (a pracným) výsledkem 0.0000659.

Kdybychom zohlednili kvantizační chybu a uvažovali toleranci $\pm 199.5$, dostali bychom ještě lepší odhad 0.0000661.

**Buffonova úloha – přesnost odhadu**

Binomické rozdělení $B(n, \frac{2}{\pi})$ při $n$ hodech se blíží normálnímu se střední hodnotou $EX = \frac{2}{\pi} \cdot n \overset{?}{=} 0.6366n$,

směrodatnou odchylkou $\sigma_{X} = \sqrt{\frac{2}{\pi} \left(1 - \frac{2}{\pi}\right) n} \overset{?}{=} 0.48 \sqrt{n}$.

Kolik hodů $n$ potřebujeme, abychom $\pi$ odhadli na 99 % s přesností 1 %?

S pravděpodobností 99 % normovaná hodnota bude v absolutní hodnotě menší než kvantil $\Phi^{-1}(0.995) \overset{?}{=} 2.576$.

To je tolerance měřená ve směrodatných odchylkách.

Má být nejvýše 1 % střední hodnoty,\

$$2.576 \sigma_{X} \leq 0.01 \cdot EX,$$

$$2.576 \sqrt{\frac{2}{\pi} \left(1 - \frac{2}{\pi}\right) n} \leq 0.01 \frac{2}{\pi} \cdot n,$$

$$\sqrt{n} \geq 2.576 \cdot 100 \sqrt{\frac{\pi}{2} - 1},$$

$$n \geq 2.576^{2} \cdot 10000 \left(\frac{\pi}{2} - 1\right) \overset{?}{=} 37 877.$$

Výsledný počet bude v mezích $24 113 \pm 241$ s pravděpodobností

$$\sum_{k=23 872}^{24 354} p_{Bi(37 877, \frac{2}{\pi})}(k) = \sum_{k=23 872}^{24 354} \left(\frac{37 877}{k}\right)^{k} \left(\frac{2}{\pi}\right)^{k} \left(1 - \frac{2}{\pi}\right)^{37 877-k} \overset{?}{=} 0.00988.$$

### 8.4 Odhad $k$-tého obecného momentu $EX^{k}$

Pomocí $k$-tého obecného momentu empirického rozdělení:

$$m_{x^{k}} := E\text{Emp}(x)^{k} = \frac{1}{n} \sum_{j=1}^{n} x_{j}^{k} \quad \text{(realizace výběrového $k$-tého obecného momentu)}.$$

Je realizací odhadu

$$M_{X^{k}} := \frac{1}{n} \sum_{j=1}^{n} X_{j}^{k} \quad \text{(výběrový $k$-tý obecný moment)}.$$

Alternativní značení: $M_{k}, m_{k}$. 
Věta: \( EM_{X^k} = EX^k \).
Výběrový \( k \)-tý obecný moment je nestranný konzistentní odhad \( k \)-tého obecného momentu (pokud \( X \) má \( k \)-tý a \( 2k \)-tý obecný moment).
Důkaz:

\[
EM_{X^k} = E \left( \frac{1}{n} \sum_{j=1}^{n} X_j^k \right) = \frac{1}{n} \sum_{j=1}^{n} EX_j^k = EX^k ,
\]
\[
DM_{X^k} = \frac{1}{n^2} n DX^k = \frac{1}{n} \left( E(X^k)^2 - (EX^k)^2 \right) = \frac{1}{n} \left( EX^{2k} - (EX^k)^2 \right) \to 0 .
\]

8.5 Odhad rozptylu
8.5.1 Odhad rozptylu při známé střední hodnotě
\[
\hat{\theta} = \frac{1}{n} \sum_{j=1}^{n} (x_j - EX)^2 .
\]
Je realizací odhadu
\[
\hat{\Theta} = \frac{1}{n} \sum_{j=1}^{n} (X_j - EX)^2 .
\]
\[
E\hat{\theta} = \frac{1}{n} \sum_{j=1}^{n} E(X_j - EX)^2 = DX .
\]
- nestranný,
- konzistentní, pokud existuje 4. centrální moment

Rozdělení odhadu rozptylu pro výběr z normálního rozdělení
\( X_j - EX \) má rozdělení \( N(0, DX) = \sigma_X N(0, 1) \)
\[
U_j := \frac{X_j - EX}{\sigma_X} \mbox{ má rozdělení } N(0, 1)
\]
\[
\sum_{j=1}^{n} U_j^2 = \frac{1}{DX} \sum_{j=1}^{n} (X_j - EX)^2 = \frac{n \hat{\Theta}}{DX} \mbox{ má}
\]

8.5.2 Rozdělení \( \chi^2 \) s \( n \) stupni volnosti, \( \chi^2(n) \)

= rozdělení náhodné veličiny \( Y = \sum_{j=1}^{n} U_j^2 \), kde \( U_j \) jsou nezávislé náhodné veličiny s normovaným normálním rozdělením \( N(0, 1) \).
\[
E\chi^2(n) = E \sum_{j=1}^{n} U_j^2 = \sum_{j=1}^{n} EU_j^2 = \sum_{j=1}^{n} DU_j = n , \quad \frac{E\chi^2(n)}{n} = 1 ,
\]
\[
D\chi^2(n) = D \sum_{j=1}^{n} U_j^2 = 2n , \quad \frac{D\chi^2(n)}{n} = \frac{2}{n} .
\]
Součet nezávislých náhodných veličin s rozděleními \( \chi^2(k), \chi^2(n) \) má rozdělení \( \chi^2(k+n) \).
Hustota

\[ f_Y(y) = \begin{cases} 
  c(n) \frac{y^{\frac{n}{2}-1}}{\left(\frac{n}{2}\right)^{\frac{n}{2}}} e^{-\frac{y}{2}} & \text{pro } y > 0, \\
  0 & \text{jinak},
\end{cases} \]

\[ c(n) = \frac{1}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}, \]

\[ \Gamma(z) = \int_{0}^{\infty} t^{z-1} e^{-t} \, dt, \]

speciálně \( \Gamma(m + 1) = m! \) pro všechna \( m \in \mathbb{N} \).

Speciálně pro \( n = 2 \) je \( c(n) = 1/2, \chi^2(2) = \text{Ex}(2) \) (exponenciální rozdělení), \( \frac{\chi^2(2)}{2} = \text{Ex}(1) \).

Důsledek: Součet \( m \) nezávislých náhodných veličin s exponenciálním rozdělením \( \text{Ex}(2) = \chi^2(2) \) má rozdělení \( \chi^2(2m) \).

Hustota rozdělení \( \chi^2 \) s 1, 2, ... , 10 stupni volnosti.

Hustota odmocniny z rozdělení \( \chi^2 \) s 1, 2, ... , 10 stupni volnosti („vzdálenosti od středu terče“).

8.5.3 Odhad rozptylu při neznámé střední hodnotě

pomocí rozptylu empirického rozdělení:

\[ \hat{\sigma}_x^2 := D_{\text{Emp}}(x) = \frac{1}{n} \sum_{j=1}^{n} (x_j - \bar{x})^2. \]

Je realizací odhadu

\[ \hat{\sigma}_X^2 := \frac{1}{n} \sum_{j=1}^{n} (X_j - \bar{X})^2 \neq \frac{1}{n} \sum_{j=1}^{n} (X_j - E_X)^2. \]
Vzorce jsou dvouprůchodové, ale lze je upravit na jednoprůchodové (numericky nevhodné):

\[
\hat{\sigma}_X^2 = \frac{1}{n} \left( \sum_{j=1}^{n} X_j^2 - 2 \bar{X} \sum_{j=1}^{n} X_j + \sum_{j=1}^{n} \bar{X}^2 \right) = \frac{1}{n} \sum_{j=1}^{n} X_j^2 - \bar{X}^2 ,
\]

\[
\hat{\sigma}_x^2 = \frac{1}{n} \sum_{j=1}^{n} x_j^2 - \bar{x}^2 = \frac{1}{n} \sum_{j=1}^{n} x_j^2 - \left( \frac{1}{n} \sum_{j=1}^{n} x_j \right)^2 .
\]

**Věta:** \( \hat{\sigma}_X^2 \) je vychýlený konzistentní odhad rozptylu (pokud původní rozdělení má rozptyl a 4. centrální moment).

**Důkaz** (pouze první části, s použitím jednoprůchodového vzorce):

\[
E\hat{\sigma}_X^2 = E\left( \frac{1}{n} \sum_{j=1}^{n} X_j^2 - \bar{X}^2 \right) = E\bar{X}^2 - E\bar{X}^2 = (EX)^2 + DX - (E\bar{X})^2 - \frac{1}{n} DX
\]

\[
= \left( 1 - \frac{1}{n} \right) DX = \frac{n-1}{n} DX \rightarrow DX \quad \text{pro} \ n \rightarrow \infty .
\]

⇒ nestranný odhad:

\[
S_X^2 := \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \bar{X}_n)^2 = \frac{n}{n-1} \hat{\sigma}_X^2 \quad \text{(výběrový rozptyl)},
\]

\[
s_x^2 := \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x}_n)^2 = \frac{n}{n-1} \hat{\sigma}_x^2 \quad \text{(realizace výběrového rozptylu)}.
\]

Alternativní značení: \( S^2, s^2 \). (Dvojka v horním indexu neznamená kvadrát.) Jednoprůchodový vzorec – praktičtější, ale numericky horší:

\[
S_X^2 = \frac{1}{n-1} \sum_{j=1}^{n} X_j^2 - \frac{n}{n-1} \bar{X}_n^2 = \frac{1}{n-1} \sum_{j=1}^{n} X_j^2 - \frac{1}{n(n-1)} \left( \sum_{j=1}^{n} X_j \right)^2 ,
\]

\[
s_x^2 = \frac{1}{n-1} \sum_{j=1}^{n} x_j^2 - \frac{n}{n-1} \bar{x}_n^2 = \frac{1}{n-1} \sum_{j=1}^{n} x_j^2 - \frac{1}{n(n-1)} \left( \sum_{j=1}^{n} x_j \right)^2 .
\]

**Věta:**

\[
ES_X^2 = DX.
\]

Výběrový rozptyl je nestranný konzistentní odhad rozptylu (pokud původní rozdělení má rozptyl a 4. centrální moment).

Rozdělení výběrového rozptylu může být podstatně složitější.

Rozdělení výběrového rozptylu pro výběr z normálního rozdělení (dle Likeš, Machek).
Problém: Ve výrazu
\[ \sum_{j=1}^{n} (X_j - \bar{X})^2 \]
jsou veličiny \( X_j - \bar{X} \) závislé.

1. Pro \( n = 2 \) a \( X_1, X_2 \) s rozdělením \( N(0, 1) \):
\[ \bar{X} = \frac{X_1 + X_2}{2}, \quad X_1 - \bar{X} = -(X_2 - \bar{X}) = \frac{X_1 - X_2}{2} \]
je rozdělení \( N(0, \frac{1}{2}) \),
\[ S_X^2 = (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 = 2 \left( \frac{X_1 - X_2}{2} \right)^2 = \left( \frac{X_1 - X_2}{\sqrt{2}} \right)^2 = U^2, \]
kde \( U = \frac{X_1 - X_2}{\sqrt{2}} \) má rozdělení \( N(0, 1) \), takže \( S_X^2 \) má rozdělení \( \chi^2(1) \).

