Pravděpodobnost a matematická statistika

Mirko Navara
Centrum strojového vnímání
katedra kybernetiky FEL ČVUT
Karlovo náměstí, budova G, místnost 104a
http://cmp.felk.cvut.cz/~navara/stat

24. listopadu 2020

Obsah

1 O čem to je?

1.1 Teorie pravděpodobnosti

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

Poskytuje model 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 zítra bude v menze dobrý oběd?

Realizace náhodného pokusu – např. losovací zařízení:

- Kostka, čtyřstěn, dvanáctistěn...
- Tužka, dlouhý hranol...
- „**Kolo štěstí.**“
- **Urna s losy**.

2.2 Laplaceova (klasická) definice pravděpodobnosti

Předpoklad: Náhodný pokus s $n \in \mathbb{N}$ různými, po dvou neslučitelnými výsledky, které jsou **stejně možné**. Jev, který nastává právě při k z těchto výsledků, má pravděpodobnost k/n.
(Urna s n 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: Ω *(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: Ω, **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: $\bar{A} = \Omega \setminus A$ *(losy, které nevyhrávají)*

$A \Rightarrow B$: $A \subseteq B$

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 $\exp \Omega$ (prozatím množina všech podmnožin množiny Ω)

Pravděpodobnost jevu A:

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

kde $|\,.\,|$ značí počet prvků množiny

2.3 Vlastnosti pravděpodobnosti

$P(A) \in \langle 0, 1\rangle$
$P(\mathbf{0}) = 0, \qquad P(\mathbf{1}) = 1$
$P(\overline{A}) = 1 - P(A)$
$A \subseteq B \Rightarrow P(A) \leq P(B)$
$A \subseteq B \Rightarrow P(B \setminus A) = P(B) - P(A)$
$A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$ **(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 $\bigcup_{i \in I} B_i = \mathbf{1}$.

Příklad: Ruleta: např.
A. $\{0\}, \{1\}, \ldots, \{36\}$;
B. sudá, lichá;
C. $\{0\}$, červená, černá.

Speciální případ pro 2 jevy: $\{C, \overline{C}\}$, $B_1 = C, B_2 = \overline{C}$.

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 nekuřáků): Mužů je v populaci 48 %, kuřáků a kuřaček dohromady 30 %. Jakých hodnot může nabývat pravděpodobnost, že náhodně vybraný člověk je muž a nekuřák?

M ... muž, $P(M) = 0.48$
K ... kuřák, $P(K) = 0.3$, $P(\overline{K}) = 0.7$
Hledaná pravděpodobnost $P(M \cap \overline{K})$ jevu opačného k $\overline{M} \cup K$,
$\max\{P(\overline{M}), P(K)\} = 0.52 \leq P(\overline{M} \cup K) \leq P(\overline{M}) + P(K) = 0.52 + 0.3 = 0.82$
$1 - 0.82 = 0.18 \leq P(M \cap \overline{K}) \leq 1 - 0.52 = 0.48$
Možné jsou všechny hodnoty z intervalu $\langle 0.18, 0.48\rangle$.

2.3.2 Náhodná veličina

je (prozatím libovolná) **funkce** $X \colon \Omega \to \mathbb{R}$

Příklady: Teplota v Klementinu zítra v 6 hodin.
Známka ze zkoušky.
Kolo štěstí. Na každém poli je napsána částka výhry.
Toto je univerzální příklad na pravděpodobnost dle Laplace.

2.3.3 Střední hodnota

Příklad: Kolo štěstí má 4 stejně velké sekce (A, B, C, D), na které připadají výhry $0, 10, 50, 100$.
Elementární jevy jsou všechny možné výsledky,
$\Omega = \{A, B, C, D\}$,
náhodná veličina X je výše výhry:

ω	A	B	C	D
$X(\omega)$	0	10	50	100

Střední hodnota je spravedlivá cena za účast ve hře:

$$\mathrm{E}X = \frac{0 + 10 + 50 + 100}{4} = \frac{160}{4} = 40\,.$$

Obecně

$$\mathrm{E}X = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} X(\omega)\,.$$

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 prováděn „rovnoměrně“).

3. problém: Nedovoluje iracionální hodnoty pravděpodobnosti.

Příklad: Kolo štěstí s nestejnými sekcemi.

2.4.1 Rozšíření Laplaceova modelu pravděpodobnosti

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].)
Losovací zařízení:
Urna s losy, které nelze rozlišit před losováním a lze rozlišit po vylosování.
Losů je n (zde 9, číslovaných $1, 2, 3, 4, 5, 6, 7, 8, 9$).
Větší počet výsledků umožní **opakované losování**, které lze provést různě.

2.5.1 Permutace (pořadí) bez opakování

Vytáhneme všechny losy, záleží na pořadí.

1.	2.	3.	4.	5.	6.	7.	8.	9.
2	7	1	8	3	9	4	5	6

$9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 9!$
Obecně $n!$ výsledků stejně pravděpodobných
Nadále postupně vytáhneme k losů (zde 6).

2.5.2 Variace s opakováním (uspořádaný výběr s vracením)

(záleží na pořadí, např. číselný zámek)

1.	2.	3.	4.	5.	6.
2	7	1	8	2	8

9^6
Obecně n^k výsledků stejně pravděpodobných

2.5.3 Variace bez opakování (uspořádaný výběr bez vracení)

(záleží na pořadí, $k \leq n$)

1.	2.	3.	4.	5.	6.
2	7	1	8	~~2~~3	~~8~~9

$9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 = \dfrac{9!}{3!}$
Obecně $\dfrac{n!}{(n-k)!}$ výsledků stejně pravděpodobných

2.5.4 Kombinace bez opakování (neuspořádaný výběr bez vracení)

(nezáleží na pořadí, $k \leq n$)

1	2	3	4	5	6	7	8	9
x	x	x				x	x	x

$$\frac{9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = \frac{9!}{3!\,6!} = \binom{9}{6}$$

Obecně $\binom{n}{k}$ výsledků stejně pravděpodobných

2.5.5 Kombinace s opakováním (neuspořádaný výběr s vracením)

(nezáleží na pořadí)

1	2	3	4	5	6	7	8	9
x	x					x	x	
	x						x	

$1\,|\,22\,|\,|\,|\,|\,|\,7\,|\,88\,|$
$\cdot\,|\,\cdot\cdot\,|\,|\,|\,|\,|\,\cdot\,|\,\cdot\cdot\,|$

$$\binom{14}{6}$$

Obecně $\binom{n+k-1}{k}$ výsledků nestejně pravděpodobných.

1	2	3	4	5	6	7	8	9
x								
x								
x								
x								
x								
x								

$$P(111111\,|\,|\,|\,|\,|\,|\,|\,|\,|) = \frac{1}{9^6}$$

$\ll$

1	2	3	4	5	6	7	8	9
x	x	x	x	x	x			

$$P(1\,|\,2\,|\,3\,|\,4\,|\,5\,|\,6\,|\,|\,|\,) = \frac{6!}{9^6}$$

výběr	s vracením (opakováním)	bez vracení (opakování)
uspořádaný (variace)	n^k s pravděpodobnostmi $\frac{1}{n^k}$	$\frac{n!}{(n-k)!}$ s pravděpodobnostmi $\frac{(n-k)!}{n!}$
neuspořádaný (kombinace)	$\binom{n+k-1}{k}$ s **různými** pravděpodobnostmi	$\frac{n!}{k!\,(n-k)!} = \binom{n}{k}$ s pravděpodobnostmi $\frac{k!\,(n-k)!}{n!}$

Z této tabulky pouze **kombinace s opakováním nejsou všechny stejně pravděpodobné** (odpovídají různému počtu variací s opakováním) a nedovolují proto použití Laplaceova modelu pravděpodobnosti.

2.5.6 Permutace s opakováním

Některé losy mohou být nerozlišitelné, např. $1, 1, 1, 2, 2, 3, 3, 4, 5$,

obecněji $\underbrace{1, \ldots, 1}_{k_1\times}, \underbrace{2, \ldots, 2}_{k_2\times}, \ \ldots, \ \underbrace{i, \ldots, i}_{k_i\times}$, kde $\sum_{j=1}^{i} k_j = n$.

(záleží na pořadí)

1.	2.	3.	4.	5.	6.	7.	8.	9.
3	1	4	1	5	3	2	1	2

$3! \cdot 2! \cdot 2!$ (obecně $k_1! \cdot \ldots \cdot k_i!$) permutací dává stejný výsledek.

$\frac{9!}{3! \cdot 2! \cdot 2!}$ (obecně $\frac{n!}{k_1! \cdot \ldots \cdot k_i!}$) výsledků stejně pravděpodobných.

Speciálně pro $i = 2$: $\dfrac{n!}{k_1! \cdot k_2!} = \dfrac{n!}{k_1! \cdot (n-k_1)!} = \binom{n}{k_1}$
(k_1-prvkové kombinace bez opakování z n prvků;
rozlišujeme pouze k_1 „vylosovaných“ a $n - k_1$ „nevylosovaných“).

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

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.*

$$\lim_{n\to\infty} \frac{n!}{(n-k)!\, n^k} = 1\,, \quad \textit{resp.} \qquad \lim_{n\to\infty} \frac{\binom{n}{k}}{\binom{n+k-1}{k}} = 1\,.$$

Důkaz.

$$\begin{aligned}
\frac{n!}{(n-k)!\, n^k} &= \frac{n\,(n-1)\cdots(n-(k-1))}{n^k} = \\
&= 1\,(1-\frac{1}{n})\cdots(1-\frac{k-1}{n}) \to 1\,, \\
\frac{\binom{n}{k}}{\binom{n+k-1}{k}} &= \frac{n\,(n-1)\cdots(n-(k-1))}{(n+(k-1))\cdots(n+1)\,n} = \\
&= \frac{1\,(1-\frac{1}{n})\cdots(1-\frac{k-1}{n})}{(1+\frac{k-1}{n})\cdots(1+\frac{1}{n})\,1} \to 1
\end{aligned}$$

(počet činitelů k je konstantní). □

Důsledek. *Pro $n \gg k$ je počet variací (resp. kombinací) bez opakování přibližně*

$$\frac{n!}{(n-k)!} \doteq n^k\,, \quad \textit{resp.} \qquad \binom{n}{k} \doteq \frac{n^k}{k!}\,.$$

Jednodušší bývají **variace s opakováním** (uspořádaný výběr s vracením) nebo **kombinace bez opakování** (neuspořádaný výběr bez vracení).

2.6 Kolmogorovova definice pravděpodobnosti

Elementární jevy = všechny možné výsledky pokusu = prvky množiny Ω.
Může jich být **nekonečně mnoho, nemusí být stejně pravděpodobné**.

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

(A1) $\emptyset \in \mathcal{A}$.
(A2) $A \in \mathcal{A} \Rightarrow \bar{A} \in \mathcal{A}$.

(A3) $(\forall n \in \mathbb{N} : A_n \in \mathcal{A}) \Rightarrow \bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}$.

Systém $\mathcal{A}$ podmnožin nějaké množiny Ω, který splňuje podmínky (A1)–(A3), se nazývá **σ-algebra**.
Důsledky: $\Omega = \overline{\emptyset} \in \mathcal{A}$,

$$(\forall n \in \mathbb{N} : A_n \in \mathcal{A}) \Rightarrow \bigcap_{n\in\mathbb{N}} A_n = \overline{\bigcup_{n\in\mathbb{N}} \overline{A_n}} \in \mathcal{A}\,.$$

Maximalista: Volme $\mathcal{A} = \exp\Omega$.
Vede k nežádoucím problémům, např. Banachův-Tarského paradox.

(A3) je uzavřenost na **spočetná** sjednocení.

Maximalista: Volme uzavřenost na **jakákoli** sjednocení.
A jsme tam, kde jsme byli...

Praktik: 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 σ-algebra

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

Obsahuje všechny intervaly otevřené, uzavřené, polouzavřené, jejich spočetná sjednocení, některé další množiny (např. Cantorovo diskontinuum), ale ne všechny. Její prvky nazýváme **borelovské množiny**.

2.6.2 Pravděpodobnost (=pravděpodobnostní míra)

je funkce $P\colon \mathcal{A} \to \langle 0, 1\rangle$, splňující podmínky

(P1) $P(\mathbf{1}) = 1$,

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

Trojice $(\Omega, \mathcal{A}, P)$, kde Ω je neprázdná množina, $\mathcal{A}$ je σ-algebra podmnožin množiny Ω a $P\colon \mathcal{A} \to \langle 0, 1\rangle$ je pravděpodobnost, se nazývá **pravděpodobnostní prostor**.

Dříve uvedené vlastnosti pravděpodobnosti jsou důsledkem (P1), (P2).
Praktik: Spokojme se s **konečnou** aditivitou.
Problémem je např. pravděpodobnost výsledku v kruhu coby sjednocení obdélníků (spočetně mnoha).

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

Maximalista: Požadujme **úplnou** aditivitu (pro jakékoli soubory po dvou neslučitelných jevů).
Pak bychom nepřipouštěli žádné spojité rozdělení (ani rovnoměrné ani normální).

Pravděpodobnost zachovává limity monotónních posloupností jevů (množin):
Nechť $(A_n)_{n\in\mathbb{N}}$ je posloupnost jevů.

$$A_1 \subseteq A_2 \subseteq \ldots \Rightarrow P\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \lim_{n\to\infty} P(A_n)\,,$$
$$A_1 \supseteq A_2 \supseteq \ldots \Rightarrow P\Big(\bigcap_{n\in\mathbb{N}} A_n\Big) = \lim_{n\to\infty} P(A_n)\,.$$

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

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

$$\frac{2}{\pi} \doteq 0.63661977236758134307553505349005744\,.$$

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 $A, \overline{B}$ (a též dvojice jevů $\overline{A}, B$ a $\overline{A}, \overline{B}$). **Důkaz:**

$$\begin{aligned} P(A \cap \overline{B}) &= P(A) - P(A \cap B) = P(A) - P(A) \cdot P(B) = \\ &= P(A) \cdot (1 - P(B)) = P(A) \cdot P(\overline{B})\,. \end{aligned}$$

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 líc,
B ... při druhém hodu padne líc,
C ... při právě jednom hodu padne líc.

$$\begin{aligned} P(A) = P(B) = P(C) &= \frac{1}{2}\,, \\ P(A \cap B) = P(A \cap C) = P(B \cap C) &= \frac{1}{4} = P(A) \cdot P(B) = P(A) \cdot P(C) = P(B) \cdot P(C)\,, \\ P(A \cap B \cap C) = 0 &\neq \frac{1}{8} = P(A) \cdot P(B) \cdot P(C)\,, \end{aligned}$$

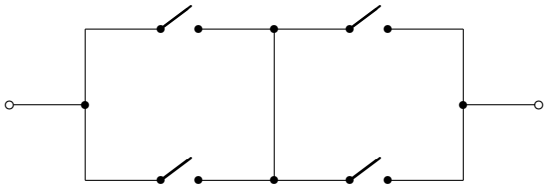
takže jevy A, B, C jsou po dvou nezávislé, ale nejsou nezávislé.

Množina jevů $\mathcal{M}$ se nazývá **nezávislá**, jestliže

$$P\Big(\bigcap_{A \in \mathcal{K}} A\Big) = \prod_{A \in \mathcal{K}} P(A)$$

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

$$0.99^2 = 0.9801\,.$$

Příklad. *(Sally Clark): Syndrom náhlého úmrtí kojenců se vyskytuje s pravděpodobností* $1/8500$. *Jaká je pravděpodobnost dvou úmrtí po sobě v jedné rodině?*
Soud uvěřil žalobci, že $1/8500^2 = 1/72\ 250\ 000$, *a matku odsoudil.*
Později osvobozena ona i 3 další ženy.
I kdyby platila nezávislost, velmi pravděpodobně by byl někdo neprávem odsouzen.

3.2 Podmíněná pravděpodobnost

Motivační příklad (alkohol za volantem):
90 % všech nehod způsobili střízliví řidiči.
Alkoholik: Když se napiju, budu mít 9× menší riziko havárie.
Statistik: To by byla pravda, kdyby opilých bylo stejně jako střízlivých.
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 naši 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:

$$P(A) = P(B) \underbrace{\frac{P(A \cap B)}{P(B)}}_{P(A|B)} + P(\bar{B}) \underbrace{\frac{P(A \cap \bar{B})}{P(\bar{B})}}_{P(A|\bar{B})}\,.$$

Funkce $P(.|B), P(.|\bar{B}) \colon \mathcal{A} \to \langle 0, 1\rangle$,

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

jsou pravděpodobnosti na $\mathcal{A}$, neboť splňují

(P1) $P(\mathbf{1}|B) = \dfrac{P(\mathbf{1} \cap B)}{P(B)} = \dfrac{P(B)}{P(B)} = 1$

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

$$\text{(P2)}\quad P\Big(\bigcup_{n\in\mathbb{N}} A_n \Big| B\Big) = \frac{P\Big(\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) \cap B\Big)}{P(B)} = \frac{P\Big(\bigcup_{n\in\mathbb{N}} (A_n \cap B)\Big)}{P(B)} = \frac{\sum_{n\in\mathbb{N}} P(A_n \cap B)}{P(B)} =$$
$$= \sum_{n\in\mathbb{N}} P(A_n|B)\,.$$

(Obdobně pro $P(.|\bar{B})$.)
Nazývají se **podmíněné pravděpodobnosti**.

$P(A|B)$ čteme např. „pravděpodobnost jevu A za podmínky B.“
Je-li $P(A|B)$ definována, jsou jevy A, B **nezávislé**, právě když $P(A|B) = P(A)$.
Podmíněné pravděpodobnosti navíc splňují $B \subseteq A \Rightarrow P(A|B) = 1\,, \qquad P(A \cap B) = 0 \Rightarrow P(A|B) = 0,$
speciálně $P(B|B) = 1\,, \qquad P(\overline{B}|B) = 0.$
(Obdobně pro $P(.|\overline{B})$.)

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:

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

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

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

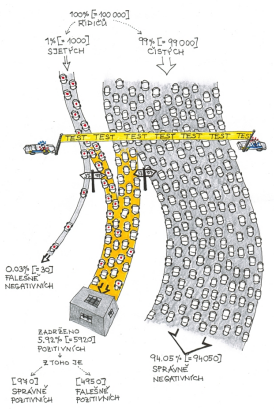
Důkaz (s využitím věty o úplné pravděpodobnosti):

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

Motivační příklad (test na drogy): Policií užívaný test DrugWipe 5S je falešně pozitivní u 5 % nezdrogovaných a falešně negativní u 3 % zdrogovaných. Jaká je (podmíněná) pravděpodobnost, že řidič s pozitivním testem je pod vlivem drog?
D ... zdrogovaný, $\qquad T$ ... pozitivní test,
$P(T|\overline{D}) = 0.05, \qquad P(T|D) = 1 - P(\overline{T}|D) = 1 - 0.03 = 0.97.$
Potřebujeme ještě znát podíl zdrogovaných řidičů na silnicích $P(D)$.
Policejní ani statistická data nemáme, tedy jen zkusíme, co bychom dostali pro
$P(D) = 0.01, \qquad P(\overline{D}) = 0.99$:

$$P(T) = P(T|D) \cdot P(D) + P(T|\overline{D}) \cdot P(\overline{D})$$
$$= 0.97 \cdot 0.01 + 0.05 \cdot 0.99 = 0.0097 + 0.0495 = 0.0592$$

$$P(D|T) = \frac{P(D) \cdot P(T|D)}{P(D) \cdot P(T|D) + P(\overline{D}) \cdot P(T|\overline{D})}$$
$$= \frac{P(D) \cdot P(T|D)}{P(T)} = \frac{0.0097}{0.0592} \doteq 0.164\,.$$



Zdroj: https://finmag.penize.cz/kaleidoskop/407932-velka-drogova-kocovina

Význam: Pravděpodobnosti $P(A|B_i)$ odhadneme z pokusů nebo z modelu, pomocí nich určíme pravděpodobnosti $P(B_i|A)$, které slouží k „optimálnímu“ odhadu, který z jevů B_i nastal jestliže jsme pozorovali A.

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, $j \in \{1, \dots, m\}$

A_i ... přijat i-tý výstupní znak, $i \in \{1, \dots, k\}$ (může být $k \neq m$)

Lze odhadnout podmíněné pravděpoděpodobnosti $P(A_i|B_j)$, že znak j bude přijat jako i.

Z apriorních pravděpodobností (vyslání znaku j) $P(B_j)$ můžeme maticovým násobením určit pravděpodobnosti přijatých znaků:

$$\begin{bmatrix} P(A_1) & P(A_2) & \cdots & P(A_k) \end{bmatrix} =$$

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

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

$$P(B_j|A_i) = \frac{P(B_j)\, P(A_i|B_j)}{P(A_i)}.$$

Rozdělení pravděpodobností vyslaných znaků je

$$\begin{bmatrix} P(B_1) & P(B_2) & \cdots & P(B_m) \end{bmatrix} =$$

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

pokud $k = m$ a příslušná inverzní matice existuje.

Motivační příklad (alkohol za volantem – pokračování):

90 % všech nehod způsobili střízliví řidiči.

99 % řidičů bylo střízlivých.

Označme jevy

A ... požil alkohol, $P(A) = 0.01$,

H ... způsobil nehodu, $P(A|H) = 0.1$.

$$0.1 = P(A|H) = \frac{P(A)\cdot P(H|A)}{P(A)\cdot P(H|A) + P(\bar{A})\cdot P(H|\bar{A})} =$$
$$= \frac{0.01\cdot P(H|A)}{0.01\cdot P(H|A) + 0.99\cdot P(H|\bar{A})} = \frac{1}{1 + 99\cdot \dfrac{P(H|\bar{A})}{P(H|A)}}\,.$$

Požitím alkoholu se zvyšuje 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

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

(Neuvažovali jsme, že účastníků nehody bývá víc a přitomnost alkoholu u jejích účastníků nemusí být nezávislá.)
Příklad: Detektor v CERNu zachytí *událost* (současný vznik Higgsova bosonu a páru top kvarků) s účinností 0.01;
falešně pozitivní je s pravděpodobností 0.000 05.
Hledaná událost se vyskytne v 6 případech z 10^{12}. Jaká je pravděpodobnost, že detekovaná událost skutečně nastala? [1]
Označme jevy
U ... událost nastala, $\quad P(U) = 6\cdot 10^{-12}$,
D ... detektor hlásí událost, $\quad P(D|U) = 0.01, \quad P(D|\bar{U}) = 5\cdot 10^{-5}$.

$$P(D) = P(U)\cdot P(D|U) + P(\bar{U})\cdot P(D|\bar{U}) =$$
$$= 6\cdot 10^{-12}\cdot 0.01 + (1 - 6\cdot 10^{-12})\cdot 5\cdot 10^{-5} = 5.000000006\cdot 10^{-5}\,,$$

$$P(U|D) = \frac{P(U)\cdot P(D|U)}{P(D)} \doteq \frac{6\cdot 10^{-14}}{5\cdot 10^{-5}} = 1.2\cdot 10^{-9}\,.$$

Příklad. *(Sally Clark — pokračování): Jaká je (podmíněná) pravděpodobnost syndromu náhlého úmrtí kojenců v rodině, kde se už vyskytl?*
Není dostatek dat pro kvalifikovaný odhad.
Alespoň jsou známy vlivy, které (podmíněné) riziko snižují (nekuřácká rodina, stabilní, dobře zajištěná).
Tytéž vlivy snižují i riziko vraždy novorozence, takže ***podmíněná*** *pravděpodobnost, že šlo o trestný čin, zůstává nízká.*

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 n výchozích 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, popsaných $2^n - 1$ parametry (součet jejich pravděpodobností 1).*
B. Stačí n pravděpodobností výchozích jevů.

[1] André Sopczak: Recognition and categorization of events resulting from proton collisions at CERN. Přednáška na katedře kybernetiky FEL ČVUT, 22. 1. 2019

n	2	3	4	5	10	16	30
$2^n - 1$	3	7	15	31	1023	65 535	1 073 741 823

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*

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

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 : 1 000. Adekvátní cena ročního pojistného (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 pojistné 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 $(\Omega, \mathcal{A}, P)$ je **měřitelná funkce** $X\colon \Omega \to \mathbb{R}$, tj. taková, že pro každý interval I platí

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

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 σ-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:
$P_X(\mathbb{R}) = 1$,
$P_X\Big(\bigcup\limits_{n\in\mathbb{N}} I_n\Big) = \sum\limits_{n\in\mathbb{N}} P_X(I_n)$, pokud jsou množiny I_n, $n \in \mathbb{N}$, po dvou disjunktní,
$P_X(\emptyset) = 0$, $\quad P_X(\mathbb{R} \setminus I) = 1 - P_X(I)$,
$I \subseteq J \Rightarrow P_X(I) \le P_X(J)$, $P_X(J \setminus I) = P_X(J) - P_X(I)$.
Popisy náhodné veličiny

prostor elementárních jevů	Ω	$\mathbb{R}$
σ-algebra jevů	$\mathcal{A}$	$\mathcal{B}(\mathbb{R})$
pravděpodobnostní míra	P	P_X
pravděpodobnostní prostor	$(\Omega, \mathcal{A}, P)$	$(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$
náhodná proměnná	$X\colon \Omega \to \mathbb{R}\,, \quad \omega \mapsto X(\omega)$	$\mathrm{id}\colon \mathbb{R} \to \mathbb{R}\,, \quad x \mapsto x$
$P(X \in I)$	$P(\{\omega \in \Omega \mid X(\omega) \in I\})$	$P_X(I)$

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

prostor elementárních jevů	$\Omega = \{1,2,3,4,5,6\}$	$\mathbb{R}$
σ-algebra jevů	$\mathcal{A} = \exp\Omega$	$\mathcal{B}(\mathbb{R})$
pravděpodobnostní míra	$P(A) = \dfrac{\lvert A\rvert}{6}$	$P_X(I) = \begin{cases} 1\,, & 0,1 \in I \\ 5/6\,, & 0 \in I\,, 1 \notin I \\ 1/6\,, & 0 \notin I\,, 1 \in I \\ 0\,, & 0,1 \notin I \end{cases}$
náhodná proměnná	$X(\omega) = \begin{cases} 1\,, & \omega = 6 \\ 0\,, & \text{jinak} \end{cases}$	$X\colon x \mapsto x$

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

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

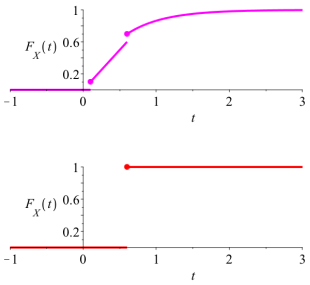
$F_X : \mathbb{R} \to \langle 0, 1\rangle$ je **distribuční funkce** náhodné veličiny X. Ta stačí, neboť

$$\begin{array}{ll} (a,b\rangle = (-\infty,b\rangle \setminus (-\infty,a\rangle\,, & P_X((a,b\rangle) = P(a < X \le b) = F_X(b) - F_X(a)\,, \\ (a,\infty) = \mathbb{R} \setminus (-\infty,a\rangle\,, & P_X((b,\infty)) = 1 - F_X(b)\,, \\ \{b\} = \bigcap\limits_{n\to\infty} (b-\frac{1}{n},b\rangle & P_X(\{a\}) = P(X=a) = \lim\limits_{n\to\infty}(F_X(b) - F_X(b-\frac{1}{n}) \\ \quad = \lim\limits_{n\to\infty}(b-\frac{1}{n},b)\rangle & \qquad = F_X(b) - \lim\limits_{a\to b-} F_X(a)\,, \\ \ldots & \ldots \end{array}$$

Vlastnosti distribuční funkce:

- neklesající,
- zprava spojitá,
- $\lim\limits_{t\to-\infty} F_X(t) = 0, \quad \lim\limits_{t\to\infty} F_X(t) = 1.$

Věta: Tyto podmínky jsou nejen **nutné**, ale i **postačující**.