2. Pro \( X_1, \ldots, X_n \) s rozdělením \( N(0, 1) \):
Náhodný vektor \( \mathbf{X} = (X_1, \ldots, X_n) \) s nezávislými složkami má rozdělení sféricky symetrické (kolem počátku), které se nezmění rotací ani ortonormální transformací souřadnic, \( \mathbf{U} = \mathbf{X} \mathbf{M} \), kde \( \mathbf{M} \in \mathbb{R}^{n \times n} \) je ortonormální matice. Matice
\[
M^* = \begin{bmatrix}
1 & 1 & 1 & \cdots & 1 & 1 \\
-1 & 1 & 1 & \ddots & 1 & 1 \\
0 & -2 & 1 & \ddots & 1 & 1 \\
0 & 0 & -3 & \ddots & 1 & 1 \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\
0 & 0 & 0 & \cdots & -(n-1) & 1
\end{bmatrix}
\]
má sloupce ortogonální, zbývá je normalizovat:
\[
M = \begin{bmatrix}
\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{12}} & \cdots & \frac{1}{\sqrt{n(n+1)}} & \frac{1}{\sqrt{n}} \\
-\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{12}} & \ddots & \frac{1}{\sqrt{n(n+1)}} & \frac{1}{\sqrt{n}} \\
0 & -\frac{2}{\sqrt{6}} & \frac{1}{\sqrt{12}} & \ddots & \frac{1}{\sqrt{n(n+1)}} & \frac{1}{\sqrt{n}} \\
0 & 0 & -\frac{3}{\sqrt{12}} & \ddots & \frac{1}{\sqrt{n(n+1)}} & \frac{1}{\sqrt{n}} \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\
0 & 0 & 0 & \cdots & -\frac{n-1}{\sqrt{n(n+1)}} & \frac{1}{\sqrt{n}}
\end{bmatrix}
\]
Transformací se zachovala nezávislost \( U_1, \ldots, U_n \), jejich rozdělení \( N(0, 1) \) i součet
\[ \sum_{j=1}^{n} X_j^2 = \sum_{j=1}^{n} U_j^2, \]
kde pro poslední člen platí \( U_n = \sqrt{n \bar{X}} \), \( U_n^2 = n \bar{X}^2 \),

\[
\sum_{j=1}^{n-1} U_j^2 = \sum_{j=1}^{n} X_j^2 - U_n^2 = \sum_{j=1}^{n} X_j^2 - n \bar{X}^2 = n \sigma_\bar{X}^2 = (n - 1) S_X^2 ,
\]

\[ n \sigma_\bar{X}^2 = (n - 1) S_X^2 \]

má rozdělení \( \chi^2(n - 1) \),

\[ S_X^2 \]

má rozdělení \( \frac{\chi^2(n - 1)}{n - 1} \),

\[ \sigma_\bar{X}^2 \]

má rozdělení \( \frac{\chi^2(n - 1)}{n} \).

Předchozí postup pro \( n = 2 \) byl speciálním případem, kde

\[ M = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} . \]

3. Pro \( X_1, \ldots, X_n \) s rozdělením \( N(EX, DX) \):

\( EX \) nemá vliv, \( DX \) se vše vynásobilo,

\[ S_X^2 \]

má rozdělení \( \frac{\chi^2(n - 1)}{n - 1} DX \),

\[ \sigma_\bar{X}^2 \]

má rozdělení \( \frac{\chi^2(n - 1)}{n} DX \),

\[ \frac{n \sigma_\bar{X}^2}{DX} = \frac{(n - 1) S_X^2}{DX} \]

má rozdělení \( \chi^2(n - 1) \).

Důsledky:

\[ ES_X^2 = \frac{n - 1}{n - 1} DX = DX , \]

\[ DS_X^2 = \frac{2(n - 1)}{(n - 1)^2} (DX)^2 = \frac{2}{n - 1} (DX)^2 \rightarrow 0 \quad \text{pro} \ n \rightarrow \infty . \]

Věta: Pro náhodný výběr \( \bar{X} = (X_1, \ldots, X_n) \) z normálního rozdělení je \( \bar{X} \) nejlepší nestranný odhad střední hodnoty, \( S_X^2 \) je nejlepší nestranný odhad rozptylu a statistiky \( \bar{X} \), \( S_X^2 \) jsou konzistentní a nezávislé.

![Diagram s výběrovými rozptyly pro rozsah výběru 2, 3, \ldots, 10.](image-url)
Rozdělení odhadu rozptylu pomocí vyběrového rozptylu $S^2_X$ pro rozsah výběru $3 = 2^1 + 1, 2^2 + 1, \ldots, 2^7 + 1 = 129$.

### 8.5.4 Eficience odhadů rozptylu pro normální rozdělení

1. eficience odhadu $S^2_X$ (z vlastností rozdělení $\chi^2$):

\[
E(S^2_X -DX)^2 = DS^2_X = \frac{2}{n-1} (DX)^2.
\]

2. eficience odhadu $\hat{\sigma}^2_X$ (DX je konstanta):

\[
E(\hat{\sigma}^2_X -DX)^2 = D \left( \frac{\hat{\sigma}^2_X}{n} - DX + \left( E \hat{\sigma}^2_X - DX \right) \right)^2 =
\]
\[
= D \left( \hat{\sigma}^2_X \right) + \left( \frac{1}{n} DX \right)^2 =
\]
\[
= \left( \frac{n-1}{n} \right)^2 \frac{2}{n-1} (DX)^2 + \frac{1}{n^2} (DX)^2 = \frac{2n-1}{n^2} (DX)^2,
\]

a protože

\[
\frac{2n-1}{n^2} < \frac{2}{n} < \frac{2}{n-1},
\]

je odhad $\hat{\sigma}^2_X$ více eficientní než $S^2_X$ (ktéřý je nejlepší nestranný!).

### 8.6 Odhad směrodatné odchylky

pomocí směrodatné odchylky empirického rozdělení:

\[
\hat{\sigma}_x := \sigma_{\text{Emp}(x)} = \sqrt{\frac{1}{n} \sum_{j=1}^{n} (x_j - \bar{x})^2}.
\]

Je realizací odhadu

\[
\hat{\sigma}_X := \sqrt{\frac{1}{n} \sum_{j=1}^{n} (X_j - \bar{X})^2}.
\]

Je vychýlený. Alternativa:

\[
S_X = \sqrt{S^2_X} = \sqrt{\frac{1}{n-1} \sum_{j=1}^{n} (X_j - \bar{X})^2} \quad \text{(výběrová směrodatná odchylka)},
\]

\[
s_x = \sqrt{\frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x}_n)^2} \quad \text{(realizace výběrové směrodatné odchylky)}.
\]
Alternativní značení: $S, s$

Věta:

$$ES_X \leq \sigma_X .$$

Rovnost obecně nenastává, takže to není nestranný odhad směrodatné odchylky!

Důkaz:

$$DX = E S^2_X = (ES_X)^2 + DS_X \geq (ES_X)^2 ,$$

$$\sigma_X \geq ES_X .$$

Věta: Výběrová směrodatná odchylka je vychýlený konzistentní odhad směrodatné odchylky (pokud původní rozdělení má rozptyl a 4. centrální moment).

8.7 Histogram a popis empirického rozdělení

V realizaci náhodného výběru $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ nezáleží na pořadí složek (ale záleží na jejich opakování). Úsporněji je popsán množinou (nejvýše $n$) hodnot $H := \{x_1, \ldots, x_n\}$ a jejich četnostmi $n_t, t \in H$. Data lze popsat tabulkou četnosti nebo grafiem zvaným histogram. Normováním dostaneme relativní četnosti $r_t := \frac{n_t}{n} = p_{\text{Emp}}(x)(t)$ (= hodnoty pravděpodobnostní funkce empirického rozdělení Emp($x$)), $t \in H$, kde $\sum_{t \in H} r_t = 1$.

Výpočet z četností je jednodušší (pokud se opakují stejné hodnoty):

$$E \text{Emp}(x) = \sum_{t \in H} t r_t = \frac{1}{n} \sum_{t \in H} t n_t = \frac{1}{n} \sum_{j=1}^{n} x_j = \bar{x} ,$$

$$E (\text{Emp}(x))^k = \sum_{t \in H} t^k r_t = \frac{1}{n} \sum_{t \in H} t^k n_t = \frac{1}{n} \sum_{j=1}^{n} x_j^k = m_{x^k} .$$

$$D \text{Emp}(x) = \sum_{t \in H} (t - \bar{x})^2 r_t = \frac{1}{n} \sum_{t \in H} (t - \bar{x})^2 n_t =$$

$$= \frac{1}{n} \sum_{j=1}^{n} (x_j - \bar{x})^2 = \hat{\sigma}^2 = \frac{n - 1}{n} s^2_x .$$

8.8 Odhad mediánu

pomocí mediánu empirického rozdělení, $q_{\text{Emp}}(\frac{1}{2})$ (výběrový medián). Poskytuje jinou informaci než výběrový průměr, mnohdy užitečnější (mj. robustnější – odolnější vůči vlivu vychýlených hodnot, outliers).

Výběrový medián minimalizuje kritérium

$$\ell_1(c) = E|c - \text{Emp}(x)| = \frac{1}{n} \sum_{i=1}^{n} |c - x_i| .$$

Navíc víme, jak se změní monotónní funkcí $h$: $q_{\text{Emp}}(h(x))(\frac{1}{2}) = h(q_{\text{Emp}}(x)(\frac{1}{2}))$.

Proč se používá méně než výběrový průměr:
• Vyšší výpočetní náročnost: seřazení hodnot má pracnost úměrnou \( n \ln n \), výběrový průměr \( n \).

• Vyšší paměťová náročnost: úměrná \( n \), u výběrového průměru stačí 2 registry.

• Obtížná decentralizace a paralelizace výpočtu.

Obecně lze odhadnout \( \alpha \)-kvantil \( q_X(\alpha) \) pomocí \( \alpha \)-kvantilu empirického rozdělení, \( q_{\text{Emp}(x)}(\alpha) \). Nesmíme však volit \( \alpha \) blízké 0 nebo 1, nemůžeme např. na základě výběru rozsahu 1000 odhadovat kvantil \( q_X(10^{-6}) \).

8.9 Intervalové odhady

Dosud jsme skutečnou hodnotu parametru \( \vartheta^* \) nahrazovali bodovým odhadem \( \hat{\Theta} \) (což je náhodná veličina). Nyní místo toho hledáme intervalový odhad, tzv. interval spolehlivosti \( I \), což je minimální interval takový, že

\[
P(\vartheta^* \in I) \geq 1 - \alpha,
\]

kde \( \alpha \in (0,1) \) je pravděpodobnost, že meze intervalu \( I \) budou překročeny; \( 1 - \alpha \) je koeficient spolehlivosti. Obvykle hledáme horní, resp. dolní jednostranný odhad, kdy

\[
I = (-\infty, \hat{\Theta}(1 - \alpha)), \text{ resp. } I = (\hat{\Theta}(\alpha), \infty),
\]

nebo (symetrický) oboustranný odhad,

\[
I = \langle \hat{\Theta} \left( \frac{\alpha}{2} \right), \hat{\Theta} \left( 1 - \frac{\alpha}{2} \right) \rangle.
\]

K tomu potřebujeme znát rozdělení odhadu \( \hat{\Theta} \).

8.10 Intervalové odhady parametrů normálního rozdělení \( N(\mu, \sigma^2) \)

8.10.1 Odhad střední hodnoty při známém rozptylu \( \sigma^2 \)

\textbf{Předpoklad}: Náhodná veličina \( X \) má normální rozdělení \( N(\mu, \sigma^2) \).

S pravděpodobností \( 1 - \alpha \) je

\[
q_{N(\mu, \sigma^2)} \left( \frac{\alpha}{2} \right) \leq X \leq q_{N(\mu, \sigma^2)} \left( 1 - \frac{\alpha}{2} \right),
\]

po znormování

\[
\Phi^{-1} \left( \frac{\alpha}{2} \right) = -\Phi^{-1} \left( 1 - \frac{\alpha}{2} \right) \leq \text{norm} X = \frac{X - \mu}{\sigma} \leq \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right),
\]

\[
X - \sigma \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right) \leq \mu \leq X + \sigma \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right),
\]

Pro známé \( \sigma \) dostáváme intervalový odhad neznámé střední hodnoty \( \mu \) z (realizace) náhodné veličiny \( X \).
Obvykle místě jedné realizace použijeme realizaci výběrového průměru \( \overline{X}_n \), který má rozdělení \( N \left( \mu, \frac{\sigma^2}{n} \right) \). Předchozí postup použijeme pro normalizovanou veličinu \( \frac{\sqrt{n}}{\sigma} (\overline{X}_n - \mu) \) s rozdělením \( N(0,1) \).

S pravděpodobností \( 1 - \alpha \) je

\[
\overline{X}_n - \frac{\sigma}{\sqrt{n}} \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right) \leq \mu \leq \overline{X}_n + \frac{\sigma}{\sqrt{n}} \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right),
\]

pro jednostranné odhady

\[
\mu \leq \overline{X}_n + \frac{\sigma}{\sqrt{n}} \Phi^{-1} (1 - \alpha),
\]

\[
\overline{X}_n - \frac{\sigma}{\sqrt{n}} \Phi^{-1} (1 - \alpha) \leq \mu,
\]

Dostali jsme intervalové odhady pro \( \mu \)

\[
\left( \overline{X} - \frac{\sigma}{\sqrt{n}} \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right), \overline{X} + \frac{\sigma}{\sqrt{n}} \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right) \right),
\]

\[
\left( -\infty, \overline{X} + \frac{\sigma}{\sqrt{n}} \Phi^{-1} (1 - \alpha) \right),
\]

\[
\left( \overline{X} - \frac{\sigma}{\sqrt{n}} \Phi^{-1} (1 - \alpha), \infty \right),
\]

Při výpočtu nahradíme výběrový průměr \( \overline{X}_n \) jeho realizací \( x_n \).