Příklad: Reálnému číslu r odpovídá náhodná veličina (značená též r) s **Diracovým** rozdělením v r:

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

(F_r je posunutá Heavisideova funkce.)

Tvrzení: $X \le Y \Rightarrow F_X \ge F_Y$,
neboli $\big(\forall\omega \in \Omega : X(\omega) \le Y(\omega)\big) \Rightarrow \big(\forall t \in \mathbb{R} : F_X(t) \ge F_Y(t)\big)$.

Důkaz.

$$\begin{gathered} \forall t \in \mathbb{R} : \big(Y(\omega) \le t \Rightarrow X(\omega) \le Y(\omega) \le t\big)\,, \\ \{\omega \in \Omega : Y(\omega) \le t\} \subseteq \{\omega \in \Omega : X(\omega) \le t\}\,, \\ F_Y(t) = P(Y \le t) = P(\{\omega \in \Omega : Y(\omega) \le t\}) \le P(\{\omega \in \Omega : X(\omega) \le t\}) = P(X \le t) = F_X(t)\,. \end{gathered}$$

□

Příklad: V zimním období je venkovní teplota, X, nižší než vnitřní, Y. Např. pro $t = 0$ tvrdíme, že pravděpodobnost, že mrzne uvnitř, je menší než pravděpodobnost, že mrzne venku.

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 $(-\infty, t\rangle$, tj.

$$P(X_1 \le t_1, X_2 \le t_2) = P(X_1 \le t_1) \cdot P(X_2 \le t_2) = F_{X_1}(t_1) \cdot F_{X_2}(t_2)$$

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

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

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 $(\Omega_1, \mathcal{A}_1, P_1)$, resp. $(\Omega_2, \mathcal{A}_2, P_2)$, přičemž $\Omega_1 \cap \Omega_2 = \emptyset$.
Nechť $c \in \langle 0, 1\rangle$.
Definujeme nový pravděpodobnostní prostor $(\Omega, \mathcal{A}, P)$, kde
$\Omega = \Omega_1 \cup \Omega_2, \qquad \mathcal{A} = \{A_1 \cup A_2 \mid A_1 \in \mathcal{A}_1,\ A_2 \in \mathcal{A}_2\}$,
$P(A_1 \cup A_2) = c\,P_1(A_1) + (1-c)\,P_2(A_2)$ pro $A_1 \in \mathcal{A}_1,\ A_2 \in \mathcal{A}_2$.
Definujeme funkci $X\colon \Omega \to \mathbb{R}$:

$$X(\omega) = \begin{cases} V(\omega) & \text{pro } \omega \in \Omega_1\,, \\ U(\omega) & \text{pro } \omega \in \Omega_2\,. \end{cases}$$

X je náhodná veličina na $(\Omega, \mathcal{A}, P)$.
X nazýváme **směs náhodných veličin** V, U **s koeficientem** c (angl. *mixture*), značíme $\mathrm{Mix}_c(V, U)$. 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 obecněji **směs náhodných veličin** $V_1, \ldots, V_n$ **s koeficienty** $c_1, \ldots, c_n \in \langle 0, 1\rangle$, $\sum_{i=1}^{n} c_i = 1$, značíme $\mathrm{Mix}_{(c_1,\ldots,c_n)}(V_1, \ldots, V_n) = \mathrm{Mix}_{\boldsymbol{c}}(V_1, \ldots, V_n)$, kde $\boldsymbol{c} = (c_1, \ldots, c_n)$. 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 $\boldsymbol{c} = (c_1, \ldots, c_n)$. 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 $\mathrm{Mix}_{(c,1-c)}(V, U) = \mathrm{Mix}_c(V, U)$ (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 $X = \mathrm{Mix}_{(c_1,\ldots,c_n)}(r_1, \ldots, r_n)$,

$$P_X(I) = P(X \in I) = \sum_{i:\, r_i \in I} c_i, \qquad F_X(t) = \sum_{i:\, r_i \le t} c_i\,.$$

Lze ji popsat též **pravděpodobnostní funkcí** $p_X : \mathbb{R} \to \langle 0, 1\rangle$,

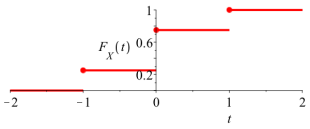
$$p_X(t) = P_X(\{t\}) = P(X = t) = \begin{cases} c_i & \text{pro } t = r_i\,, \\ 0 & \text{jinak} \end{cases}$$

(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
$\Omega_X = \{t \in \mathbb{R} : P_X(\{t\}) \neq 0\} = \{t \in \mathbb{R} : P(X = t) \neq 0\}$.

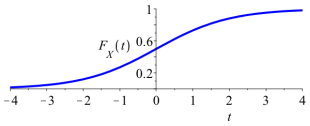


Diskrétní náhodnou veličinu popisuje **pravděpodobnostní funkce** $p_X(t) = P_X(\{t\}) = P(X = t)$.
Splňuje

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

4.4.2 Spojité náhodné veličiny

Mají spojitou distribuční funkci.



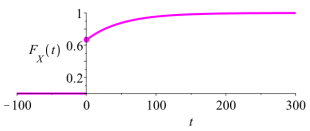
Náhodná veličina X je **absolutně spojitá**, jestliže existuje nezáporná funkce $f_X : \mathbb{R} \to \langle 0, \infty)$ (**hustota** náhodné veličiny X) taková, že

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

Hustota splňuje $\int_{-\infty}^{\infty} f_X(u)\,\mathrm{d}u = 1$.
Není určena jednoznačně, ale dvě hustoty f_X, g_X téže náhodné veličiny splňují
$\int_I (f_X(x) - g_X(x))\,\mathrm{d}x = 0$ pro všechny intervaly I.
Lze volit $f_X(t) = \frac{\mathrm{d}F_X(t)}{\mathrm{d}t}$, 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 neuvažujeme.

4.4.3 Smíšené náhodné veličiny

Motivační příklad: Náhodné veličiny jako množství srážek, spotřeba elektřiny apod. je vhodné modelovat spojitým rozdělením, jsou-li nenulové, nicméně nulové mohou být s nezanedbatelnou nenulovou pravděpodobností.



Podobně i u dalších náhodných veličin se mohou uplatnit meze či jiné preferované hodnoty, nastávající s nenulovou pravděpodobností.

Směs předchozích dvou případů;
$\Omega_X \neq \emptyset$, $P_X(\mathbb{R} \setminus \Omega_X) = P(X \notin \Omega_X) \neq 0$.
Nelze je popsat ani pravděpodobnostní funkcí (existuje, ale neurčuje celé rozdělení) ani hustotou (neexistuje, nevychází konečná).
Každou náhodnou veličinu se smíšeným rozdělením lze **jednoznačně** vyjádřit ve tvaru $X = \mathrm{Mix}_c(V, U)$, kde V je spojitá, U je diskrétní a $c \in (0, 1)$:
Nespojitostí je spočetně 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,

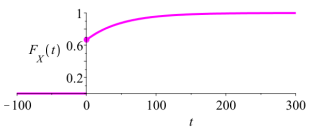
$$X = \mathrm{Mix}_{(c, c_1, c_2, \ldots)}(V, r_1, r_2, \ldots)\,,$$
$$U = \underbrace{\frac{1}{\sum_n c_n}}_{1-c} \mathrm{Mix}_{(c_1, c_2, \ldots)}(r_1, r_2, \ldots)\,,$$

je diskrétní složka náhodné veličiny X,

$$X = \mathrm{Mix}_{(c, 1-c)}(V, U)\,.$$

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

$$F_X(t) = \begin{cases} 1 - \frac{1}{3}\exp(-\frac{t}{50}), & t \geq 0, \\ 0 & \text{jinak.} \end{cases}$$



Po 2/3 dní neprší, diskrétní složka je $U = 0$ s váhou $c_1 := 2/3$.
Motivační příklad (dešťové srážky):
Srážkový úhrn v mm za 24 hodin má rozdělení s distribuční funkcí

$$F_X(t) = \begin{cases} 1 - \frac{1}{3}\exp(-\frac{t}{50}), & t \geq 0, \\ 0 & \text{jinak.} \end{cases}$$

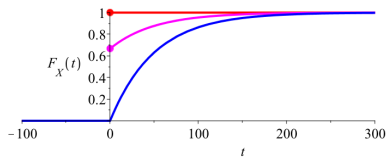
Po 2/3 dní neprší, diskrétní složka je $U = 0$ s váhou $c_1 := 2/3$.
Spojitá složka V má váhu $c := 1/3$, distribuční funkci

$$F_V(t) = \begin{cases} 1 - \exp(-\frac{t}{50}), & t \geq 0, \\ 0 & \text{jinak} \end{cases}$$

a hustotu

$$f_V(t) = \begin{cases} \frac{1}{50} \exp(-\frac{t}{50}), & t \geq 0, \\ 0 & \text{jinak} \end{cases}$$

(exponenciální rozdělení).



4.4.4 Směsi náhodných veličin stejného typu

$$X = \mathrm{Mix}_{(c,1-c)}(V, U)\,.$$

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 směsi 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. *(Předpokládáme, že lepšímu prospěchu odpovídá nižší průměr známek.)*
Neostrá nerovnost v definici 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

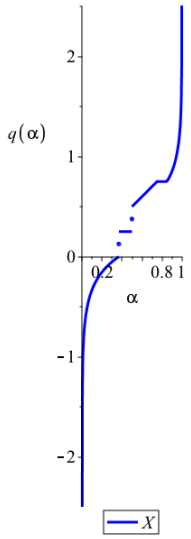
$$P(X < t) \leq \alpha \leq P(X \leq t)\,,$$

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,

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



Číslo $q_X(\alpha)$ se nazývá **α-kvantil** náhodné veličiny X a funkce $q_X\colon (0,1) \to \mathbb{R}$ je **kvantilová funkce** náhodné veličiny X. Speciálně $q_X(\frac{1}{2})$ je **medián**, další kvantily 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,
- $q_X(\alpha) = \frac{1}{2}\left(\lim\limits_{\beta\to\alpha-} q_X(\beta) + \lim\limits_{\beta\to\alpha+} q_X(\beta)\right)$.

Věta: Tyto podmínky jsou **nutné** i **postačující**.

Obrácený převod:

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

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 $t_k,\ k = 1,\dots,n$, 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 (spočetně) nekonečně mnoha hodnot, musíme některé vynechat, zejména ty, které jsou málo pravděpodobné. Pro každé $\varepsilon > 0$ lze vybrat konečně mnoho hodnot $t_k,\ k = 1,\dots,n$, tak, že $P_X(\mathbb{R}\setminus\{t_1,\ \dots,t_n\}) = P(X \notin \{t_1,\ \dots,t_n\}) \le \varepsilon$. 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 $t_k,\ k = 1,\dots,n$, ale jen za předpokladu, že je „dostatečně hladká“. Zajímají nás z ní spíše integrály typu

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

   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 hustě.

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

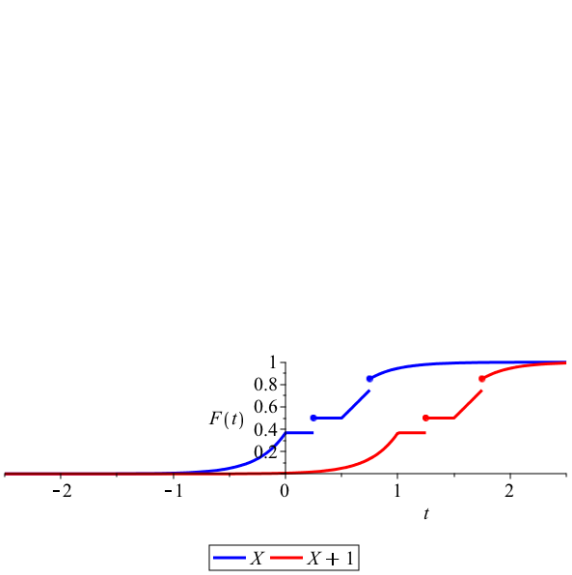
Mnohé obecnější případy se snažíme vyjádřit alespoň jako směsi náhodných veličin s rozděleními známého typu, abychom vystačili s konečně mnoha parametry.

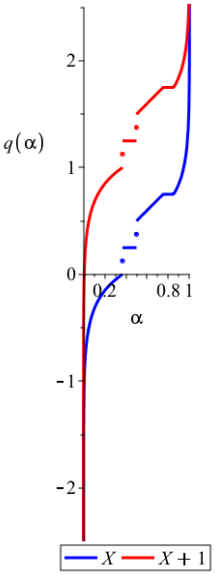
3. **Smíšená**: Jako u spojité náhodné veličiny. Tento popis je však pro diskrétní část zbytečně nepřesný. Můžeme použít rozklad na diskrétní a spojitou část.