Díky centrální limitní větě je odhad použitelný i pro výběr z jiného než normálního rozdělení, pokud má nenulový rozptyl a rozsah výběru je velký.

8.10.2 Odhad střední hodnoty při neznámém rozptylu

\( \mu \) odhadneme výběrovým průměrem \( \overline{X}_n \) s rozdělením \( N \left( \mu, \frac{\sigma^2}{n} \right) \),

\( \sigma^2 \) odhadneme výběrovým rozptylem \( S_X^2 \) s rozdělením \( \frac{\sigma^2}{n-1} \chi^2(n-1) \).

Testujeme analogicky náhodnou veličinu \( \frac{\sqrt{n}}{S_X} (\overline{X}_n - \mu) \), její rozdělení však není normální, ačkoli \( \overline{X}_n, S_X \) jsou nezávislé.

8.10.3 Studentovo t-rozdělení (autor: Gossett)

s \( \eta \) stupně volnosti je rozdělení náhodné veličiny \( \frac{U}{\sqrt{V/\eta}} \),

kde \( U \) má rozdělení \( N(0,1) \),

\( V \) má rozdělení \( \chi^2(\eta) \),

\( U, V \) jsou nezávislé.

Značení: \( t(\eta) \).

Hustota:

\[
f_{t(\eta)}(x) = c(\eta) \left( 1 + \frac{x^2}{\eta} \right)^{-\frac{1+\eta}{2}}, \quad c(\eta) = \frac{\Gamma \left( \frac{1+\eta}{2} \right)}{\sqrt{\eta \pi} \Gamma \left( \frac{\eta}{2} \right)}.
\]
Symetrie kolem nuly \( \Rightarrow q_{t(\eta)}(1 - \alpha) = -q_{t(\eta)}(\alpha) \).

\( t(1) \) je Cauchyho rozdělení, které nemá střední hodnotu,

\[
f_{t(1)}(x) = \frac{1}{\pi} \frac{1}{1 + x^2}.
\]

Pro velký počet stupňů volnosti se nahrazuje normovaným normálním rozdělením.

Hustota normovaného normálního rozdělení a Studentova rozdělení s 5 stupni volnosti.

Hustota normovaného normálního rozdělení a \textit{normovaného} Studentova rozdělení s 5 stupni volnosti.

### 8.10.4 Odhad střední hodnoty při neznámém rozptylu 2

V našem případě:

\[
U = \frac{\sqrt{n}}{\sigma} (\bar{X}_n - \mu) \text{ má } N(0, 1),
\]

\[
V = \frac{(n - 1) S^2_{\bar{X}}}{\sigma^2} \text{ má } \chi^2(n - 1), \ \eta = n - 1,
\]

\[
\frac{U}{\sqrt{\frac{V}{\eta}}} = \frac{\sqrt{n}}{S_{\bar{X}}} (\bar{X}_n - \mu) \text{ má } t(n - 1).
\]

Z toho vyplývají intervalové odhady

\[
\left\langle \bar{X}_n - \frac{S_{\bar{X}}}{\sqrt{n}} q_{t(n-1)}(1 - \frac{\alpha}{2}), \bar{X}_n + \frac{S_{\bar{X}}}{\sqrt{n}} q_{t(n-1)}(1 - \frac{\alpha}{2}) \right\rangle,
\]

\[
\left\langle -\infty, \bar{X}_n + \frac{S_{\bar{X}}}{\sqrt{n}} q_{t(n-1)}(1 - \alpha) \right\rangle,
\]

\[
\left\langle \bar{X}_n - \frac{S_{\bar{X}}}{\sqrt{n}} q_{t(n-1)}(1 - \alpha), \infty \right\rangle.
\]

Při výpočtu nahradíme výběrový průměr \( \bar{X}_n \) jeho realizací \( \tilde{x}_n \) a výběrovou směrodatnou odchylku \( S_{\bar{X}} \) její realizací \( s_x \).
Díky centrální limitní větě je odhad použitelný i pro výběr z jiného než normálního rozdělení, pokud má nenulový rozptyl a rozsah výběru je velký (pak můžeme místo Studentova rozdělení použít normální).

8.10.5 Odhad rozptylu

\( \sigma^2 \) odhadneme výběrovým rozptylem \( S^2_X \) s rozdělením \( \frac{\sigma^2}{n-1} \chi^2(n-1) \);

\( \frac{(n-1)S^2_X}{\sigma^2} \) má rozdělení \( \chi^2(n-1) \);

S pravděpodobností \( 1 - \alpha \) je

\[
q_{\chi^2(n-1)}\left(\frac{\alpha}{2}\right) \leq \frac{(n-1)S^2_X}{\sigma^2} \leq q_{\chi^2(n-1)}\left(1 - \frac{\alpha}{2}\right),
\]

\[
\frac{(n-1)S^2_X}{q_{\chi^2(n-1)}\left(1 - \frac{\alpha}{2}\right)} \leq \sigma^2 \leq \frac{(n-1)S^2_X}{q_{\chi^2(n-1)}\left(\frac{\alpha}{2}\right)},
\]

pro jednostranné odhady

\[
\sigma^2 \leq \frac{(n-1)S^2_X}{q_{\chi^2(n-1)}\left(\alpha\right)},
\]

\[
\frac{(n-1)S^2_X}{q_{\chi^2(n-1)}(1 - \alpha)} \leq \sigma^2.
\]

Dostali jsme intervalové odhady pro \( \sigma^2 \)

\[
\left\langle \frac{(n-1)S^2_X}{q_{\chi^2(n-1)}\left(1 - \frac{\alpha}{2}\right)}, \frac{(n-1)S^2_X}{q_{\chi^2(n-1)}\left(\frac{\alpha}{2}\right)} \right\rangle,
\]

\[
\left\langle -\infty, \frac{(n-1)S^2_X}{q_{\chi^2(n-1)}(\alpha)} \right\rangle,
\]

\[
\left\langle \frac{(n-1)S^2_X}{q_{\chi^2(n-1)}(1 - \alpha)}, \infty \right\rangle.
\]

Při výpočtu nahradíme výběrový rozptyl \( S^2_X \) jeho realizací \( s^2_x \).

Díky centrální limitní větě je odhad použitelný i pro výběr z jiného než normálního rozdělení, pokud má nenulový rozptyl a rozsah výběru je velký (pak můžeme místo rozdělení \( \chi^2(n-1) \) použít normální \( N(n-1, 2(n-1)) \)).

8.10.6 Intervalové odhady spojitých rozdělení, která nejsou normální

převádíme obvykle na normální rozdělení nelineární neklesající transformací

\[
h(t) = \Phi^{-1}(F_X(t))
\]

\( F_X(X) \) má rovnoměrné rozdělení na \( (0, 1) \).

Použijeme intervalový odhad pro normální rozdělení a transformujeme jej zpět podle vzorce

\[
h^{-1}(u) = q_X^{-1}(\Phi(u)).
\]
Někdy se transformuje na jiné rozdělení, např. Studentovo s vhodným počtem stupňů volnosti.

**Poznámka:** Je možné najít platné intervalové odhady, i když neexistuje střední hodnota nebo rozptyl.

### 8.11 Obecné odhady parametrů

**Motivační příklad (volební předpověď):**
Máme odhadnout výsledky voleb (kompletní).
Dosavadní postupy nám dovolovaly pouze odhad (včetně intervalového) výsledků jedné strany.
Hledanými parametry mohou být výsledky všech zúčastněných stran, vyjádřené vektorem čísel z ⟨0, 1⟩, jejichž součet je 1.

**Motivační příklad (směs normálních rozdělení):**
Spojité rozdělení, jehož hustota má více maxim, aproximujeme směsí normálních,

\[
\text{Mix}_{(c_1, \ldots, c_m)} \left( \text{N}(\mu_1, \sigma_1^2), \ldots, \text{N}(\mu_m, \sigma_m^2) \right),
\]

s hustotou

\[
\sum_{i=1}^{m} c_i \frac{1}{\sigma_i \sqrt{2\pi}} \exp \left( -\frac{(t - \mu_i)^2}{2 \sigma_i^2} \right).
\]

Vektor parametrů je \((c_1, \ldots, c_m; \mu_1, \ldots, \mu_m; \sigma_1, \ldots, \sigma_m)\) s omezujícími podmínkami

\[
\begin{align*}
\sigma_i &\geq 0, & i &= 1, \ldots, m, \\
0 &\leq c_i \leq 1, & i &= 1, \ldots, m, \\
\sum_{i=1}^{m} c_i &= 1.
\end{align*}
\]

**Formulace úlohy**
Rozdělení náhodné veličiny \(X\) závisí na vektoru parametrů \(\vartheta = (\vartheta_1, \ldots, \vartheta_i) \in \Pi\), kde \(\Pi \subseteq \mathbb{R}^i\) je **parametrický prostor**, tj. množina všech přípustných hodnot parametrů; pravděpodobnostní funkci značíme \(p_X(t; \vartheta) = p_X(t; \vartheta_1, \ldots, \vartheta_i)\) atd.
Hledáme odhad \(\hat{\Theta} = (\hat{\Theta}_1, \ldots, \hat{\Theta}_i)\), resp. realizaci odhadu \(\hat{\vartheta} = (\hat{\vartheta}_1, \ldots, \hat{\vartheta}_i)\) pomocí realizace \(x = (x_1, \ldots, x_n)\).

#### 8.11.1 Metoda momentů

( angl. moment matching )
Pro \(k = 1, 2, \ldots\) lze \(k\)-tý obecný moment vypočítat z modelu jako funkci \(\vartheta\),

\[
\text{EX}^k(\vartheta) = \text{EX}^k(\vartheta_1, \ldots, \vartheta_i)
\]

a současně odhadnout pomocí výběrového \(k\)-tého obecného momentu

\[
m_x^k = \frac{1}{n} \sum_{j=1}^{n} x_j^k.
\]
Metoda momentů doporučuje realizaci odhadu \( \hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_i) \) takovou, že

\[
EX^k(\hat{\theta}_1, \ldots, \hat{\theta}_i) = \frac{1}{n} \sum_{j=1}^{n} x_j^k.
\]

K jednoznačnému určení \( i \) proměnných obvykle použijeme (prvních) \( i \) rovnic pro \( k = 1, 2, \ldots, i \).

**Použitelnost metody momentů**

**Možné problémy:**

1. Potřebujeme existenci použitých momentů (zatímco jejich „odhady“ jsou definovány vždy).

2. Nelze použít pro nenumerická data (např. volební předpověď), pokud je nelze smysluplně očíslovat.


4. Snažíme se použít tolik rovnic (obyčejně pro nejmenší možná \( k \)), abychom dostali konečný nenulový počet řešení.

5. Může být více než jedno řešení (např. soustavy kvadratických rovnic).


7. Soustava může být špatně podmíněná (typicky pro velký počet parametrů).

8. Můžeme dospět k řešení, které nesplňuje předpoklady, \( \hat{\theta} \notin \Pi \) (např. parametry nemohou být libovolná čísla) \( \Rightarrow \) vždy kontrolujte řešení!

9. Všem rovnicím je přikládána stejná důležitost, což bývá nežádoucí (typicky pro velký počet parametrů), přitom i jejich fyzikální rozměry bývají různé.

**Výhody:**

1. Shoda momentů zajišťuje „podobné“ rozdělení modelu i dat.

2. Lze použít pro diskrétní, spojité i smíšené rozdělení.
8.11.2 Metoda maximální věrohodnosti \((\text{angl. likelihood})\)

**Pro diskrétní rozdělení**

Pravděpodobnost realizace je funkce \( L: \Pi \rightarrow \langle 0, 1 \rangle, \Pi \subseteq \mathbb{R}^i \), parametrů \( \vartheta = (\vartheta_1, \ldots, \vartheta_i) \), zvaná **věrohodnost realizace diskrétního rozdělení**, 

\[
L(\vartheta) := p_X(x; \vartheta) = P(X_1 = x_1 \land \ldots \land X_n = x_n; \vartheta) = \prod_{j=1}^{n} P(X_j = x_j; \vartheta) = \prod_{j=1}^{n} p_X(x_j; \vartheta).
\]

Hledáme takové hodnoty \( \widehat{\vartheta} = (\widehat{\vartheta}_1, \ldots, \widehat{\vartheta}_i) \), které maximalizují věrohodnost, resp. její logaritmus (**angl. log-likelihood**), 

\[
\ell(\vartheta) := \ln L(\vartheta) = \sum_{j=1}^{n} \ln p_X(x_j; \vartheta).
\]

(Nutno vyloučit případ \( p_X(x_j; \vartheta) = 0 \), který však nevede na maximum.)

**Poznámka:** Odhad na základě maxima věrohodnosti odpovídá Bayesovskému odhadu ve speciálním případě, kdy všechny hodnoty parametrů mají stejnou apriorní pravděpodobnost (resp. hustotu pravděpodobnosti). Používá se, pokud apriorní pravděpodobnosti parametrů neznáme.

**Poznámka:** Často určujeme věrohodnost až na násobek konstantou, kterou by bylo obtížné určit. Na výsledek optimalizace to nemá vliv.

**Pro spojité rozdělení**

Každá realizace má nulovou pravděpodobnost, proto místo ní použijeme hustotu pravděpodobnosti, což ale vede na zcela **jiný pojem**

\[
\Lambda(\vartheta) := f_X(x; \vartheta) = \prod_{j=1}^{n} f_X(x_j; \vartheta).
\]

Nicméně i tato funkce \( \Lambda: \Pi \rightarrow \langle 0, \infty \rangle, \Pi \subseteq \mathbb{R}^i \), se nazývá **věrohodnost realizace spojitého rozdělení**. Pro korektní definici potřebujeme spojitou hustotu (alespoň na oboru hodnot, jichž náhodná veličina nabývá); taková hustota je nejvýše jedna.

\[
\lambda(\vartheta) := \ln \Lambda(\vartheta) = \sum_{j=1}^{n} \ln f_X(x_j; \vartheta).
\]

(Nutno vyloučit případ \( f_X(x_j; \vartheta) = 0 \), který však nevede na maximum.)

**Pro smíšené rozdělení**

není věrohodnost definována!

Použitelnost metody maximální věrohodnosti

Možné problémy:
1. Může být více než jedno řešení.  
   (Může se stát, že různé hodnoty parametrů popisují totéž rozdělení – vadí to?)

2. Řešení nemusí existovat (pokud věrohodnostní funkce je nespojitá nebo parametrický prostor neuzavřený).

3. Může být obtížné řešení nalézt. (Uvíznutí v lokálním extrému; parametrický prostor nemusí být spojitý.)

4. Hodnoty věrohodnosti mohou být velmi malé.

5. Fyzikální rozměr věrohodnosti může být ve spojitém případě kuriózní.

6. **Nelze použít pro smíšené rozdělení!**

**Výhody:**

1. Hledání optima je o něco snazší než řešení soustavy rovnic – vždy nějakou aproximaci najdeme.

2. Různým datům je dán společný (srovnatelný) význam.

3. Lze použít i na nenumerická data.

4. Někdy vybíráme jen z konečného počtu modelů, pak metoda maximální věrohodnosti je použitelná, zatímco postupy založené na řešení rovnic nikoli.

8.11.3 **Příklady na odhady parametrů**

**Cvičení.** Odhadněte diskrétní rozdělení z četností hodnot v realizaci:

<table>
<thead>
<tr>
<th>výsledek s</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>∑</th>
</tr>
</thead>
<tbody>
<tr>
<td>pravděpodobnost</td>
<td>a</td>
<td>b</td>
<td>c</td>
<td>1</td>
</tr>
<tr>
<td>četnost n_s</td>
<td>10</td>
<td>12</td>
<td>10</td>
<td>32</td>
</tr>
</tbody>
</table>

**Řešení.** **Metoda momentů:**

\[
\begin{align*}
a + b + c &= 1, \\
E[X] &= \sum_s s p_s = a + 2b + 3c = \frac{1}{n} \sum_s s n_s = \frac{10 + 2 \cdot 12 + 3 \cdot 10}{32} = 2, \\
E[X^2] &= \sum_s s^2 p_s = a + 4b + 9c = \frac{1}{n} \sum_s s^2 n_s = \frac{10 + 4 \cdot 12 + 9 \cdot 10}{32} = \frac{37}{8},
\end{align*}
\]

\[
\begin{align*}
a &= c = \frac{5}{16}, \quad b = \frac{6}{16}.
\end{align*}
\]
Metoda maximální věrohodnosti:

\[ L(a, b) = a^{10} \cdot b^{12} \cdot (1 - a - b)^{10}, \]

\[ \ell(b) = \ln \ell(a, b) = 10 \ln a + 12 \ln b + 10 \ln (1 - a - b), \]

\[ 0 = \frac{\partial}{\partial a} \ell(a, b) = \frac{10}{a} - \frac{10}{1 - a - b} = \frac{10}{a} - \frac{10}{c}, \]

\[ 0 = \frac{\partial}{\partial b} \ell(a, b) = \frac{12}{b} - \frac{10}{1 - a - b} = \frac{12}{b} - \frac{10}{c}, \]

\[ a = c = \frac{5}{6} b, \]

a z jednotkového součtu pravděpodobností opět

\[ a = c = \frac{5}{16}, \quad b = \frac{6}{16}. \]

Obě metody vedly na empirické rozdělení. Tento výsledek není náhodný:

Věta. Pokud na diskrétní rozdělení nejsou kladeny žádné omezující podmínky, pak empirické rozdělení je jeho odhadem podle metody momentů i maximálně věrohodným odhadem.

Důkaz. Označme \( u_1, \ldots, u_i \) (\( i \leq n \)) všechny různé hodnoty, které se vyskytly v realizaci \( x \), \( n_s \) četnost a \( r_s = n_s/n \) relativní četnost hodnoty \( u_s \). Máme odhadnout pravděpodobnosti

\[ q_s = p_X (u_s), \quad s = 1, \ldots, i, \quad \sum_{s=1}^{i} q_s = 1. \]

Připomeňme, že empirické rozdělení \( \text{Emp}(x) \) nabývá hodnot \( u_1, \ldots, u_i \) s pravděpodobnostmi po řadě \( r_1, \ldots, r_i \).

Metoda maximální věrohodnosti:

\[ L(q_1, \ldots, q_i) = \prod_{j=1}^{n} p_X (x_j) = \prod_{s=1}^{i} (p_X (u_s))^{n_s} = \prod_{s=1}^{i} q_s^{n_s}, \]

\[ \ell(q_1, \ldots, q_i) = \ln L(q_1, \ldots, q_i) = \sum_{s=1}^{i} n_s \ln q_s. \]

Použijeme metodu Lagrangeových multiplikátorů, tj. přičteme \( c \)-násobek podmínky jednotkového součtu neznámých a hledáme globální maximum funkce

\[ h(c, q_1, \ldots, q_i) = \sum_{s=1}^{i} n_s \ln q_s + c \left(1 - \sum_{s=1}^{i} q_s\right), \]

\[ 0 = \frac{\partial}{\partial q_s} h(c, q_1, \ldots, q_i) = \frac{n_s}{q_s} - c. \]
Hodnota \( \frac{n_s}{q_s} = c \) je nezávislá na \( s \in \{1, \ldots, i\} \). Určíme ji z podmínky 1 = \( \sum_{s=1}^{i} q_s = \frac{1}{c} \sum_{s=1}^{i} n_s = \frac{n}{c} \)

\[
\frac{n_s}{q_s} = c = n, \\
q_s = \frac{n_s}{n} = r_s
\]

(empirické rozdělení).

**Metoda momentů:**

\[
EX^k = \sum_{s=1}^{i} q_s u_s^k = \frac{1}{n} \sum_{j=1}^{n} x_j^k = \frac{1}{n} \sum_{s=1}^{i} n_s u_s^k = \sum_{s=1}^{i} r_s u_s^k.
\]

Řešením je \( q_s = r_s \) (empirické rozdělení). Je to jediné řešení, neboť matice soustavy

\[
\begin{bmatrix}
  u_1^1 & u_1^2 & \cdots & u_1^i \\
  u_2^1 & u_2^2 & \cdots & u_2^i \\
  \vdots & \vdots & \ddots & \vdots \\
  u_i^1 & u_i^2 & \cdots & u_i^i
\end{bmatrix}
\]

(tzv. Vandermondova matice) je regulární, právě když čísla \( u_1, \ldots, u_i \) jsou navzájem různá. \( \Box \)

**Cvičení.** Odhadněte meze \( a, b \) spojitého rovnoměrného rozdělení z realizace

1. \((3, 7, 5, 8, 1)\),  
2. \((0, 0, 0, 0, 5)\),

**Řešení. Metoda momentů:**

\[
EX = \frac{a + b}{2} = \frac{1}{n} \sum_{s=1}^{n} x_s, \\
EX^2 = (EX)^2 + DX = \left( \frac{a + b}{2} \right)^2 + \frac{(b - a)^2}{12 DX} = \frac{a^2 + ab + b^2}{3} = \frac{1}{n} \sum_{s=1}^{n} x_s^2.
\]

1. **Soustava**

\[
EX = \frac{a + b}{2} = \frac{24}{5}, \\
EX^2 = \frac{a^2 + ab + b^2}{3} = \frac{148}{5},
\]

má 2 řešení

\[
a \approx 0.36, \quad b \approx 9.24, \\
a \approx 9.24, \quad b \approx 0.36
\]
První řešení je jediné správné.

2. Soustava

\[ \begin{align*}
EX &= \frac{a + b}{2} = 1, \\
EX^2 &= \frac{a^2 + ab + b^2}{3} = 5,
\end{align*} \]

má 2 řešení

\[ a = -2.5, \quad b = 4.5, \]
\[ a = 4.5, \quad b = -2.5. \]

Žádné není správné, neboť \( 5 \notin (a, b). \)

**Metoda maximální věrohodnosti:** Spojitá hustota je konstantní \( \frac{1}{b-a} \) na intervalu \( (a, b) \).

\[ L(a, b) = \prod_{j=1}^{n} \frac{1}{b-a} = \left( \frac{1}{b-a} \right)^n, \]

pokud \( x_j \in (a, b) \) pro všechna \( j \); jinak je nulová. Věrohodnost je maximální, pokud \( b - a \) je minimální, tj.

\[ a = \min_j x_j, \quad b = \max_j x_j. \]

1. \( a = 1, \quad b = 8. \)
2. \( a = 0, \quad b = 5. \)

**Příklad.** Z realizace náhodného výběru \( x = (x_1, \ldots, x_n) \) z normálního rozdělení \( N(\mu, \sigma^2) \) odhadněte parametry \( \mu \) a \( r = \sigma^2 \).

**Řešení:** **Metoda momentů:** Použijeme první dva obecné momenty,

\[ EX = \mu, \quad EX^2 = (EX)^2 + DX = \mu^2 + \sigma^2 = \mu^2 + r. \]

Pro odhady \( \hat{\mu}, \hat{\sigma} \) máme soustavu rovnic

\[ \hat{\mu} = \frac{1}{n} \sum_{j=1}^{n} x_j, \]
\[ \hat{\mu}^2 + \hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^{n} x_j^2. \]

**Řešení:**

\[ \hat{\mu} = \bar{x}, \]
\[ \hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^{n} x_j^2 - \hat{\mu}^2 = \frac{1}{n} \sum_{j=1}^{n} x_j^2 - \bar{x}^2 = \frac{\sigma_X^2}{n} = D\text{Emp}(x). \]
Metoda maximální věrohodnosti:

\[ \Lambda(\mu, r) = \prod_{j=1}^{n} f_{N(\mu, r)}(x_j) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi r}} \exp \left( -\frac{(x_j - \mu)^2}{2r} \right), \]

\[ \lambda(\mu, r) = \ln \Lambda(\mu, r) = -\frac{1}{2r} \sum_{j=1}^{n} (x_j - \mu)^2 - \frac{n}{2} \ln r - \frac{n}{2} \ln 2\pi, \]

\[ 0 = \frac{\partial}{\partial \mu} \lambda(\mu, r) = \frac{1}{r} \sum_{j=1}^{n} (x_j - \hat{\mu}) = \frac{1}{r} \left( \sum_{j=1}^{n} x_j - n \hat{\mu} \right) = \frac{n}{r} (\bar{x} - \hat{\mu}), \]

\[ \Rightarrow \hat{\mu} = \bar{x}, \]

\[ 0 = \frac{\partial}{\partial r} \lambda(\mu, r) = \frac{1}{2r^2} \sum_{j=1}^{n} (x_j - \hat{\mu})^2 - \frac{n}{2r^2} = \frac{n}{2r^2} \left( \frac{1}{n} \sum_{j=1}^{n} (x_j - \bar{x})^2 - \hat{r} \right), \]

\[ \Rightarrow \hat{r} = \frac{1}{n} \sum_{j=1}^{n} (x_j - \bar{x})^2 = \frac{\sigma^2}{\hat{\sigma}^2} = D_{\text{Emp}}(\bar{x}). \]

Motivační příklad (směs normálních rozdělení – pokračování):