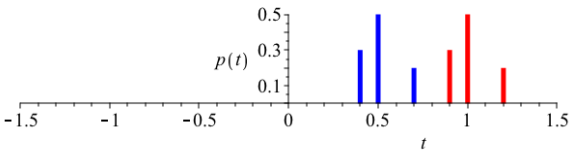
4.7 Operace s náhodnými veličinami

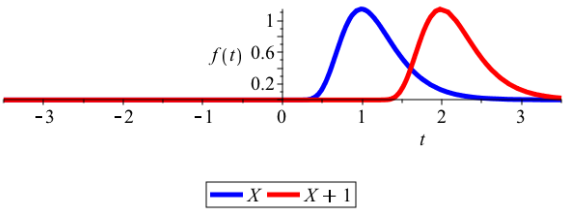
Zde $I, J \subseteq \mathbb{R}$ jsou intervaly nebo spočetná sjednocení intervalů.
Přičtení konstanty r odpovídá posunutí ve směru vodorovné osy:

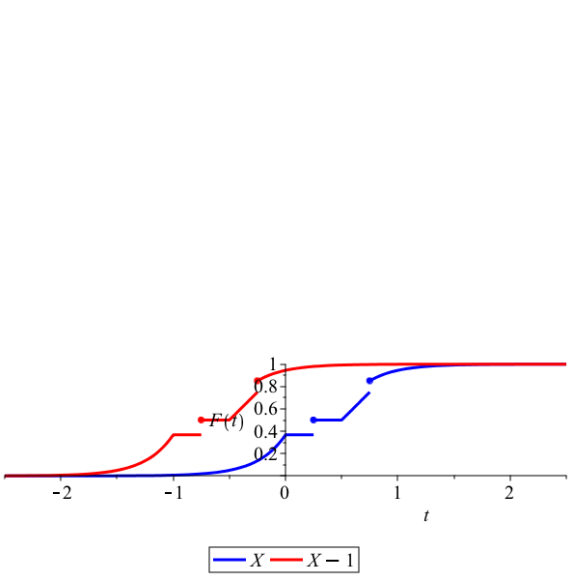
$$\begin{aligned} P_{X+r}(I+r) &= P_X(I)\,, & P_{X+r}(J) &= P_X(J-r)\,,\\ F_{X+r}(t+r) &= F_X(t)\,, & F_{X+r}(u) &= F_X(u-r)\,,\\ q_{X+r}(\alpha) &= q_X(\alpha) + r\,. \end{aligned}$$

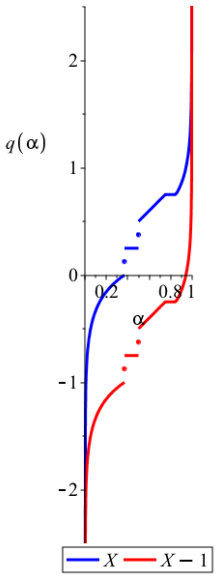


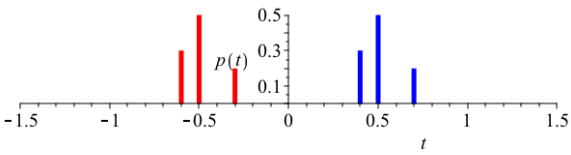


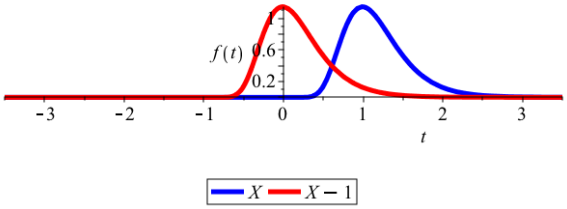










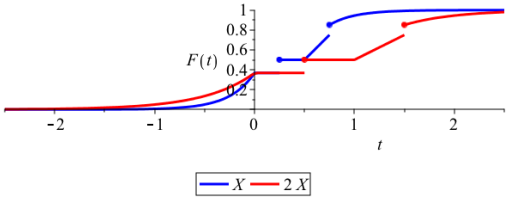


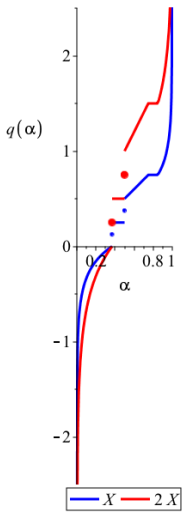
Vynásobení nenulovou konstantou r odpovídá podobnost ve směru vodorovné osy.

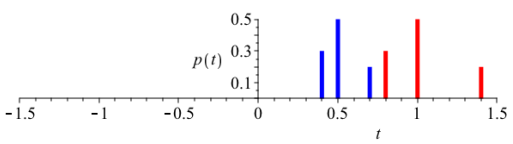
$$P_{rX}(rI) = P_X(I), \qquad P_{rX}(J) = P_X(\tfrac{J}{r}).$$

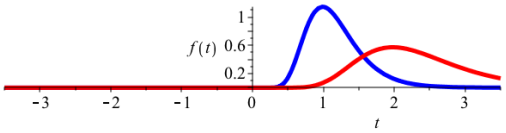
Pro distribuční funkci musíme rozlišit případy:

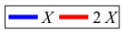
- $r > 0$: $\quad F_{rX}(rt) = F_X(t), \qquad F_{rX}(u) = F_X(\tfrac{u}{r}), \qquad q_{rX}(\alpha) = r\, q_X(\alpha),$

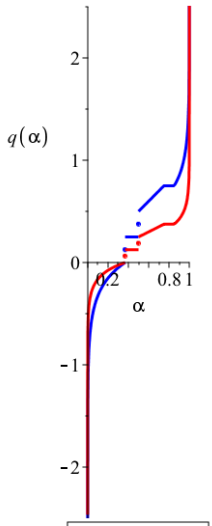


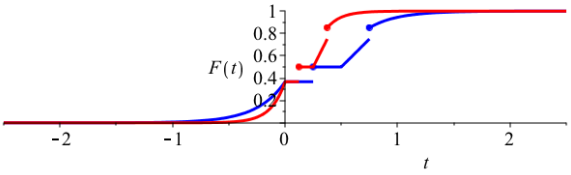


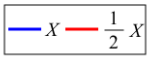


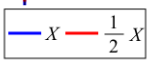


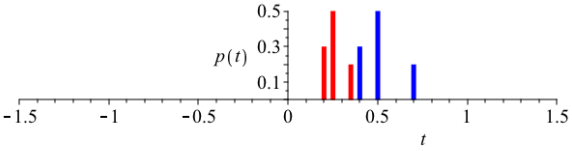


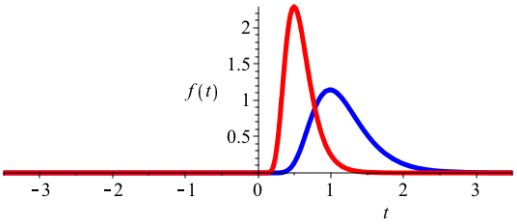


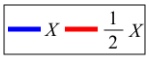




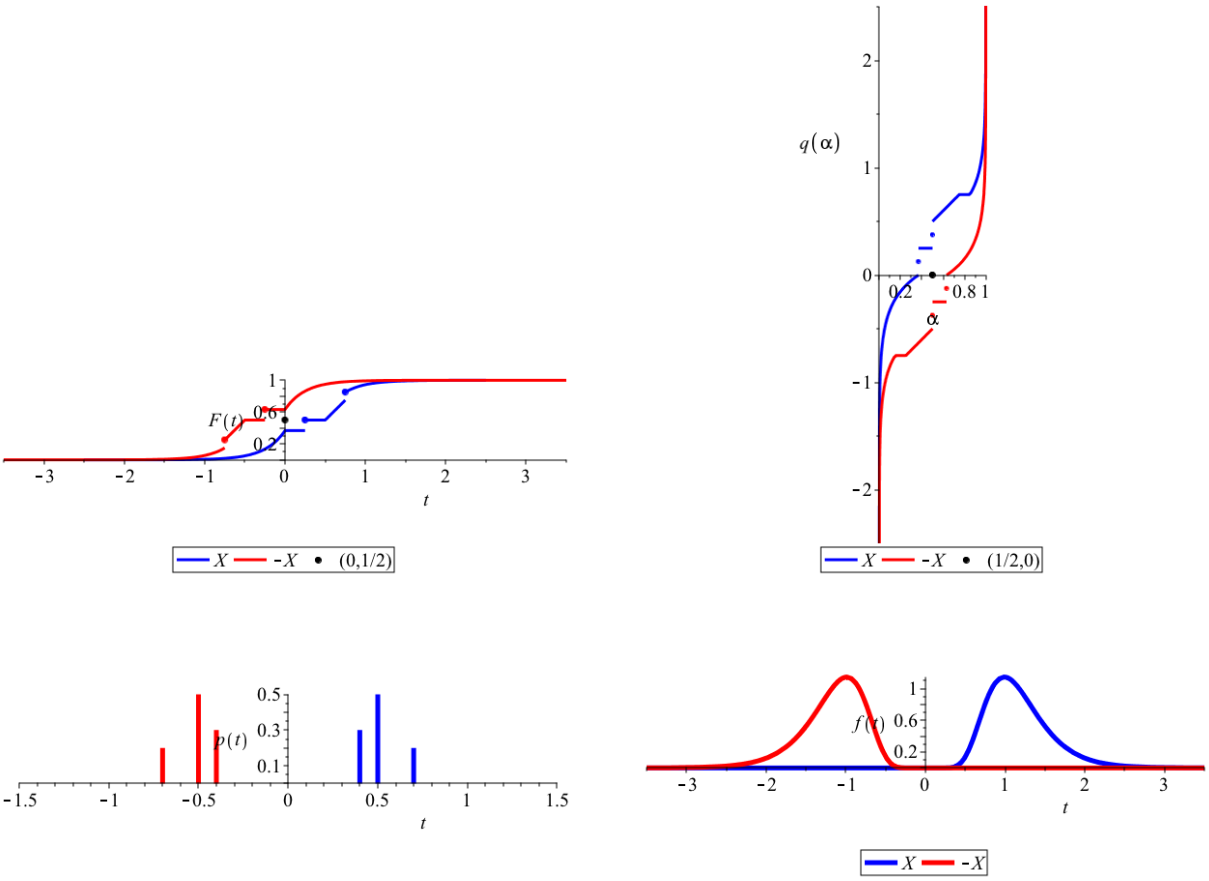








- $r = -1$: $F_{-X}(-t) = P(-X \le -t) = P(X \ge t) = 1 - P(X < t)$, **v bodech spojitosti** distribuční funkce
  $F_{-X}(-t) = 1 - P(X < t) = 1 - P(X \le t) = 1 - F_X(t)$,
  $F_{-X}(u) = 1 - F_X(-u)$, v bodech nespojitosti limita zprava
  (středová symetrie grafu podle bodu $(0, \frac{1}{2})$ s opravou na spojitost zprava),
  $q_{-X}(\alpha) = -q_X(1-\alpha)$.



- $r < 0$: kombinace předchozích případů.

Zobrazení spojitou rostoucí funkcí h:
$P_{h(X)}(h(I)) = P_X(I), \qquad F_{h(X)}(h(t)) = F_X(t), \qquad F_{h(X)}(u) = F_X(h^{-1}(u)),$
$q_{h(X)}(\alpha) = h(q_X(\alpha))$ **v bodech spojitosti kvantilové funkce**.
Zobrazení neklesající funkcí h: $F_{h(X)}(u) = \sup\{F_X(t) \mid h(t) \leq u\}$.
Zobrazení nerostoucí funkcí h lze řešit jako zobrazení náhodné veličiny $-X$ neklesající funkcí $g(t) = h(-t)$.
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 rozděleními vstupních náhodných veličin a koeficienty směsi,*

$$h(\mathrm{Mix}_c(U, V)) = \mathrm{Mix}_c(h(U), h(V))\,.$$

(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 $\langle 0, 1\rangle$. („*Kolo štěstí.*“)
2. Náhodná veličina $q_Y(X) \sim Y$ (kde $\sim$ znamená „má stejné rozdělení“). Stačí na každou realizaci náhodné veličiny X aplikovat funkci q_Y.
(„*Na kolo štěstí napíšeme vzestupně výsledky, každý s šířkou odpovídající jeho pravděpodobnosti.*“)

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.
Místo mnoha hodnot chceme jednu „průměrnou“, která charakterizuje celou populaci.
Značení: E. nebo $\mu_\cdot$

$$\mathrm{E}X = \int_0^1 q_X(\alpha)\,\mathrm{d}\alpha\,, \qquad \text{(pokud existuje)}.$$

Interpretace:

- Pro Diracovo rozdělení integrujeme konstantu přes interval délky 1.
- Pro diskrétní náhodnou veličinu U integrujeme po částech konstantní funkci; kde délka každého úseku je pravděpodobnost příslušné hodnoty,

$$\mathrm{E}U = \sum_{t\in\mathbb{R}} t\cdot p_U(t) = \sum_{t\in\Omega_U} t\cdot p_U(t)\,.$$

- Obecný případ je limitou předchozího, pokud „diskretizační chyba“ $\to 0$.