Spojité rozdělení, jehož hustota má více maxim, aproximujeme směsí normálních,

\[ \text{Mix}_{c_1, \ldots, c_m} \left( N(\mu_1, \sigma_1^2), \ldots, N(\mu_m, \sigma_m^2) \right), \]

s hustotou

\[ \sum_{i=1}^{m} c_i \frac{1}{\sigma_i \sqrt{2\pi}} \exp \left( -\frac{(t - \mu_i)^2}{2\sigma_i^2} \right). \]

Vektor parametrů je \((c_1, \ldots, c_m; \mu_1, \ldots, \mu_m; \sigma_1, \ldots, \sigma_m)\) s omezujícími podmínkami

\[ \sigma_i \geq 0, \quad i = 1, \ldots, m, \]

\[ 0 \leq c_i \leq 1, \quad i = 1, \ldots, m, \]

\[ \sum_{i=1}^{m} c_i = 1. \]
Pokus o řešení:

\[ f_X(t) = \sum_{i=1}^{m} c_i f_{N(\mu_i,\sigma^2)}(t) = \sum_{i=1}^{m} c_i \frac{1}{\sqrt{2\pi} \sigma} \exp \left( -\frac{(t - \mu_i)^2}{2\sigma^2} \right), \]

\[ \Lambda(\mu, c) = \prod_{j=1}^{n} f_X(x_j) = \prod_{j=1}^{n} \sum_{i=1}^{m} c_i f_{N(\mu_i,\sigma^2)}(x_j) = \]

\[ = \prod_{j=1}^{n} \left( \frac{1}{\sqrt{2\pi} \sigma} \sum_{i=1}^{m} c_i \exp \left( -\frac{(x_j - \mu_i)^2}{2\sigma^2} \right) \right) = \]

\[ = \left( \frac{1}{\sqrt{2\pi} \sigma} \right)^n \prod_{j=1}^{n} \sum_{i=1}^{m} c_i \exp \left( -\frac{(x_j - \mu_i)^2}{2\sigma^2} \right), \]

\[ \lambda(\mu, c) = \sum_{j=1}^{n} \ln \sum_{i=1}^{m} c_i f_{N(\mu_i,\sigma^2)}(x_j) = \]

\[ = -n \ln \left( \sqrt{2\pi} \sigma \right) + \sum_{j=1}^{n} \ln \sum_{i=1}^{m} c_i \exp \left( -\frac{(x_j - \mu_i)^2}{2\sigma^2} \right). \]

Věrohodnost se těžko maximalizuje přímo, používá se iteráční metoda:

**EM algoritmus**


**Stupeň příslušnosti** \( x_j \) ke \( i \)-té složce směsi popíšeme koeficientem \( \alpha_{j,i} \in (0,1) \), přičemž

\[ \sum_{i=1}^{m} \alpha_{j,i} = 1, \quad \sum_{j=1}^{n} \alpha_{j,i} > 0. \]

1. Zvolíme náhodně různé střední hodnoty složek směsi \( \mu_i \) a nemulové koeficienty \( c_i, \ i = 1, ..., m \), splňující \( \sum_{i=1}^{m} c_i = 1 \).

E. Stanovíme stupně příslušnosti

\[ \alpha_{j,i} := \frac{c_i f_{N(\mu_i,\sigma^2)}(x_j)}{\sum_{i'=1}^{m} c_{i'} f_{N(\mu_{i'},\sigma^2)}(x_j)} = \frac{c_i \exp \left( -\frac{(x_j - \mu_i)^2}{2\sigma^2} \right)}{\sum_{i'=1}^{m} \left( c_{i'} \exp \left( -\frac{(x_j - \mu_{i'})^2}{2\sigma^2} \right) \right)}. \]

(jmenovatel je normalizační faktor).

M. Aktualizujeme koeficienty složek směsi

\[ c_i := \frac{\sum_{j=1}^{n} \alpha_{j,i}}{\sum_{i'=1}^{m} \sum_{j=1}^{n} \alpha_{j,i'}} = \frac{1}{n} \sum_{j=1}^{n} \alpha_{j,i} \]
a střední hodnoty složek jako těžiště hodnot realizace vážených stupni příslušnosti,

\[ \mu_i := \frac{\sum_{j=1}^{n} \alpha_{j,i} x_j}{\sum_{j=1}^{n} \alpha_{j,i}} = \frac{\sum_{j=1}^{n} \alpha_{j,i} x_j}{n c_i} . \]

2. Opakujeme EM, dokud to přináší podstatnou změnu výsledků.
Podobně lze postupovat i pro neznámé rozptyly jednotlivých složek směsi.

Věta: V průběhu EM algoritmu věrohodnost neklesá.
Toto je jen velmi speciální ukázka EM algoritmu; lze jej snadno rozšířit na více dimenzí a jiné typy směsí.
Použití pro parametry směsi rozdělení je typické, ne však jediné možné.

Problém: Uváznutí v lokálním extrému.
EM algoritmus rozšiřuje možnosti použití metody maximální věrohodnosti.

**Použití empirického rozdělení \( \text{Emp}(x) \) v odhadech**

<table>
<thead>
<tr>
<th>veličina</th>
<th>realizace odhadu</th>
<th>nestranný</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{EX} )</td>
<td>( \text{E Emp}(x) = \frac{1}{n} \sum x_i = \bar{x} )</td>
<td>+</td>
</tr>
<tr>
<td>( \text{EX}^k )</td>
<td>( \text{E}(\text{Emp}(x)^k) = \frac{1}{n} \sum x_i^k = m_k )</td>
<td>-</td>
</tr>
<tr>
<td>( \text{DX} )</td>
<td>( \text{D Emp}(x) = \frac{1}{n} \sum (x_i - \bar{x})^2 )</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>( \frac{n}{n-1} \text{D Emp}(x) = \frac{1}{n-1} \sum (x_i - \bar{x})^2 = s_x^2 )</td>
<td>-</td>
</tr>
<tr>
<td>( \sigma X )</td>
<td>( \sigma_{\text{Emp}}(x) = \sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2} )</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>( \sqrt{\frac{n}{n-1}} \sigma_{\text{Emp}}(x) = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2} = s_x )</td>
<td>-</td>
</tr>
<tr>
<td>( \text{qX} \left( \frac{1}{2} \right) )</td>
<td>( \text{qEmp}(x) \left( \frac{1}{2} \right) )</td>
<td>?</td>
</tr>
<tr>
<td>( \text{qX} \left( \alpha \right) )</td>
<td>( \text{qEmp}(x) \left( \alpha \right) )</td>
<td>?</td>
</tr>
<tr>
<td>( \text{pX} ) bez omezení</td>
<td>( \text{pEmp}(x) )</td>
<td>+</td>
</tr>
<tr>
<td>( \text{pX} ) s omezením</td>
<td>( \text{pEmp}(x) )</td>
<td>?</td>
</tr>
<tr>
<td>( \text{fX} )</td>
<td>( \text{?} )</td>
<td>?</td>
</tr>
</tbody>
</table>

? = záleží na okolnostech

9 Testování hypotéz

9.1 Základní pojmy a principy testování hypotéz

*(doporučená literatura: [Jaroš a kol.]*)
Máme posoudit hypotézu o hodnotě nějakého parametru rozdělení \( \vartheta \) (pomocí kritérií čili testovací statistiky \( T \), resp. její realizace \( t \)).

Předpoklad: Parametr \( \vartheta \) nabývá pouze 2 hodnot, 0 pro „normální“ populaci, 1 pro „anomální“ prvky. O prvky máme rozhodnout, ke které skupině patří (tj. odhadnout \( \vartheta \)). K tomu použijeme
testovací statistiku \( T \) (resp. její realizaci \( t \)). Ta závisí na \( \theta \). Předpokládejme, že obě skupiny mají známá rozdělení statistiky \( T \), která pro anomální skupinu nabývá „větší“ hodnot. (Některé hodnoty statistiky \( T \) se mohou vyskytnout v obou skupinách, takže klasifikace nemůže být bezchybná.) Zvolíme práh \( \kappa \in \mathbb{R} \) a prvek klasifikujeme následovně:

\[
\begin{align*}
\text{pro } T \leq \kappa & \quad \text{normální}, \\
\text{pro } T > \kappa & \quad \text{anomální}.
\end{align*}
\]

**Příklad:** Máme zastavit používání léku pro podezření z nežádoucích účinků?

**Nulová hypotéza** \( H_0 \): Výrobce je nevinen, riziko se nezvyšuje.

**Alternativní hypotéza** \( H_1 \): Výrobce je vinen, riziko se zvyšuje.

**Chyba 1. druhu** (obviníme nevinného): Zamítáme nulovou hypotézu, která platí. Normální je klasifikován jako anomální s pravděpodobností \( \alpha(\kappa) \) (nerostoucí funkce \( \kappa \)).

**Chyba 2. druhu** (osvobodíme vinného): Nezamítáme nulovou hypotézu, která neplatí. Anomální je klasifikován jako normální s pravděpodobností \( \beta(\kappa) \) (neklesající funkce \( \kappa \)).

**ROC křivka** (angl. ROC curve, receiver operating characteristic) vyjadřuje závislost pravděpodobnosti chyby prvního druhu \( \alpha \) (vodorovně) a síly testu \( 1 - \beta \) (svisle), parametrem křivky je kritická hodnota \( \kappa \). Volbou kritické hodnoty se chceme co nejvíc přibližit bodu \((0, 1)\), tj. bezchybné klasifikaci. Nicméně vybereme bod, v němž se pravděpodobnost chyby prvního druhu rovná zvolenému číslu \( \alpha \) (tj. s danou vodorovnou souřadnicí).

![Typický průběh ROC křivky](image)

Možná kritéria pro volbu prahu \( \kappa \):

- \( \alpha(\kappa) = \beta(\kappa) \),
- \( \min_\kappa (\alpha(\kappa) + \beta(\kappa)) \),
- \( \min_\kappa e(\alpha(\kappa), \beta(\kappa)) \), např. \( \min_\kappa (a \alpha(\kappa) + b \beta(\kappa)) \), tj. minimalizace výplatná funkce,
- \( \alpha(\kappa) = \text{předem zvolená malá hodnota} \).

Většinou se používá poslední možnost, a to z důvodů

- technických (snazší úloha),
- nepotřebujeme znát rozdělení anomální skupiny,
obvykle máme více než dvě možné hodnoty parametru, což situaci komplikuje.

Volbou přísnosti kritéria snižujeme riziko jedné chyby na úkor zvýšení rizika druhé chyby.

Dohodnuté východisko: **Kritickou hodnotu** testu \( \kappa \) stanovíme tak, aby chyba 1. druhu nastávala s danou pravděpodobností \( \alpha \) zvanou **hladina významnosti** (nebo s menší pravděpodobností, nelze-li dosáhnout rovnost).

Podle tradice v oboru se nejčastěji užívají hodnoty 1\% nebo 5\% (vždy \( \alpha \ll \frac{1}{2} \)).

Hodnoty kritéria, která přesahují kritickou hodnotu (odpovídají výsledkům málo pravděpodobným při platnosti nulové hypotézy) považujeme za **statisticky významné** a nulovou hypotézu zamítáme.

V opačném případě **nulovou hypotézu nezamítáme**, ale ani **nepotvrzuje**, neboť tím bychom se mohli dopustit chyby 2. druhu s blíže neurčenou pravděpodobností \( \beta \).

**Slovníček pojmů** (pro porozumění jiným textům, zde se téměř nepoužijí)

<table>
<thead>
<tr>
<th>skutečnost</th>
<th>anomální</th>
<th>normální</th>
<th>celkem</th>
</tr>
</thead>
<tbody>
<tr>
<td>pozitivní</td>
<td>( TP )</td>
<td>( FP )</td>
<td>( P' )</td>
</tr>
<tr>
<td>negativní</td>
<td>( FN )</td>
<td>( TN )</td>
<td>( N' )</td>
</tr>
<tr>
<td>celkem</td>
<td>( P )</td>
<td>( N )</td>
<td></td>
</tr>
</tbody>
</table>

(Položky v tabulce mohou být pravděpodobnosti empirického nebo skutečného rozdělení nebo empirické četnosti.)