Střední hodnota může být definována zvlášť pro

- **diskrétní** náhodnou veličinu U: $\mathrm{E}U = \sum_{t\in\mathbb{R}} t\cdot p_U(t)\,,$
- **spojitou** náhodnou veličinu V: $\mathrm{E}V = \int_{-\infty}^{\infty} t\cdot f_V(t)\,\mathrm{d}t\,,$
- **směs** náhodných veličin $X = \mathrm{Mix}_c(V,U)$: $\mathrm{E}X = c\,\mathrm{E}V + (1-c)\,\mathrm{E}U\,.$
  (Může být V diskrétní, U spojitá. To **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).
Vzorec používající kvantilovou funkci lze zobecnit na střední hodnotu jakékoli (měřitelné) funkce h náhodné veličiny:

$$\mathrm{E}(h(X)) = \int_0^1 h(q_X(\alpha))\,\mathrm{d}\alpha\,.$$

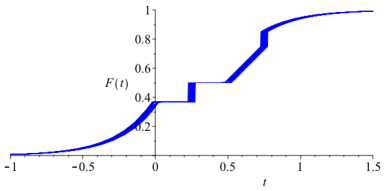
Speciálně pro **diskrétní** náhodnou veličinu U

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

pro **spojitou** náhodnou veličinu V

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

(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

$$\begin{aligned} \mathrm{E}r &= r\,, && \text{speciálně} && \mathrm{E}(\mathrm{E}X) &= \mathrm{E}X\,,\\ \mathrm{E}(X+Y) &= \mathrm{E}X + \mathrm{E}Y\,, && \text{speciálně} && \mathrm{E}(X+r) &= \mathrm{E}X + r\,,\\ \mathrm{E}(X-Y) &= \mathrm{E}X - \mathrm{E}Y\,, && && && \\ \mathrm{E}(r\,X) &= r\,\mathrm{E}X\,, && \text{obecněji} && \mathrm{E}(r\,X + s\,Y) &= r\,\mathrm{E}X + s\,\mathrm{E}Y\,. \end{aligned}$$

*(To **je** linearita střední hodnoty.)*

$$\mathrm{E}(\mathrm{Mix}_c(V,U)) = c\,\mathrm{E}V + (1-c)\,\mathrm{E}U\,.$$

*(To **není** linearita střední hodnoty.)*

Pouze pro **nezávislé** náhodné veličiny

$$\mathrm{E}(X \,\tilde{\cdot}\, Y) = \mathrm{E}X \cdot \mathrm{E}Y\,,$$

kde $\tilde{\cdot}$ značí součin nezávislých náhodných veličin.

5.2 Rozptyl (disperze)

Motivační příklad (stejné podnebí):
Chceme najít místo s podobným podnebím.
Průměrná teplota nestačí. Důležité je i kolísání teplot.
Značení: $\sigma^2_{\cdot}$, $\mathrm{D}.$, $\mathrm{var}\,.$

$$\begin{aligned} \mathrm{D}X = \mathrm{E}((X-\mathrm{E}X)^2) &= \mathrm{E}(X^2) - (\mathrm{E}X)^2\,,\\ \mathrm{E}(X^2) &= (\mathrm{E}X)^2 + \mathrm{D}X\,. \end{aligned} \tag{1}$$

Vlastnosti:

$$\mathrm{D}X = \int_0^1 (q_X(\alpha) - \mathrm{E}X)^2\,\mathrm{d}\alpha\,.$$

$$\begin{aligned} \mathrm{D}X &\geq 0\,,\\ \mathrm{D}r &= 0\,,\\ \mathrm{D}(X+r) &= \mathrm{D}X\,,\\ \mathrm{D}(r\,X) &= r^2\,\mathrm{D}X\,. \end{aligned}$$

$$\begin{aligned} \mathrm{D}(\mathrm{Mix}_c(V,U)) &= \mathrm{E}(\mathrm{Mix}_c(V,U)^2) - (\mathrm{E}(\mathrm{Mix}_c(V,U)))^2\\ &= c\,\mathrm{E}(V^2) + (1-c)\,\mathrm{E}(U^2) - (c\,\mathrm{E}V + (1-c)\,\mathrm{E}U)^2\\ &= c\,(\mathrm{D}V + (\mathrm{E}V)^2) + (1-c)\,(\mathrm{D}U + (\mathrm{E}U)^2)\\ &\quad - (c^2\,(\mathrm{E}V)^2 + 2\,c\,(1-c)\,\mathrm{E}V\,\mathrm{E}U + (1-c)^2\,(\mathrm{E}U)^2)\\ &= c\,\mathrm{D}V + (1-c)\,\mathrm{D}U + c\,(1-c)\,(\mathrm{E}V)^2\\ &\quad - 2\,c\,(1-c)\,\mathrm{E}V\,\mathrm{E}U + c\,(1-c)\,(\mathrm{E}U)^2\\ &= c\,\mathrm{D}V + (1-c)\,\mathrm{D}U + c\,(1-c)\,(\mathrm{E}V - \mathrm{E}U)^2\,. \end{aligned}$$

Pouze pro **nezávislé** náhodné veličiny ($\tilde{+}, \tilde{-}$ značí součet a rozdíl nezávislých)

$$\mathrm{D}(X \,\tilde{+}\, Y) = \mathrm{D}X + \mathrm{D}Y, \qquad \mathrm{D}(X \,\tilde{-}\, Y) = \mathrm{D}X + \mathrm{D}Y\,.$$

5.3 Směrodatná odchylka

Značení: σ.
Má **stejný fyzikální rozměr** jako původní náhodná veličina (rozptyl nikoli).

$$\sigma_X = \sqrt{\mathrm{D}X} = \sqrt{\mathrm{E}((X - \mathrm{E}X)^2)}$$

Vlastnosti:

$$\sigma_X = \sqrt{\int_0^1 (q_X(\alpha) - \mathrm{E}X)^2 \,\mathrm{d}\alpha}.$$

$$\begin{aligned} \sigma_X &\geq 0\,, \\ \sigma_r &= 0\,, \\ \sigma_{X+r} &= \sigma_X\,, \\ \sigma_{r\,X} &= |r|\,\sigma_X\,. \end{aligned}$$

Pouze pro **nezávislé** náhodné veličiny

$$\sigma_{X \pm Y} = \sigma_{X - Y} = \sqrt{\mathrm{D}X + \mathrm{D}Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}\,.$$

5.4 Obecné a centrální momenty

$k \in \mathbb{N}$
k-tý **obecný moment** *(značení nezavádíme)*: $\mathrm{E}(X^k)$, speciálně:
pro $k = 1$: $\mathrm{E}X$,
pro $k = 2$: $\mathrm{E}(X^2) = (\mathrm{E}X)^2 + \mathrm{D}X$.
Alternativní značení: m_k, μ'_k.

k-tý **centrální moment** *(značení nezavádíme)*: $\mathrm{E}((X - \mathrm{E}X)^k)$, speciálně:
pro $k = 1$: 0,
pro $k = 2$: $\mathrm{D}X$.
Alternativní značení: μ_k.

Pomocí kvantilové funkce:

$$\mathrm{E}(X^k) = \int_0^1 (q_X(\alpha))^k \,\mathrm{d}\alpha\,.$$

$$\mathrm{E}((X - \mathrm{E}X)^k) = \int_0^1 (q_X(\alpha) - \mathrm{E}X)^k \,\mathrm{d}\alpha\,.$$

5.5 Normovaná náhodná veličina

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

$$\mathrm{norm}\, X = \frac{X - \mathrm{E}X}{\sigma_X}$$

(pokud má vzorec smysl). Zpětná transformace je

$$X = \mathrm{E}X + \sigma_X\ \mathrm{norm}\, X\,.$$

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.

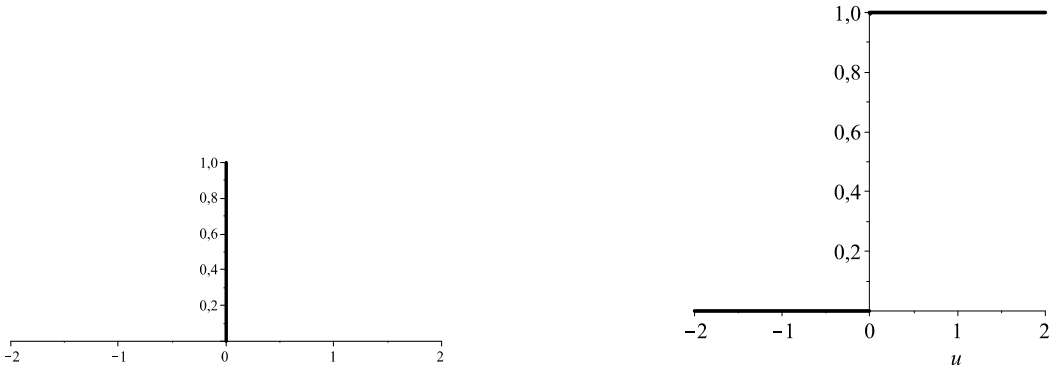
5.6 Základní typy diskrétních rozdělení

5.6.1 Diracova

s parametrem $r \in \mathbb{R}$, kterým je jediný možný výsledek,

$$p_X(r) = 1\,, \qquad \mathrm{E}X = r\,, \qquad \mathrm{D}X = 0\,.$$

Všechna diskrétní rozdělení jsou směsi Diracových rozdělení.

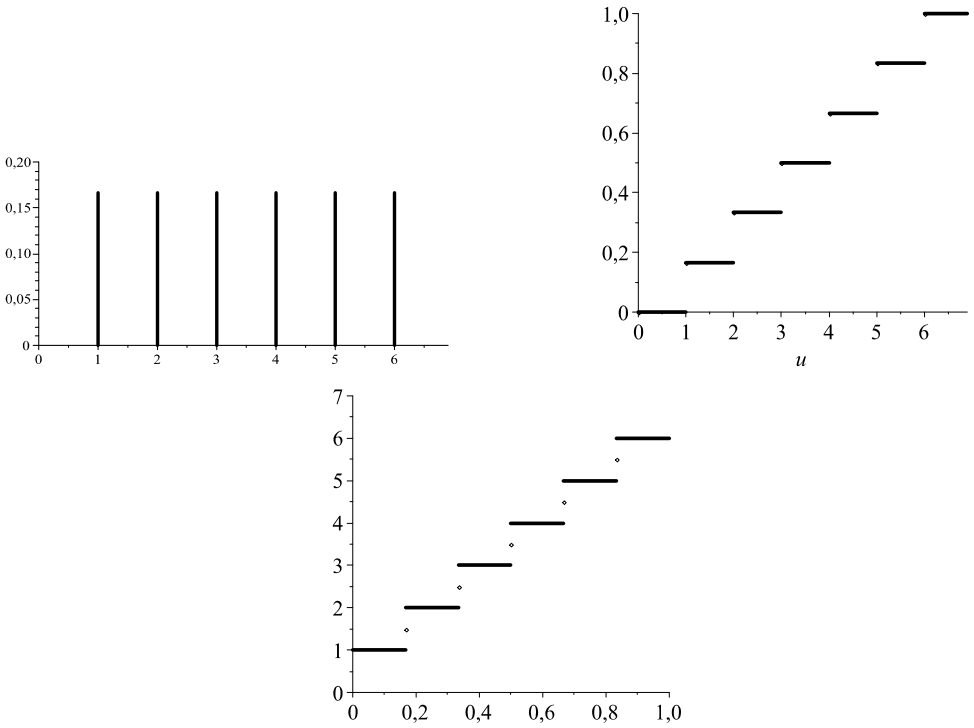


Pravděpodobnostní a distribuční funkce Diracova rozdělení pro $r = 0$

5.6.2 Rovnoměrná

s parametrem $m \in \mathbb{N}$, kterým je počet možných výsledků se stejnou pravděpodobností $1/m$.
Speciálně pro obor hodnot $\{1, 2, \ldots, m\}$ dostáváme

$$p_X(k) = \frac{1}{m}\,, \qquad k \in \{1, 2, \ldots, m\}\,,$$
$$\mathrm{E}X = \frac{m+1}{2}\,, \qquad \mathrm{D}X = \frac{1}{12}\,(m+1)\,(m-1)\,.$$

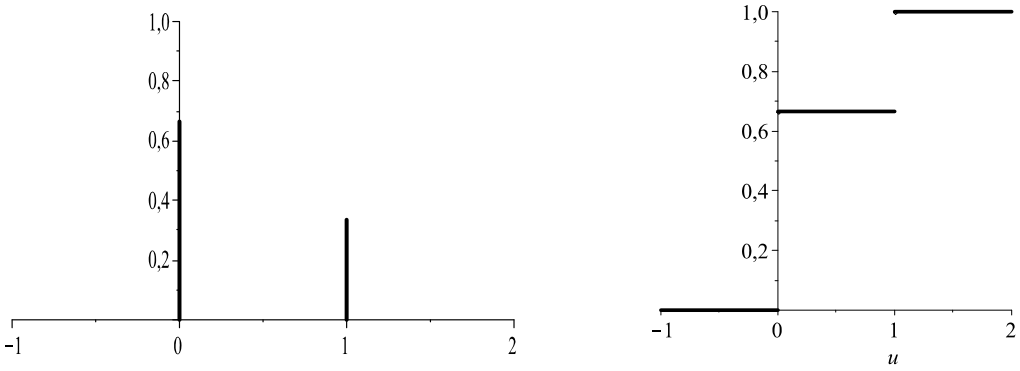


Pravděpodobnostní a distribuční funkce rovnoměrného rozdělení na $\{1, 2, \ldots, 6\}$

5.6.3 Alternativní (Bernoulliova) $\mathrm{Alt}(q)$

s parametrem $q \in (0,1)$, kterým je pravděpodobnost výsledku 1, s pravděpodobností $1-q$ je výsledek 0, $\mathrm{Alt}(q) = \mathrm{Mix}_q(1,0)$,

$$p_X(1) = q\,, \qquad p_X(0) = 1-q\,,$$
$$\mathrm{E}X = q\,, \qquad \mathrm{D}X = q\,(1-q)\,.$$



Pravděpodobnostní a distribuční funkce alternativního rozdělení pro $q = 1/3$

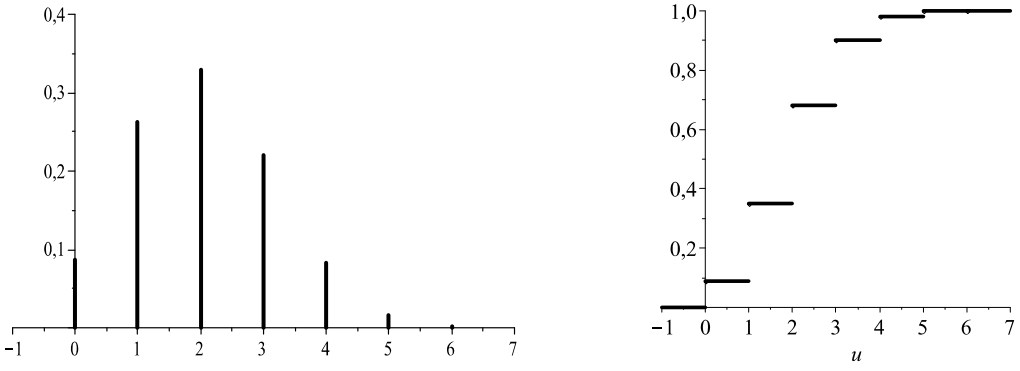
5.6.4 Binomická $\mathrm{Bi}(m,q)$

s parametry $m \in \mathbb{N}$, $q \in (0,1)$, jsou součtem m nezávislých náhodných veličin s alternativními rozděleními $\mathrm{Alt}(q)$, $\mathrm{Bi}(m,q) = \underbrace{\mathrm{Alt}(q) \tilde{+} \ldots \tilde{+} \mathrm{Alt}(q)}_{m\times}$,

$\mathrm{Bi}(1,q) = \mathrm{Alt}(q)$,

$$p_X(k) = \binom{m}{k} q^k\,(1-q)^{m-k}\,, \qquad k \in \{0,1,2,\ldots,m\}\,,$$
$$\mathrm{E}X = m\,q\,, \qquad \mathrm{D}X = m\,q\,(1-q)\,.$$

Součet **nezávislých** náhodných veličin s rozděleními $\mathrm{Bi}(m,q)$, $\mathrm{Bi}(n,q)$ má binomické rozdělení $\mathrm{Bi}(m+n,q)$.
Parametry by mohly být též m a $\lambda = m\,q = \mathrm{E}X$, značíme $\mathrm{Bi}^*(m,\lambda) = \mathrm{Bi}(m,q)$.
Výpočetní složitost výpočtu $p_X(k)$ je $O(k)$, celého rozdělení $O(m^2)$.



Pravděpodobnostní a distribuční funkce binomického rozdělení $\mathrm{Bi}\,(6,1/3)$

5.6.5 Poissonova $\mathrm{Po}(\lambda)$

s parametrem $\lambda \in (0,\infty)$ jsou limitní případy binomických rozdělení $\mathrm{Bi}(m,q_m) = \mathrm{Bi}^*(m,\lambda)$ pro $m \to \infty$ při konstantním $\lambda = m\,q_m > 0$ (tedy $q_m \to 0$),
$\mathrm{Po}(\lambda) = \lim_{m\to\infty} \mathrm{Bi}^*(m,\lambda)$.

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

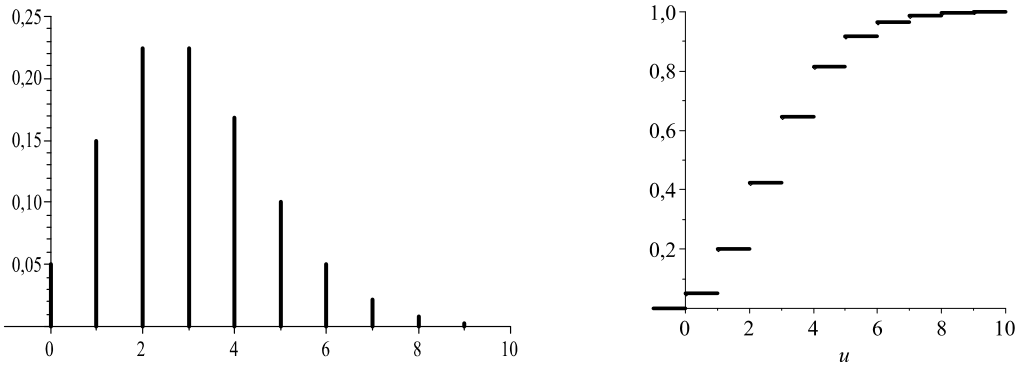
hodnota	0	1	2	3	4	5	6
$\mathrm{Bi}^*(30,3) = \mathrm{Bi}(30,0.1)$	0.042	0.141	0.228	0.236	0.177	0.102	0.047
$\mathrm{Bi}^*(100,3) = \mathrm{Bi}(100,0.03)$	0.047	0.147	0.225	0.227	0.171	0.101	0.050
$\mathrm{Po}(3)$	0.050	0.149	0.224	0.224	0.168	0.101	0.050

$$p_X(k) = \frac{\lambda^k}{k!}\,\mathrm{e}^{-\lambda}\,, \qquad k \in \{0,1,2,\ldots\}\,.$$

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

$$\mathrm{E}X = \lambda\,, \qquad \mathrm{D}X = \lambda\,.$$

„Střední hodnota se rovná rozptylu;“ jedná se ***vždy o bezrozměrné celočíselné*** *náhodné veličiny (počet výskytů).*



Pravděpodobnostní a distribuční funkce Poissonova rozdělení $\mathrm{Po}\,(3)$

Poissonova rozdělení jako limitní případy binomických

Pro $m \to \infty$ při konstantním $m\,q_m = \lambda$, tj. $q_m = \frac{\lambda}{m}$:

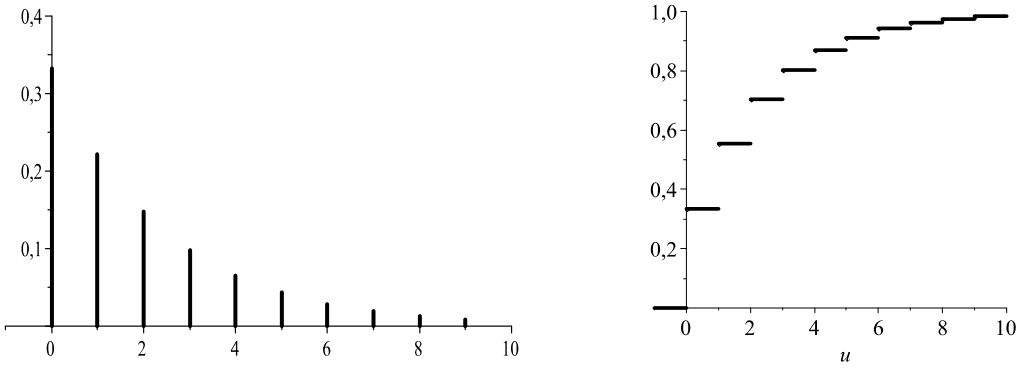
$$\begin{aligned} p_{\mathrm{Bi}(m,q_m)}(k) &= \tbinom{m}{k}\, q_m^k\,(1-q_m)^{m-k} = \\ = p_{\mathrm{Bi}^*(m,\lambda)}(k) &= \tbinom{m}{k}\,\left(\tfrac{\lambda}{m}\right)^k\,\left(1-\tfrac{\lambda}{m}\right)^{m-k} \\ &\to \frac{\lambda^k}{k!}\,\mathrm{e}^{-\lambda} = p_{\mathrm{Po}(\lambda)}(k)\,. \end{aligned}$$

Součet **nezávislých** náhodných veličin s Poissonovými rozděleními $\mathrm{Po}(\lambda)$, $\mathrm{Po}(\mu)$ má Poissonovo rozdělení $\mathrm{Po}(\lambda+\mu)$.

5.6.6 Geometrická

s parametrem $q \in (0,\infty)$ udávají
počet **úspěchů** do prvního **neúspěchu**, je-li v každém pokusu stejná pravděpodobnost **úspěchu** $q \in (0,1)$, neboli
počet **neúspěchů** do prvního **úspěchu**, je-li v každém pokusu stejná pravděpodobnost **neúspěchu** $q \in (0,1)$, neboli
počet jedniček na začátku posloupnosti nezávislých náhodných veličin s alternativními rozděleními $\mathrm{Alt}(q)$,

$$p_X(k) = q^k\,(1-q)\,, \qquad k \in \{0,1,2,\ldots\}\,,$$
$$\mathrm{E}X = \frac{q}{1-q}\,, \qquad \mathrm{D}X = \frac{q}{(1-q)^2}\,.$$



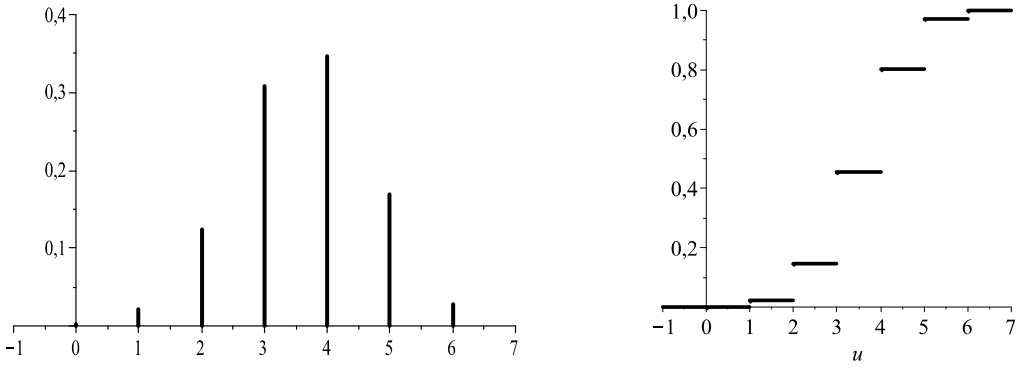
Pravděpodobnostní a distribuční funkce geometrického rozdělení pro $q = 2/3$

5.6.7 Hypergeometrická

s parametry $M, J, m \in \mathbb{N}$, $m \le J \le M$.
Např. M losů, z nichž J vyhrává; náhodná veličina je počet vyhrávajících z m vybraných,

$$p_X(j) = \frac{\binom{J}{j}\binom{M-J}{m-j}}{\binom{M}{m}}, \qquad j \in \{0, 1, 2, \ldots, m\},$$
$$\mathrm{E}X = \frac{m\,J}{M}, \qquad \mathrm{D}X = \frac{m\,J\,(M-J)\,(M-m)}{M^2\,(M-1)}.$$

Výpočetní složitost výpočtu $p_X(j)$ je $O(m)$, celého rozdělení $O(m^2)$.



Pravděpodobnostní a distribuční funkce hypergeometrického rozdělení pro $M = 25$, $J = 15$, $m = 6$

Binomická rozdělení jako limitní případy hypergeometrických

Pro $M \gg m$ je $\binom{M}{m} \doteq \frac{M^m}{m!}$ (Věta 2.5.6).
Hypergeometrická rozdělení pro $M \to \infty$ při konstantním $\frac{J_M}{M} = q$, tj. $\frac{M-J_M}{M} = 1 - q$:

$$p_X(j) = \frac{\binom{J_M}{j}\binom{M-J_M}{m-j}}{\binom{M}{m}} \to \frac{\frac{J_M^j}{j!} \cdot \frac{(M-J_M)^{m-j}}{(m-j)!}}{\frac{M^m}{m!}} =$$
$$= \frac{m!}{j!\,(m-j)!} \cdot \frac{J_M^j}{M^j} \cdot \frac{(M-J_M)^{m-j}}{M^{m-j}} = \binom{m}{j} q^j\,(1-q)^{m-j} = p_{\mathrm{Bi}(m,q)}(j).$$

5.7 Základní typy spojitých rozdělení

5.7.1 Rovnoměrná $\mathrm{R}(a,b)$

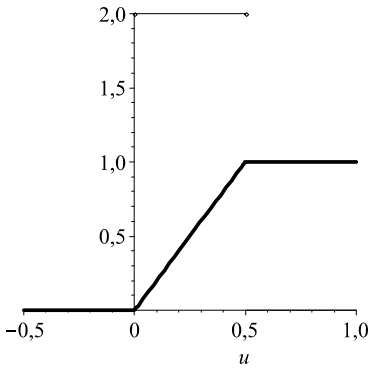
s parametry $a, b \in \mathbb{R}$, $a < b$.