\( TP \) skutečné pozitivní (*true positive*)

\( FP \) falešně pozitivní, chyba 1. druhu (*false positive, type I error*)

\( TN \) skutečně negativní (*true negative*)

\( FN \) falešně negativní, chyba 2. druhu (*false negative, type II error*)

\[ \alpha = \frac{FP}{N} = \frac{FP}{TN+FP} = \text{pravděpodobnost chyby 1. druhu} \]

\[ \beta = \frac{FN}{P} = \frac{FN}{TP+FN} = \text{pravděpodobnost chyby 2. druhu} \]
\[
\frac{TP}{TP + FN} = \frac{TP}{P} = 1 - \beta
\] senzitivita, síla, míra skutečně pozitivních

\[
\frac{TN}{TN + FP} = \frac{TN}{N} = 1 - \alpha
\] specificity, míra skutečně negativních

\[
\frac{FP}{TN + FP} = \frac{FP}{N} = \alpha
\] míra falešně pozitivních

\[
\frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{P + N}
\] nesprávně přesnost

\[
\frac{TP + FN}{TP + TN + FP + FN} = \frac{P}{P + N}
\] prevalence

\[
\frac{TP}{TP + FP} = \frac{TP}{P'}
\] přesnost, prediktivní hodnota pozitivního testu

\[
\frac{TN}{TN + FN} = \frac{TN}{N'}
\] prediktivní hodnota negativního testu

**Jednoduchá hypotéza:** nulové hypotéze odpovídá jediná hodnota parametru.

**Složená hypotéza:** nulové hypotéze odpovídá více hodnot parametru.

**Jednoduchá alternativa:** alternativní hypotéze odpovídá jediná hodnota parametru.

**Složená alternativa:** alternativní hypotéze odpovídá více hodnot parametru.

Často se formuluje nulová a alternativní hypotéza tak, že nejsou navzájem svými negacemi a nepokrývají prostor všech možných hodnot parametru. Vzniká tím jen chaos (viz většina ostatní literatury). Snadno se mu vyhneme, když budeme formulovat nulovou hypotézu jako negaci alternativní hypotézy.

Je-li např. \( H_1 : \vartheta > c \), pak nevolíme \( H_0 : \vartheta = c \), ale \( H_0 : \vartheta \leq c \). (Největší riziko chyby 1. druhu obvykle odpovídá případu \( \vartheta = c \), takže postup je stejný.)

U složené hypotézy požadujeme, aby pravděpodobnost chyby 1. druhu byla nejvýše \( \alpha \) pro všechny hodnoty parametru vyhovující nulové hypotéze.

(Statistická významnost neznamená významnost praktickou.)

**Řešení:** Nulovou hypotézu zamítneme, právě když hodnota kritéria získaná z realizace nepadne do intervalu spolehlivosti pro koeficient spolehlivosti \( 1 - \alpha \), tj. kritická hodnota je mezi intervalového odhadu.

**Obrácený problém:** Při jaké mezní hladině významnosti by pozorovaná hodnota byla kritická; tomu říkáme dosažená významnost; stačí ji porovnat s předem zvolenou hladinou významnosti testu. (Čím nižší číslo, tím významnější výsledek.) Programy obvykle dávají za výsledek dosaženou významnost (obvykle se značí \( P \) a říká se jí pouze *significance*). Výhody: hladinu významnosti není třeba předem zadat, a navíc se dovíme, jak daleko od ní jsme byli.

**Typický tvar testu:** Pro mezní případ nulové hypotézy, \( \vartheta = c \), odvodíme rozdělení testovací statistiky \( T \), která s \( \vartheta \) roste. Kvantily tohoto rozdělení určují intervalový odhad s koeficientem spolehlivosti \( 1 - \alpha \). Nulovou hypotézu zamítneme, pokud realizace \( t \) statistiky \( T \) padne mimo tento interval:
V literatuře se setkáme i s následujícími případy hypotéz, které se však řeší stejně:

<table>
<thead>
<tr>
<th>$H_0$</th>
<th>$H_1$</th>
<th>zamítáme pro</th>
<th>dosažená významnost</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\vartheta \leq c$</td>
<td>$\vartheta &gt; c$</td>
<td>$t &gt; q_T(1 - \alpha)$</td>
<td>$1 - F_T(t)$</td>
</tr>
<tr>
<td>$\vartheta \geq c$</td>
<td>$\vartheta &lt; c$</td>
<td>$t &lt; q_T(\alpha)$</td>
<td>$F_T(t)$</td>
</tr>
<tr>
<td>$\vartheta = c$</td>
<td>$\vartheta \neq c$</td>
<td>$t &gt; q_T(1 - \frac{\alpha}{2})$ nebo $t &lt; q_T(\frac{\alpha}{2})$</td>
<td>$2 \min (F_T(t), 1 - F_T(t))$</td>
</tr>
</tbody>
</table>

V nulové hypotéze daná hodnota $c$ nahrazuje neznámou střední hodnotu $\mu$. Realizaci testové statistiky

$$t = \frac{\bar{x} - c}{\sigma \sqrt{n}}$$

porovnáváme s kvantily normovaného normálního rozdělení:

<table>
<thead>
<tr>
<th>$H_0$</th>
<th>$H_1$</th>
<th>zamítáme pro</th>
<th>dosažená významnost</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu = c$</td>
<td>$</td>
<td>t</td>
<td>&gt; \Phi^{-1}(1 - \frac{\alpha}{2})$</td>
</tr>
<tr>
<td>$\mu \leq c$</td>
<td>$t &gt; \Phi^{-1}(1 - \alpha)$</td>
<td>$1 - \Phi(t)$</td>
<td></td>
</tr>
<tr>
<td>$\mu \geq c$</td>
<td>$t &lt; -\Phi^{-1}(1 - \alpha)$</td>
<td>$\Phi(t)$</td>
<td></td>
</tr>
</tbody>
</table>

Díky centrální limitní větě je odhad použitelný i pro výběr z jiného než normálního rozdělení, pokud má nenuvolový rozptyl a rozsah výběru je velký.

### 9.2.2 Při neznámém rozptylu

$$t = \frac{\bar{x} - c}{s_x \sqrt{n}}$$

porovnáváme s kvantily Studentova rozdělení s $n - 1$ stupni volnosti:

<table>
<thead>
<tr>
<th>$H_0$</th>
<th>zamítáme pro</th>
<th>dosažená významnost</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu = c$</td>
<td>$</td>
<td>t</td>
</tr>
<tr>
<td>$\mu \leq c$</td>
<td>$t &gt; q_{t(n-1)}(1 - \alpha)$</td>
<td>$1 - F_{t(n-1)}(t)$</td>
</tr>
<tr>
<td>$\mu \geq c$</td>
<td>$t &lt; -q_{t(n-1)}(1 - \alpha)$</td>
<td>$F_{t(n-1)}(t)$</td>
</tr>
</tbody>
</table>

Díky centrální limitní větě je odhad použitelný i pro výběr z jiného než normálního rozdělení, pokud má nenuvolový rozptyl a rozsah výběru je velký (pak můžeme místo Studentova rozdělení použít normální).
9.3 Testy rozptylu normálního rozdělení

S pravděpodobností \(1 - \alpha\) je

\[
q_{\chi^2(n-1)} \left( \frac{\alpha}{2} \right) \leq \frac{(n-1)S^2_X}{\sigma^2} \leq q_{\chi^2(n-1)} \left( 1 - \frac{\alpha}{2} \right),
\]

V nulové hypotéze daná hodnota \(c\) nahrazuje neznámý rozptyl \(\sigma^2\).

Realizaci testové statistiky

\[
t = \frac{(n-1) s^2_X}{c}
\]

porovnáváme s kvantily \(\chi^2\)-rozdělení s \(n - 1\) stupni volnosti:

<table>
<thead>
<tr>
<th>(H_0)</th>
<th>zamítáme pro</th>
<th>dosažená významnost</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\sigma^2 = c)</td>
<td>(t &lt; q_{\chi^2(n-1)} \left( \frac{\alpha}{2} \right)) nebo (t &gt; q_{\chi^2(n-1)} \left( 1 - \frac{\alpha}{2} \right))</td>
<td>2 min (F_{\chi^2(n-1)}(t), 1 - F_{\chi^2(n-1)}(t))</td>
</tr>
<tr>
<td>(\sigma^2 \leq c)</td>
<td>(t &gt; q_{\chi^2(n-1)} \left( 1 - \alpha \right))</td>
<td>(1 - F_{\chi^2(n-1)}(t))</td>
</tr>
<tr>
<td>(\sigma^2 \geq c)</td>
<td>(t &lt; q_{\chi^2(n-1)} \left( \alpha \right))</td>
<td>(F_{\chi^2(n-1)}(t))</td>
</tr>
</tbody>
</table>

Díky centrální limitní větě je odhad použitelný i pro výběr z jiného než normálního rozdělení, pokud má nenulový rozptyl a rozsah výběru je velký (pak můžeme místo \(\chi^2\)-rozdělení použít normální).

9.4 Porovnání dvou normálních rozdělení

Předpoklad: Nezávislé výběry

\((X_1, \ldots, X_m)\) z rozdělení \(N(EX, DX)\),
\((Y_1, \ldots, Y_n)\) z rozdělení \(N(EY, DY)\).

9.4.1 Testy rozptylu dvou normálních rozdělení [Fisher]

Je-li \(DX = DY\), pak \(S^2_X = S^2_Y\). Testovací statistikou je

\[
T = \frac{S^2_X}{S^2_Y}.
\]

F-rozdělení (Fisherovo-Snedecorovo rozdělení) s \(\xi\) a \(\eta\) stupni volnosti je rozdělení náhodné veličiny

\[
F = \frac{U}{V},
\]

kde \(U, V\) jsou nezávislé náhodné veličiny s rozdělením \(\chi^2(\xi)\), resp. \(\chi^2(\eta)\).

Značení: \(F(\xi, \eta)\)

Hustota pro \(x > 0\):

\[
f_{F(\xi, \eta)}(x) = c(\xi, \eta) x^{\frac{\xi - 1}{2}} \left( 1 + \frac{\xi}{\eta} x \right)^{-\frac{\xi + \eta}{2}},
\]
$$c(\xi, \eta) = \frac{\Gamma\left(\frac{\xi+\eta}{2}\right)}{\Gamma\left(\frac{\xi}{2}\right) \Gamma\left(\frac{\eta}{2}\right)} \left(\frac{\xi}{\eta}\right)^{\frac{\xi}{2}}$$

Je-li $DX = DY = \sigma^2$, pak dosadíme

$$U := \frac{(m-1)S^2_X}{\sigma^2} \text{ má } \chi^2(m-1),$$

$$V := \frac{(n-1)S^2_Y}{\sigma^2} \text{ má } \chi^2(n-1),$$

$$\xi := m-1, \eta := n-1,$$

$$F = \frac{U}{V} = \frac{(m-1)S^2_X}{(m-1)\sigma^2} \frac{(n-1)S^2_Y}{(n-1)\sigma^2} = \frac{S^2_X}{S^2_Y} = T.$$}

Testujeme realizaci

$$t = \frac{s^2_x}{s^2_y},$$

na rozdělení $F(m-1, n-1)$:

<table>
<thead>
<tr>
<th>$H_0$</th>
<th>zamítáme pro</th>
<th>dosažená významnost</th>
</tr>
</thead>
<tbody>
<tr>
<td>$DX \leq DY$</td>
<td>$t &gt; q_{F(m-1,n-1)}(1-\alpha)$</td>
<td>$1 - F_{F(m-1,n-1)}(t)$</td>
</tr>
<tr>
<td>$DX \geq DY$</td>
<td>$t &lt; q_{F(m-1,n-1)}(\alpha)$</td>
<td>$F_{F(m-1,n-1)}(t)$</td>
</tr>
<tr>
<td>$DX = DY$</td>
<td>$t &lt; q_{F(m-1,n-1)}(\frac{\alpha}{2})$ nebo $t &gt; q_{F(m-1,n-1)}(1-\frac{\alpha}{2})$</td>
<td>$2 \min(F_{F(m-1,n-1)}(t), 1 - F_{F(m-1,n-1)}(t))$</td>
</tr>
</tbody>
</table>

Pro každou hladinu významnosti potřebujeme dvoudimenzionální tabulku kvantilů indexovanou $\xi, \eta$; obvykle je tabelována jen polovina, druhou je třeba dopočítat podle vzorce

$$q_{F(\xi,\eta)}(\beta) = \frac{1}{q_{F(\eta,\xi)}(1-\beta)}.$$ (Pozor na opačné pořadí indexů!)