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

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

$$q_X(\alpha) = a + (b-a)\,\alpha\,,$$

$$\mathrm{E}X = \frac{a+b}{2}\,, \qquad \mathrm{D}X = \frac{1}{12}\,(b-a)^2\,, \qquad \sigma_X = \frac{1}{2\sqrt{3}}\,(b-a)\,.$$



Hustota a distribuční funkce rovnoměrného rozdělení $\mathrm{R}\,(0, 1/2)$

5.7.2 Normální (Gaussova) $\mathrm{N}(\mu,\sigma^2)$

A. Normované $\mathrm{N}(0,1)$ s hustotou

$$\varphi(t) = f_{\mathrm{N}(0,1)}(t) = \frac{1}{\sqrt{2\pi}}\,\exp(\frac{-t^2}{2})$$

Distribuční funkce je transcendentní (Gaussův integrál) Φ,

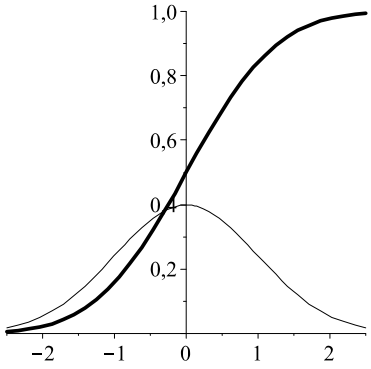
$$\Phi(u) = F_{\mathrm{N}(0,1)}(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}}\,\exp(\frac{-t^2}{2})\,\mathrm{d}t\,,$$

kvantilová funkce Φ^{-1} je inverzní k Φ.

B. Obecná $\mathrm{N}(\mu,\sigma^2) = \mu + \sigma\ \mathrm{N}(0,1)$
s parametry $\mu \in \mathbb{R}$ (střední hodnota), $\sigma^2 \in (0,\infty)$ (rozptyl),

$$f_{\mathrm{N}(\mu,\sigma^2)}(t) = \frac{1}{\sigma\,\sqrt{2\pi}}\,\exp(\frac{-(t-\mu)^2}{2\,\sigma^2})\,, \qquad \mathrm{E}X = \mu\,, \qquad \mathrm{D}X = \sigma^2\,.$$

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



Hustota a distribuční funkce normovaného normálního rozdělení $\mathrm{N}(0,1)$

5.7.3 Logaritmickonormální $\mathrm{LN}(\mu,\sigma^2) = \exp(\mathrm{N}(\mu,\sigma^2))$

s parametry $\mu \in \mathbb{R}$, $\sigma^2 \in (0,\infty)$ jsou rozdělení náhodných veličin $X = \exp(Y)$, kde Y má $\mathrm{N}(\mu,\sigma^2)$,

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

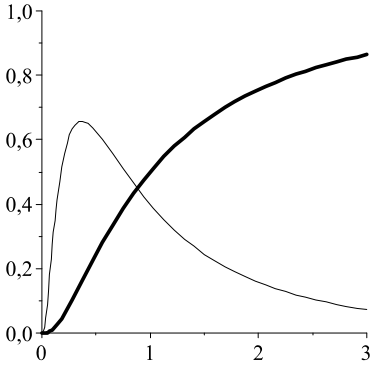
$$F_X(u) = \begin{cases} F_{\mathrm{N}(\mu,\sigma^2)}(\ln u) & \text{pro } u>0\,, \\ 0 & \text{jinak,} \end{cases}$$

$$\mathrm{E}X = \exp(\mu+\frac{\sigma^2}{2})\,,$$

$$\mathrm{D}X = (\exp(2\,\mu+\sigma^2))\,(\exp(\sigma^2)-1)\,,$$

$$\sigma_X = (\exp(\mu+\frac{\sigma^2}{2}))\,\sqrt{\exp(\sigma^2)-1}\,.$$

Součin dvou **nezávislých** veličin s logaritmickonormálním rozdělením $\mathrm{LN}(\mu_1,\sigma_1^2)$, $\mathrm{LN}(\mu_2,\sigma_2^2)$ má logaritmickonormální rozdělení $\mathrm{LN}(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$.

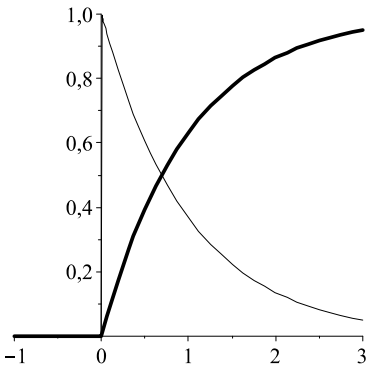


Hustota a distribuční funkce logaritmickonormálního rozdělení $\mathrm{LN}(0,1)$

5.7.4 Exponenciální $\mathrm{Ex}(\tau)$

s parametrem $\tau \in (0,\infty)$ (střední hodnota); např. rozdělení času do první poruchy, jestliže (podmíněná) pravděpodobnost poruchy za časový interval $\langle t, t+\delta\rangle$ závisí jen na δ, nikoli na t:

$$
\begin{aligned}
f_X(t) &= \begin{cases} \frac{1}{\tau}\,\exp(-\frac{t}{\tau}) & \text{pro } t>0, \\ 0 & \text{jinak,} \end{cases} \\
F_X(u) &= \begin{cases} 1-\exp(-\frac{u}{\tau}) & \text{pro } u>0\,, \\ 0 & \text{jinak,} \end{cases} \\
q_X(\alpha) &= -\tau\,\ln(1-\alpha)\,, \\
\mathrm{E}X &= \tau\,, \\
\mathrm{D}X &= \tau^2\,, \\
\sigma_X &= \tau\,, \\
\mathrm{Ex}(\tau) &= \tau\,\mathrm{Ex}(1)\,.
\end{aligned}
$$



Hustota a distribuční funkce exponenciálního rozdělení $\mathrm{Ex}\,(1)$

5.8 Čebyševova nerovnost

Motivační příklad (10 000 hodů mincí):
Při 10 000 hodech mincí má počet líců X binomické rozdělení $\mathrm{Bi}(10\,000, \frac{1}{2})$ se
střední hodnotou $\mathrm{E}X = 5\,000$,
směrodatnou odchylkou $\sigma_X = \sqrt{10\,000 \cdot \frac{1}{4}} = 50$.
Jaká je pravděpodobnost, že se výsledek bude lišit od střední hodnoty o nejméně $200 = 4\,\sigma_X$?

$$
\begin{aligned}
&\sum_{k=0}^{4\,800} p_{\mathrm{Bi}(10\,000,0.5)}(k) + \sum_{k=5\,200}^{10\,000} p_{\mathrm{Bi}(10\,000,0.5)}(k) = \\
&= 1 - \sum_{k=4\,801}^{5\,199} p_{\mathrm{Bi}(10\,000,0.5)}(k) = 1 - \sum_{k=4\,801}^{5\,199} \binom{10\,000}{k} \frac{1}{2^{10\,000}} \doteq 0.0000659\,.
\end{aligned}
$$

Věta:

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

kde $\mathrm{norm}\, X = \frac{X - \mathrm{E}X}{\sigma_X}$ (pokud má výraz smysl).

Důkaz pomocí kvantilové funkce:

$$
\begin{aligned}
1 = \mathrm{D}(\mathrm{norm}\, X) &= \mathrm{E}((\mathrm{norm}\, X)^2) - \underbrace{(\mathrm{E}(\mathrm{norm}\, X))^2}_{0} = \int_0^1 (q_{\mathrm{norm}\, X}(\alpha))^2\,\mathrm{d}\alpha \\
&\geq \int_I (q_{\mathrm{norm}\, X}(\alpha))^2\,\mathrm{d}\alpha\,,
\end{aligned}
$$

kde $I = \{\alpha \in (0,1) : |q_{\mathrm{norm}\,X}(\alpha)| \geq \delta\}$ jsou 2 intervaly o celkové délce $P(|\mathrm{norm}\,X| \geq \delta)$,

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

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

$$\begin{aligned}
\forall \delta > 0 : P(|\mathrm{norm}\,X| < \delta) &\geq 1 - \frac{1}{\delta^2}\,,\\
\forall \delta > 0 : P\Big(\Big|\frac{X - \mathrm{E}X}{\sigma_X}\Big| \geq \delta\Big) &\leq \frac{1}{\delta^2}\,,\\
\forall \varepsilon > 0 : P(|X - \mathrm{E}X| \geq \varepsilon) &\leq \frac{\sigma_X^2}{\varepsilon^2} = \frac{\mathrm{D}X}{\varepsilon^2}\,,\\
\forall \varepsilon > 0 : P(|X - \mathrm{E}X| < \varepsilon) &\geq 1 - \frac{\sigma_X^2}{\varepsilon^2} = 1 - \frac{\mathrm{D}X}{\varepsilon^2}\,.
\end{aligned}$$

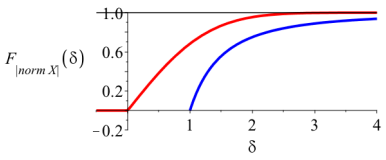
Motivační příklad (10 000 hodů mincí – pokračování):
Při 10 000 hodech mincí má počet líců X binomické rozdělení $\mathrm{Bi}(10\,000, 0.5)$ se
střední hodnotou $\mathrm{E}X = 5\,000$,
směrodatnou odchylkou $\sigma_X = \sqrt{10\,000 \cdot 0.25} = 50$.
Jaká je pravděpodobnost odchylky od střední hodnoty o více než $4\,\sigma_X$?

$$\text{nejvýše } (\tfrac{1}{4})^2 = \tfrac{1}{16}\,.$$

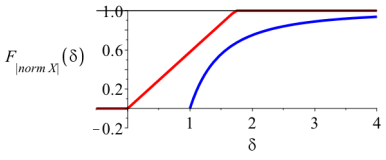
Buffonova úloha – přesnost odhadu
Při n hodech jehlou má počet protnutí linky X binomické rozdělení $\mathrm{Bi}(n, \frac{2}{\pi})$.
Kolik hodů potřebujeme, abychom π odhadli na 99 % s přesností 1 %?
Uf! Šetřme elektřinu, ne hlavu!

$$\begin{aligned}
0.01 \geq \frac{\sigma_X^2}{\varepsilon^2} = \frac{\sigma_X^2}{(0.01\,\mathrm{E}X)^2} = \frac{\frac{2}{\pi}\,(1 - \frac{2}{\pi})\,n}{(0.01\,\frac{2}{\pi}\,n)^2} = \frac{10\,000\,(\frac{\pi}{2} - 1)}{n}\,,\\
n \geq 1\,000\,000\,(\frac{\pi}{2} - 1) \doteq 570\,796\,.
\end{aligned}$$

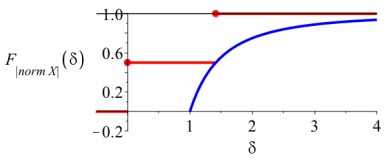
Uf! Šetřme ruce, ne hlavu!



Distribuční funkce absolutní hodnoty normovaného normálního rozdělení $|\,\mathrm{N}(0,1)|$ (červeně) ve srovnání s mezí dle Čebyševovy nerovnosti (modře)



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



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

6 Náhodné vektory

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

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

Lze psát

$$\boldsymbol{X}(\omega) = (X_1(\omega), \ldots, X_n(\omega))\,,$$

kde zobrazení $X_k\colon \Omega \to \mathbb{R}$, $k = 1, \ldots, n$, jsou náhodné veličiny.
Náhodný vektor lze považovat za vektor náhodných veličin $\boldsymbol{X} = (X_1, \ldots, X_n)$.
Je popsaný pravděpodobnostmi

$$\begin{aligned} P_{\boldsymbol{X}}(I_1 \times \ldots \times I_n) &= P(X_1 \in I_1, \ldots, X_n \in I_n) = \\ &= P(\{\omega \in \Omega \mid X_1(\omega) \in I_1, \ldots, X_n(\omega) \in I_n\})\,, \end{aligned}$$

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

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

definované pro libovolnou borelovskou množinu I v $\mathbb{R}^n$ (speciálně pro libovolné sjednocení spočetně mnoha n-rozměrných intervalů) a určující **rozdělení náhodného vektoru** $\boldsymbol{X}$.
Úspornější reprezentace: Stačí intervaly tvaru $I_k = (-\infty, t_k\rangle$, $t_k \in \mathbb{R}$,

$$\begin{aligned} P(X_1 \in (-\infty, t_1\rangle, \ldots, X_n \in (-\infty, t_n\rangle) &= P(X_1 \leq t_1, \ldots, X_n \leq t_n) = \\ &= P_{\boldsymbol{X}}((-\infty, t_1\rangle \times \ldots \times (-\infty, t_n\rangle) = \\ &= F_{\boldsymbol{X}}(t_1, \ldots, t_n)\,. \end{aligned}$$

$F_{\boldsymbol{X}}\colon \mathbb{R}^n \to \langle 0, 1\rangle$ je **sdružená distribuční funkce** náhodného vektoru $\boldsymbol{X}$. Je

- neklesající (ve všech proměnných),
- zprava spojitá (ve všech proměnných),
- $\lim\limits_{t_1 \to \infty, \ldots, t_n \to \infty} F_{\boldsymbol{X}}(t_1, \ldots, t_n) = 1$,
- $\forall k \in \{1, \ldots, n\}\ \ \forall t_1, \ldots, t_{k-1}, t_{k+1}, \ldots, t_n : \lim\limits_{t_k \to -\infty} F_{\boldsymbol{X}}(t_1, \ldots, t_n) = 0$.