Lépe je uvažovat $\frac{s^2_Y}{s^2_X}$ místo $\frac{s^2_X}{s^2_Y}$, takže rozlišíme 2 případy:

1. Pro $s^2_x \geq s^2_y$ testujeme

$$t = \frac{s^2_x}{s^2_y} \geq 1$$

na rozdělení $F(m-1, n-1)$:

<table>
<thead>
<tr>
<th>$H_0$</th>
<th>zamítáme pro</th>
<th>dosažená významnost</th>
</tr>
</thead>
<tbody>
<tr>
<td>$DX \leq DY$</td>
<td>$t &gt; q_{F(m-1,n-1)}(1-\alpha)$</td>
<td>$1 - F_{F(m-1,n-1)}(t)$</td>
</tr>
<tr>
<td>$DX \geq DY$</td>
<td>nezamítáme</td>
<td>žádná</td>
</tr>
<tr>
<td>$DX = DY$</td>
<td>$t &gt; q_{F(m-1,n-1)}(1-\frac{\alpha}{2})$</td>
<td>$2 \min(F_{F(m-1,n-1)}(t), 1 - F_{F(m-1,n-1)}(t))$</td>
</tr>
</tbody>
</table>

2. Pro $s^2_x \leq s^2_y$ testujeme

$$t' = \frac{1}{t} = \frac{s^2_y}{s^2_x} \geq 1$$

na rozdělení $F(n-1, m-1)$ (pozor na pořadí počtů stupňů volnosti!):
<table>
<thead>
<tr>
<th>$H_0$</th>
<th>znamítame pro</th>
<th>dosazená významnost</th>
</tr>
</thead>
<tbody>
<tr>
<td>$DX \leq DY$</td>
<td>nezamítame</td>
<td>žádná</td>
</tr>
<tr>
<td>$DX \geq DY$</td>
<td>$t' &gt; q_{F(n-1,m-1)} (1 - \alpha)$</td>
<td>$1 - F_{F(n-1,m-1)} (t')$</td>
</tr>
<tr>
<td>$DX = DY$</td>
<td>$t' &gt; q_{F(n-1,m-1)} (1 - \frac{\alpha}{2})$</td>
<td>$2 \left(1 - F_{F(n-1,m-1)} (t') \right)$</td>
</tr>
</tbody>
</table>

Díky centrální limitní větě je odhad použitelný i pro výběr z jiného než normálního rozdělení, pokud má nenulový rozptyl a rozsah výběru je velký.

### 9.4.2 Testy středních hodnot dvou normálních rozdělení se stejným známým rozptylem $\sigma^2$

$X_m$ má $N \left( EX, \frac{\sigma^2}{m} \right)$,

$Y_n$ má $N \left( EY, \frac{\sigma^2}{n} \right)$,

$X_m - Y_n$ má $N \left( EX - EY, \sigma^2 \left( \frac{1}{m} + \frac{1}{n} \right) \right)$.

Za předpokladu $EX = EY$:

$$ T := \frac{X_m - Y_n}{\sigma \sqrt{\frac{1}{m} + \frac{1}{n}}} \text{ má } N(0,1). $$

Testujeme realizaci $t$ na $N(0,1)$ (viz kapitola 9.2.1).

### 9.4.3 Testy středních hodnot dvou normálních rozdělení s různými známými rozptyly $\sigma^2_X, \sigma^2_Y$

$X_m$ má $N \left( EX, \frac{\sigma^2_X}{m} \right)$,

$Y_n$ má $N \left( EY, \frac{\sigma^2_Y}{n} \right)$,

$X_m - Y_n$ má $N \left( EX - EY, \left( \frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n} \right) \right)$.

Za předpokladu $EX = EY$:

$$ T := \frac{X_m - Y_n}{\sqrt{\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}}} \text{ má } N(0,1). $$

Testujeme realizaci $t$ na $N(0,1)$ (viz kapitola 9.2.1).

Díky centrální limitní větě je odhad použitelný i pro výběr z jiného než normálního rozdělení, pokud má nenulový rozptyl a rozsah výběru je velký.
9.4.4 Testy středních hodnot dvou normálních rozdělení se stejným neznámým rozptylem $\sigma^2$

Nejprve ověříme předpoklad $DX = DY = \sigma^2$ (viz kapitola [9.4.1]). (Ve skutečnosti nemůžeme předpoklad ověřit, jedině vyvrátit; pokusíme se o to, a pokud se to nepodaří, pokračujeme. Bez tohoto předpokladu by byl další postup složitější, viz např. [Mood a kol.].)

Nejprve ověříme předpoklad $D_X = D_Y = \sigma^2$ (viz kapitola 9.4.1).

$X_m$ má $N\left(\mathrm{EX}, \frac{\sigma^2}{m}\right)$,

$Y_n$ má $N\left(\mathrm{EY}, \frac{\sigma^2}{n}\right)$,

$X_m - Y_n$ má $N\left(\mathrm{EX} - \mathrm{EY}, \sigma^2 \left(\frac{1}{m} + \frac{1}{n}\right)\right)$.

Za předpokladu $\mathrm{EX} = \mathrm{EY}$:

$$\frac{X_m - Y_n}{\sigma \sqrt{\frac{1}{m} + \frac{1}{n}}} \quad \text{má} \quad N(0, 1).$$

Pro měření významnosti potřebujeme odhad rozptylu. Máme dva odhady $S^2_X, S^2_Y$ téže hodnoty $\sigma^2$; použijeme jejich vážený průměr takový, abychom znali i jeho rozdělení.

$$\frac{(m-1)S^2_X}{\sigma^2} \quad \text{má} \quad \chi^2(m-1),$$

$$\frac{(n-1)S^2_Y}{\sigma^2} \quad \text{má} \quad \chi^2(n-1),$$

$$\frac{(m-1)S^2_X + (n-1)S^2_Y}{\sigma^2} \quad \text{má} \quad \chi^2(m+n-2)$$

se střední hodnotou $m + n - 2$,

$$\frac{(m-1)S^2_X + (n-1)S^2_Y}{(m+n-2)\sigma^2} = \frac{S^2}{\sigma^2}$$

má střední hodnotu 1 a

$$S^2 := \frac{(m-1)S^2_X + (n-1)S^2_Y}{m+n-2}$$

je nestranný odhad $\sigma^2$, vedoucí na odhad směrodatné odchylky

$$S := \sqrt{\frac{(m-1)S^2_X + (n-1)S^2_Y}{m+n-2}}.$$  

Ten použijeme místo neznámé směrodatné odchylky $\sigma$ a výsledné kritérium

$$T := \frac{X_m - Y_n}{S \sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{X_m - Y_n}{\sigma \sqrt{\frac{1}{m} + \frac{1}{n}}} \quad \text{má} \quad t(m+n-2).$$

Testujeme realizaci $t$ na rozdělení $t(m+n-2)$ (viz kapitola [9.2.2]).

Díky centrální limitní větě je odhad použitelný i pro výběr z jiného než normálního rozdělení, pokud má nenulový rozptyl a rozsah výběru je velký (pak můžeme místo Studentova rozdělení použít normální).
9.4.5 Testy středních hodnot dvou normálních rozdělení - párový test
(inspirováno [SH10], volně upraveno)

**Příklad:** Máme porovnat průměrnou teplotu na dvou místech.

Standardní test středních hodnot dvou normálních rozdělení je slabý kvůli velkému rozptylu, který však má společnou příčinu (vyjádřenou náhodnými veličinami $Z_j$) a projevuje se synchronně v obou výběrech; proto výběry nelze popsat jako stejně rozložené a navzájem nezávislé.

Situaci můžeme popsat následujícím modelem:

$$X_j = Z_j + U_j,$$
$$Y_j = Z_j + V_j - c,$$

kde náhodné veličiny $U_1, U_n, V_1, V_n$ jsou nezávislé, $U_1, U_n$ mají rozdělení $N(0, \sigma^2_U)$, $V_1, V_n$ mají rozdělení $N(0, \sigma^2_V)$ a $c \in \mathbb{R}$. Testujeme hypotézu o hodnotě $c$ (nejčastěji testujieme předpoklad $c = 0$).

Náhodné veličiny $\Delta_j := X_j - Y_j = U_j - V_j + c$ ($j = 1, \ldots, n$) s rozdělením $N(c, \sigma^2_{\Delta})$ (kde $\sigma^2_{\Delta} := \sigma^2_U + \sigma^2_V$) jsou nezávislé. Obecněji nám stačí:

**Předpoklad:** Náhodné veličiny $\Delta_j := X_j - Y_j$ jsou nezávislé a mají rozdělení $N(c, \sigma^2_{\Delta})$.

Pokud rozptyl $\sigma^2_{\Delta}$ známe, testujeme

$$T := \frac{\bar{\Delta} - c}{\sigma_{\Delta} \sqrt{n}} = \frac{\bar{X} - \bar{Y} - c}{\sigma_{\Delta} \sqrt{n}}$$

na rozdělení $N(0,1)$ dle kapitoly 9.2.1.

Pokud rozptyl neznáme, testujeme

$$T := \frac{\bar{\Delta} - c}{S_{\Delta} \sqrt{n}}$$

na rozdělení $t(n-1)$ dle kapitoly 9.2.2.

Díky centrální limitní větě je odhad použitelný i pro výběr z jiného než normálního rozdělení, pokud má nenulový rozptyl a rozsah výběru je velký (pak můžeme místo Studentova rozdělení použít normální).

9.5 Korelace, její odhad a testování
(dle [Likeš, Machek])

Na základě realizace dvojrozměrného náhodného výběru $((x_1, y_1), \ldots, (x_n, y_n))$ můžeme koreliaci

$$\varrho(X, Y) = \frac{E((X - EX)(Y - EY))}{\sigma_X \sigma_Y} \in (-1, 1)$$

odhadnout pomocí korelace empirického rozdělení neboli realizace výběrového koeficientu korelace

$$r_{x,y} = r_{\text{Emp}(x, y)} = \frac{\sum_{j=1}^{n} (x_j - \bar{x}) (y_j - \bar{y})}{\sqrt{\left(\sum_{j=1}^{n} (x_j - \bar{x})^2\right) \left(\sum_{j=1}^{n} (y_j - \bar{y})^2\right)}} \in (-1, 1),$$
což je kosinus úhlu vektorů

\[(x_1 - \bar{x}, \ldots, x_n - \bar{x}), (y_1 - \bar{y}, \ldots, y_n - \bar{y}) \in \mathbb{R}^n.\]

Jednopříchozý vzorec:

\[
r_{x,y} = \frac{n \sum_{j=1}^{n} x_j y_j - \left(\sum_{j=1}^{n} x_j\right) \left(\sum_{j=1}^{n} y_j\right)}{\sqrt{n \sum_{j=1}^{n} x_j^2 - \left(\sum_{j=1}^{n} x_j\right)^2} \sqrt{n \sum_{j=1}^{n} y_j^2 - \left(\sum_{j=1}^{n} y_j\right)^2}} = \frac{n}{n-1} \frac{1}{s_{x} s_{y}} (\bar{x} \bar{y}).
\]

**Výběrový koeficient korelace** je odpovídající odhad

\[
R_{X,Y} = \frac{\sum_{j=1}^{n} (X_j - \bar{X})(Y_j - \bar{Y})}{\sqrt{\left(\sum_{j=1}^{n} (X_j - \bar{X})^2\right) \left(\sum_{j=1}^{n} (Y_j - \bar{Y})^2\right)}}.
\]

**9.5.1 Test nekorelovanosti dvou normálních rozdělení**

**Předpoklad:** Dvojrozměrná náhodná veličina \((X, Y)\) má (dvojrozměrné) normální rozdělení, \(n \geq 3\).

Testovací statistikou je

\[
T = \frac{R_{X,Y} \sqrt{n-2}}{\sqrt{1 - R_{X,Y}^2}},
\]

za předpokladu nekorelovanosti má rozdělení \(t(n-2)\), dále postupujeme dle kapitoly **9.2.2** (pro oboustranný test i jednostranné testy).

Díky centrální limitní větě je odhad použitelný i pro výběr z jiného než normálního rozdělení, pokud má nenulový rozptyl a rozsah výběru je velký (pak můžeme místo Studentova rozdělení použít normální).

**9.6 \(\chi^2\)-test dobré shody**

**9.6.1 Základní podoba testu**

Slouží k testování hypotézy, že náhodná veličina má předpokládané rozdělení. Protože umíme hypotézy jen zamítit, nikdy nepotvrdíme, že takové rozdělení opravdu má.

Testujeme **diskrétní rozdělení** (mohlo vzniknout diskretizací spojitého).

**\(H_0\)**: Náhodná veličina má diskrétní rozdělení do \(k\) tříd s nemovými pravděpodobnostmi \(p_1, \ldots, p_k\).
Testujeme pomocí realizace náhodného výběru rozsahu $n$. Není důležité pořadí výsledků, pouze jejich četnosti $N_i$, resp. realizace četností $n_i$ nebo realizace relativních četností $\frac{n_i}{n}$ ($i = 1, \ldots, k$). Porovňáváme je s teoretickými četnostmi $n_p$.

Speciální případ: Pro $k = 2$ mají $N_1, N_2$ binomická rozdělení $\text{Bi}(n, p_1), \text{Bi}(n, p_2)$, která lze pro velká $n$ přibližně nahradit normálními,

$$N(n p_1, n p_1 (1 - p_1)) = N(n p_1, n p_1 p_2),$$

$$N(n p_2, n p_2 (1 - p_2)) = N(n p_2, n p_1 p_2).$$

Kvadráty normovaných veličin

$$(\text{norm } N_1) = \frac{(N_1 - n p_1)}{n p_1 p_2},$$

$$(\text{norm } N_2) = \frac{(N_2 - n p_2)}{n p_1 p_2}$$

mají přibližně rozdělení $\chi^2(1)$. Jsou to tytéž náhodné veličiny, neboť

$$N_2 - n p_2 = n - N_1 - n (1 - p_1) = -(N_1 - n p_1),$$

takže rozdělení přibližně $\chi^2(1)$ má náhodná veličina

$$(\text{norm } N_1) = p_2 (\text{norm } N_1) + p_1 (\text{norm } N_1)$$

$$= p_2 (\text{norm } N_1) + p_1 (\text{norm } N_2)$$

$$= \frac{(N_1 - n p_1)}{n p_1} + \frac{(N_2 - n p_2)}{n p_2} = \sum_{i=1}^{2} \frac{(N_i - n p_i)}{n p_i}.$$

Obecně pro libovolné $k$ je testovací statistikou

$$T := \sum_{i=1}^{k} \frac{(N_i - n p_i)}{n p_i},$$

jejíž rozdělení se pro $n \to \infty$ blíží $\chi^2(k-1)$. Její realizace

$$t := \sum_{i=1}^{k} \frac{(n_i - n p_i)}{n p_i}.$$

Dosažená významnost: $1 - F_{\chi^2(k-1)}(t)$. Nulovou hypotézu zamítáme pro $t > q_{\chi^2(k-1)}(1 - \alpha)$, tj. $1 - F_{\chi^2(k-1)}(t) < \alpha$.

Modifikace: Pokud chceme naopak odhalit překvapivě dobrou shodu s modelem, použijeme dolní intervalový odhad a nulovou hypotézu zamítáme pro $t < q_{\chi^2(k-1)}(\alpha)$. Dosažená významnost $F_{\chi^2(k-1)}(t)$.

Cvičení. Tabulka udává rozdělení (podmíněné) pravděpodobnosti, že volič strany zastoupené v parlamentu volil danou stranu. Posuďte na 5% hladině významnosti hypotézu, že stejné rozdělení mají i poslanci.

<table>
<thead>
<tr>
<th>relativní preference</th>
<th>0.376</th>
<th>0.344</th>
<th>0.136</th>
<th>0.077</th>
<th>0.067</th>
</tr>
</thead>
<tbody>
<tr>
<td>počet poslanců</td>
<td>81</td>
<td>74</td>
<td>26</td>
<td>13</td>
<td>6</td>
</tr>
</tbody>
</table>
Rešení. Doplníme tabulku (poslední sloupec uvádí celkový údaj):

<table>
<thead>
<tr>
<th>relativní preference</th>
<th>0.376</th>
<th>0.344</th>
<th>0.136</th>
<th>0.077</th>
<th>0.067</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>počet poslanců</td>
<td>81</td>
<td>74</td>
<td>26</td>
<td>13</td>
<td>6</td>
<td>200</td>
</tr>
<tr>
<td>teor. četnost</td>
<td>75.2</td>
<td>68.8</td>
<td>27.2</td>
<td>15.4</td>
<td>13.4</td>
<td>200</td>
</tr>
<tr>
<td>příspěvek k $\chi^2$</td>
<td>0.447</td>
<td>0.393</td>
<td>0.052</td>
<td>0.374</td>
<td>4.086</td>
<td>5.353</td>
</tr>
</tbody>
</table>

Hodnotu kritéria $5.353$ porovnáme s kvantilem $q_{\chi^2(4)}(0.95) \approx 9.4877$ a hypotézu nezamítáme (poněkud překvapivý závěr vzhledem k tomu, že poslední dvě strany mají téměř stejnou podporu voličů, ale poslední má více než $2 \times$ méně poslanců).

9.6.2 Modifikace

Problém: Testujeme na rozdělení, kterému se skutečně jen limitně blíží. Tím se dopouštíme blíže neurčené dodatečné chyby. Teoretické četnosti tříd nesmí být příliš malé (aspoň 5), aby nás předpoklad byl oprávněný.

Modifikace: Vychází-li teoretická četnost některých tříd příliš malá, sloučíme je s jinými třídami (pokud možno „blízkými“).

Poznámka: Pokud data předpokládané rozdělení nemají a rozsah výběru zvětšíme $m \times$, systematický příspěvek ke kritériu se rovněž zvýší $m^2 / m = m \times$. Proto se nedivme velkým hodnotám kritéria (velké síle testu) pro rozsáhlé výběry.

Problém: Zkoumané rozdělení může záviset na neznámých parametrech.

Modifikace 1: Parametry odhadneme na základě jiného náhodného výběru.

Modifikace 2: Parametry odhadneme na základě stejného náhodného výběru, který používáme k testu dobré shody. Tím jsme však snížili počet stupňů volnosti, takže musíme testovat na rozdělení $\chi^2(k - 1 - q)$, kde $q$ je počet odhadnutých parametrů.

Problém: Chceme testovat shodu se spojitým nebo smíšeným rozdělením.

Modifikace: Rozdělení napřípad diskretizujeme, tj. všechny možné výsledky rozdělíváme do $k$ disjunktních tříd. Prvky v jedné třídě si mají být „blízké“, jinak snížujeme sílu testu. Všechny teoretické četnosti musí být dostatečně velké a nejlépe zhruba stejně.

Poznámka: Zásadně musíme pracovat s jednotkami (objekty), z nichž každá zvlášt (a nezávisle) je zařazena do nějaké třídy. Nelze počítat s tisíci, procenty, spojitým množstvím atd.

9.6.3 $\chi^2$-test nezávislosti dvou rozdělení

(dle [Likeš, Machek])

$H_0$: Dvě diskrétní náhodné veličiny (jejichž rozdělení neznáme) jsou nezávislé.

$X$ nabývá $k$ hodnot s pravděpodobnostmi $p_1, \ldots, p_k$,

$Y$ nabývá $m$ hodnot s pravděpodobnostmi $q_1, \ldots, q_m$.

Realizace dvojrozměrného náhodného výběru $((x_1, y_1), \ldots, (x_n, y_n))$ obsahuje dvojice realizací náhodných veličin $X, Y$; potřebujeme pouze četnosti $N_{ij}$, resp. jejich realizace $n_{ij}$ ($i = 1, \ldots, k$, $j = 1, \ldots, m$).
Ty bývají uspořádány do tzv. kontingenční tabulek. Počet tříd je \( km \).

Za předpokladu nezávislosti jsou pravděpodobnosti výsledků \( p_i q_j \) (\( i = 1, \ldots, k \); \( j = 1, \ldots, m \)),

\[
T := \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(N_{ij} - n p_i q_j)^2}{n p_i q_j}
\]

se blíží \( \chi^2(km - 1) \).

Použijeme realizaci odhadu

\[
t := \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(n_{ij} - n p_i q_j)^2}{n p_i q_j},
\]

kde neznámé parametry \( p_i, q_j \) odhadneme pomocí maxima věrohodnosti nebo parametry empirického rozdělení,

\[
p_i = \frac{1}{n} \sum_{j=1}^{m} n_{ij}, \quad q_j = \frac{1}{n} \sum_{i=1}^{k} n_{ij}.
\]

Z nich je jen \((k - 1) + (m - 1)\) nezávislých (neboť \( \sum_{i=1}^{k} p_i = 1, \sum_{j=1}^{m} q_j = 1 \)), takže výsledný počet stupňů volnosti je \( km - 1 - (k - 1) - (m - 1) = (k - 1)(m - 1) \) a testujeme \( t \) na \( \chi^2((k - 1)(m - 1)) \). Nulovou hypotézu zamítáme pro \( t > q_{\chi^2((k-1)(m-1))}(1 - \alpha) \). Dosažená významnost: \( 1 - F_{\chi^2((k-1)(m-1))}(t) \).

9.6.4 \( \chi^2 \)-test dobré shody dvou rozdělení

\( H_0 \): Dva náhodné výběry pocházejí ze stejného diskrétního rozdělení.

Rozsahy výběru jsou \( m, n \), četnosti výsledků \( m_i, n_i \) (\( i = 1, \ldots, k \)).

Sjednocení obou výběrů považujeme za \( m + n \) realizaci náhodné veličiny \( X \) s neznámými pravděpodobnostmi \( p_i \) (\( i = 1, \ldots, k \)), jejichž maximálně věrohodný odhad je

\[
p_i = \frac{m_i + n_i}{m + n}.
\]

Zavedeme druhou náhodnou veličinu \( Y \) se dvěma hodnotami (např. 1, 2), které označují příslušnost k prvnímu, resp. druhému z původních výběrů. Např. z výběrů

\[
(4, 5, 6, 4), \quad (6, 6, 6, 5, 4)
\]

vytvoříme dvojrozměrný výběr

\[
((4, 1), (5, 1), (6, 1), (4, 1), (6, 2), (6, 2), (6, 2), (5, 2), (5, 2), (4, 2)).
\]

Za předpokladu platnosti nulové hypotézy je \( X \) nezávislý na \( Y \), což můžeme testovat stejně jako v předchozí metodě (na rozdělení \( \chi^2(k - 1) \)).

Nulovou hypotézu zamítáme pro \( t > q_{\chi^2(k-1)}(1 - \alpha) \). Dosažená významnost: \( 1 - F_{\chi^2(k-1)}(t) \).
Praktičtější (ekvivalentní) vzorec Mood a kol.:

\[ t = \left( \frac{1}{m} + \frac{1}{n} \right) \sum_{i=1}^{k} \frac{(m_i - m p_i)^2}{p_i} \]

### 9.7 Neparametrické testy

Jsou použitelné bez ohledu na typ rozdělení, jsou však slabší.

#### 9.7.1 Znaménkový test

Rozlišujeme pouze znamení odchylky od zvolené hodnoty \( c \). Tím ztrácíme kvantitativní informaci a tedy i možnost testovat např. střední hodnotu. Místo ní testujeme medián \( q_X(\frac{1}{2}) \).

- **H_0:** \( q_X(\frac{1}{2}) = c \)

Při platnosti nulové hypotézy by kladné i záporné odchylky měly být stejně pravděpodobné. Nulové odchylky z výběru předem vyloučíme. Testovací statistikou \( T \) je počet kladných odchylek, který testujeme na binomické rozdělení Bin \( (n, \frac{1}{2}) \). Nulovou hypotézu zamítáme pro

\[ t < q_{\text{Bin}}(n, \frac{1}{2}) \left( \frac{\alpha}{2} \right) \text{ nebo } t > q_{\text{Bin}}(n, \frac{1}{2}) \left( 1 - \frac{\alpha}{2} \right) \]

(Podobně pro jednostranné testy.) Výpočet kvantilů je pracný, ale kritické hodnoty jsou tabelovány (v závislosti na \( n \) a hladině významnosti).

Dosažená významnost se počítá o trochu snáze.

Pro velká \( n \) používáme centrální limitní větu a testujeme

\[ T_0 := \frac{2T - n}{\sqrt{n}} \]

na \( N(0,1) \).

Lze použít i k porovnání dvou mediánů u párového testu.

**Příklad použití:** Odhad smrtné dávky látky.

**Na rozdíl od střední hodnoty medián vždy existuje (je však problém, jak ho definovat, aby byl jednoznačný).**

**Jeho výpočetní složitost je větší, řádu \( n \ln n \).**

#### 9.7.2 Wilcoxonův test (jednovýběrový)

- **H_0:** \( X \) má rozdělení symetrické kolem hodnoty \( c \)
  
  (V tom případě je \( c \) mediánem i střední hodnotou.)

Z realizace \((x_1, \ldots, x_n)\) vypočteme posloupnost \((z_1, \ldots, z_n)\), kde \( z_j = x_j - c \). Seřadíme ji vzestupně podle absolutních hodnot \( |z_j| = |x_j - c| \), čímž \( j \)-těmu prvku přiřadíme pořadí \( r_j \).

Je-li více stejných rozdílů, přiřadíme jim stejné pořadí rovné aritmetickému průměru. Testovací statistikou je

\[ T_1 := \sum_{j: z_j > 0} r_j \]

nebo

\[ T_2 := \min \left( \sum_{j: z_j > 0} r_j, \sum_{j: z_j < 0} r_j \right), \]

porovnáme s tabulkou kritických hodnot pro tento test.
Literatura