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

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

$$0 \leq P_{\boldsymbol{X}}((a, b\rangle \times (c, d\rangle) = F_{\boldsymbol{X}}(b, d) - F_{\boldsymbol{X}}(a, d) - F_{\boldsymbol{X}}(b, c) + F_{\boldsymbol{X}}(a, b)\,,$$

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

$$F_{\boldsymbol{X}}(t_1, \ldots, t_n) = \prod_{k=1}^{n} F_{X_k}(t_k)\,.$$

6.1 Diskrétní náhodný vektor

má všechny složky diskrétní. Lze jej popsat též **sdruženou pravděpodobnostní funkcí** $p_{\boldsymbol{X}} : \mathbb{R}^n \to \langle 0, 1\rangle$

$$p_{\boldsymbol{X}}(t_1, \ldots, t_n) = P(X_1 = t_1, \ldots, X_n = t_n)\,,$$

která je nenulová 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:

$$p_{\boldsymbol{X}}(t_1, \ldots, t_n) = \prod_{i=1}^{n} p_{X_i}(t_i)\,.$$

6.2 Spojitý náhodný vektor

má všechny složky spojité. Lze jej popsat též **sdruženou hustotou pravděpodobnosti** což je (každá) nezáporná funkce $f_{\boldsymbol{X}} : \mathbb{R}^n \to \langle 0, \infty)$ taková, že

$$F_{\boldsymbol{X}}(t_1, \ldots, t_n) = \int_{-\infty}^{t_1} \ldots \int_{-\infty}^{t_n} f_{\boldsymbol{X}}(u_1, \ldots, u_n)\,\mathrm{d}u_1 \ldots \mathrm{d}u_n\,,$$

pro všechna $t_1, \ldots, t_n \in \mathbb{R}$. Pokud to jde, volíme

$$f_{\boldsymbol{X}}(t_1, \ldots, t_n) = \frac{\partial}{\partial t_1}\frac{\partial}{\partial t_2} \ldots \frac{\partial}{\partial t_n} F_{\boldsymbol{X}}(t_1, \ldots, t_n) = \mathrm{D}_1\mathrm{D}_2 \ldots \mathrm{D}_n F_{\boldsymbol{X}}(t_1, \ldots, t_n)\,.$$

Speciálně pro intervaly $\langle a_i, b_i\rangle$ dostáváme

$$\begin{aligned} P(X_1 \in \langle a_1, b_1\rangle, \ldots, X_n \in \langle a_n, b_n\rangle) &= P_{\boldsymbol{X}}(\langle a_1, b_1\rangle \times \ldots \times \langle a_n, b_n\rangle) \\ &= \int_{a_1}^{b_1} \ldots \int_{a_n}^{b_n} f_{\boldsymbol{X}}(u_1, \ldots, u_n)\,\mathrm{d}u_1 \ldots \mathrm{d}u_n \end{aligned}$$

Spojité náhodné veličiny $X_1, \ldots, X_n$ jsou **nezávislé**, právě když

$$f_{\boldsymbol{X}}(t_1, \ldots, t_n) = \prod_{i=1}^{n} f_{X_i}(t_i)\,.$$

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.

Někdy připouštíme i „náhodné veličiny“, jejichž hodnoty jsou jiné než numerické. Mohou to být např. náhodné množiny. Jindy nabývají konečně mnoha hodnot, kterým ponecháme jejich přirozené označení, např. „rub“, „líc“, „kámen“, „nůžky“, „papír“ apod.
Na těchto hodnotách nemusí být definovaná žádná aritmetika ani uspořádání.
Mohli bychom všechny hodnoty očíslovat, ale není žádný důvod, proč bychom to měli udělat právě určitým způsobem (který by ovlivnil následné numerické výpočty).
(Příklad: Číslování politických stran ve volbách.)

6.4 Číselné charakteristiky náhodného vektoru

Střední hodnota

- náhodného vektoru $\boldsymbol{X} = (X_1, \dots, X_n)$: $\quad \mathrm{E}\boldsymbol{X} := (\mathrm{E}X_1, \dots, \mathrm{E}X_n)$
- komplexní náhodné veličiny: $X = \Re(X) + \mathrm{i}\,\Im(X)$: $\quad \mathrm{E}X := \mathrm{E}\Re(X) + \mathrm{i}\,\mathrm{E}\Im(X)$
- nenumerické náhodné veličiny: nemá smysl

Rozptyl náhodného vektoru $\boldsymbol{X} = (X_1, \dots, X_n)$: $\quad \mathrm{D}\boldsymbol{X} := (\mathrm{D}X_1, \dots, \mathrm{D}X_n)$
Je-li U náhodná veličina, $a, b \in \mathbb{R}$, pak $a\,U + b$ má charakteristiky

$$\mathrm{E}(a\,U + b) = a\,\mathrm{E}U + b\,, \qquad \mathrm{D}(a\,U + b) = a^2\,\mathrm{D}U\,.$$

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.

$$\begin{aligned}
\mathrm{E}(X+Y) &= \mathrm{E}X + \mathrm{E}Y\,,\\
\mathrm{D}(X+Y) &= \mathrm{E}((X+Y)^2) - (\mathrm{E}(X+Y))^2\\
&= \mathrm{E}(X^2 + Y^2 + 2\,X\,Y) - (\mathrm{E}X + \mathrm{E}Y\,)^2\\
&= \mathrm{E}(X^2) + \mathrm{E}(Y^2) + 2\,\mathrm{E}(X\,Y) - ((\mathrm{E}X)^2 + (\mathrm{E}Y)^2 + 2\,\mathrm{E}X\,\mathrm{E}Y)\\
&= \underbrace{\mathrm{E}(X^2) - (\mathrm{E}X)^2}_{\mathrm{D}X} + \underbrace{\mathrm{E}(Y^2) - (\mathrm{E}Y)^2}_{\mathrm{D}Y} + 2\,\underbrace{(\mathrm{E}(X\,Y) - \mathrm{E}X\,\mathrm{E}Y)}_{\mathrm{cov}(X,Y)}\\
&= \mathrm{D}X + \mathrm{D}Y + 2\,\mathrm{cov}(X,Y)\,,
\end{aligned}$$

kde $\mathrm{cov}(X,Y) := \mathrm{E}(X\,Y) - \mathrm{E}X\,\mathrm{E}Y$ je **kovariance** náhodných veličin X, Y, též

$$\mathrm{cov}(X,Y) = \mathrm{E}((X - \mathrm{E}X)\,(Y - \mathrm{E}Y))\,,$$

neboť

$$\begin{aligned}
\mathrm{E}((X - \mathrm{E}X)\,(Y - \mathrm{E}Y)) &= \mathrm{E}(X\,Y - X\,\mathrm{E}Y - Y\,\mathrm{E}X + \mathrm{E}X\,\mathrm{E}Y)\\
&= \mathrm{E}(X\,Y) - \mathrm{E}X\,\mathrm{E}Y \underbrace{-\mathrm{E}X\,\mathrm{E}Y + \mathrm{E}X\,\mathrm{E}Y}_{0}\,.
\end{aligned}$$

Pro existenci kovariance je postačující existence rozptylů $\mathrm{D}X, \mathrm{D}Y$.
Vlastnosti kovariance:
$\mathrm{cov}(X,X) = \mathrm{D}X, \quad \mathrm{cov}(Y,X) = \mathrm{cov}(X,Y),$
$\mathrm{cov}(a\,X + b, c\,Y + d) = a\,c\,\mathrm{cov}(X,Y) \qquad (a,b,c,d \in \mathbb{R})$
(srovnejte s vlastnostmi rozptylu jako speciálního případu),
speciálně $\quad \mathrm{cov}(X,-X) = -\mathrm{D}X.$
Pro **nezávislé** náhodné veličiny X, Y je $\mathrm{cov}(X,Y) = 0$.

Použitím kovariance pro **normované** náhodné veličiny vyjde **korelace**:

$$\varrho(X,Y) = \mathrm{cov}(\mathrm{norm}\,X, \mathrm{norm}\,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X\,\sigma_Y} = \mathrm{E}(\mathrm{norm}\,X \cdot \mathrm{norm}\,Y)$$

(předpokládáme, že směrodatné odchylky ve jmenovateli jsou **nenulové**).
Speciálně $\varrho(X,X) = 1$.
Vlastnosti korelace
$\varrho(X,X) = 1, \quad \varrho(X,-X) = -1, \quad \varrho(X,Y) \in \langle -1, 1\rangle,$
$\varrho(Y,X) = \varrho(X,Y),$
$\varrho(aX + b, cY + d) = \mathrm{sign}(ac)\,\varrho(X,Y) \qquad (a,b,c,d \in \mathbb{R}, \quad a \neq 0 \neq c)$
(až na znaménko nezáleží na prosté lineární transformaci).
Důsledek: $\quad \varrho(aX + b, X) = \mathrm{sign}(a).$
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 $\boldsymbol{X} = (X_1, \ldots, X_n)$ je definována **kovarianční matice**

$$\Sigma_{\boldsymbol{X}} = \begin{bmatrix} \operatorname{cov}(X_1, X_1) & \operatorname{cov}(X_1, X_2) & \cdots & \operatorname{cov}(X_1, X_n) \\ \operatorname{cov}(X_2, X_1) & \operatorname{cov}(X_2, X_2) & \cdots & \operatorname{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(X_n, X_1) & \operatorname{cov}(X_n, X_2) & \cdots & \operatorname{cov}(X_n, X_n) \end{bmatrix}$$
$$= \begin{bmatrix} \mathrm{D}X_1 & \operatorname{cov}(X_1, X_2) & \cdots & \operatorname{cov}(X_1, X_n) \\ \operatorname{cov}(X_1, X_2) & \mathrm{D}X_2 & \cdots & \operatorname{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(X_1, X_n) & \operatorname{cov}(X_2, X_n) & \cdots & \mathrm{D}X_n \end{bmatrix}.$$

Je symetrická pozitivně semidefinitní, na diagonále má rozptyly.
Podobně je definována **korelační matice**

$$\varrho_{\boldsymbol{X}} = \begin{bmatrix} 1 & \varrho(X_1, X_2) & \cdots & \varrho(X_1, X_n) \\ \varrho(X_1, X_2) & 1 & \cdots & \varrho(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \varrho(X_1, X_n) & \varrho(X_2, X_n) & \cdots & 1 \end{bmatrix}.$$

Je symetrická pozitivně semidefinitní.

6.4.1 Vícerozměrné normální rozdělení $\mathrm{N}(\mu, \Sigma)$

popisuje speciální případ náhodného vektoru, jehož složky mají normální rozdělení a mohou být korelované. Má hustotu

$$f_{\mathrm{N}(\mu,\Sigma)}(\boldsymbol{t}) := \frac{1}{\sqrt{(2\,\pi)^n \, \det \boldsymbol{T}^{-1}}} \exp(-\frac{1}{2} \, (\boldsymbol{t} - \mu) \, \boldsymbol{T} \, (\boldsymbol{t} - \mu)^{\mathrm{T}}),$$

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

Parametry rozdělení:
$\mu = (\mu_1, \ldots, \mu_n) \in \mathbb{R}^n$ je střední hodnota náhodného vektoru,
$\Sigma := \boldsymbol{T}^{-1}$ 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 $\mathrm{N}(\mu_i, \Sigma_{ii})$;
pomocí těchto parametrů píšeme

$$f_{\mathrm{N}(\mu,\Sigma)}(\boldsymbol{t}) := \frac{1}{\sqrt{(2\,\pi)^n \, \det \Sigma}} \exp(-\frac{1}{2} \, (\boldsymbol{t} - \mu) \, \Sigma^{-1} \, (\boldsymbol{t} - \mu)^{\mathrm{T}}).$$

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 přinést 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

$(\Omega, \mathcal{A}, P)$ pravděpodobnostní prostor,
$\mathcal{L}$ lineární prostor všech náhodných veličin na $(\Omega, \mathcal{A}, P)$, tj. $\mathcal{A}$-měřitelných funkcí $\Omega \to \mathbb{R}$,
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,
$\bullet\colon \mathcal{L}_2 \times \mathcal{L}_2 \to \mathbb{R}$,

$$X \bullet Y := \mathrm{E}(X\,Y)\,,$$

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)*,

$$||X|| := \sqrt{X \bullet X} = \sqrt{\mathrm{E}(X^2)}$$

je **norma**,

$$d(X,Y) := ||X - Y|| = \sqrt{\mathrm{E}((X-Y)^2)}$$

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.
$\mathrm{E}X$ je kolmý průmět X do $\mathcal{R}$*(pokud ztotožňujeme toto reálné číslo s příslušnou konstantní náhodnou veličinou, jinak souřadnice ve směru $\mathcal{R}$)*,
$X - \mathrm{E}X$ je kolmý průmět X do $\mathcal{N}$,
$\operatorname{norm} X = \frac{X - \mathrm{E}X}{\sigma_X}$ je jednotkový vektor ve směru kolmého průmětu X do $\mathcal{N}$,
$\sigma_X = ||X - \mathrm{E}X||$ je vzdálenost X od $\mathcal{R}$.
Z kolmosti vektorů $X - \mathrm{E}X \in \mathcal{N}$, $\mathrm{E}X \in \mathcal{R}$ a Pythagorovy věty plyne (1)

$$\begin{aligned} X \bullet X = ||X||^2 &= ||X - \mathrm{E}X||^2 + ||\mathrm{E}X||^2\,, \\ \mathrm{E}(X^2) &= \mathrm{D}X + (\mathrm{E}X)^2\,. \end{aligned}$$

7.1 Lineární podprostor $\mathcal{N}$ náhodných veličin s nulovými středními hodnotami

Speciálně pro náhodné veličiny z $\mathcal{N}$:

$$\begin{aligned} \sigma_X^2 &= X \bullet X\,, \\ \sigma_X &= ||X||\,, \\ \operatorname{cov}(X,Y) &= X \bullet Y\,, \\ \varrho(X,Y) &= \frac{\operatorname{cov}(X,Y)}{\sigma_X\,\sigma_Y} = \frac{X \bullet Y}{||X||\;||Y||} = \cos\angle(X,Y)\,. \end{aligned}$$

Důsledek: Náhodné veličiny X, Y **s nulovými středními hodnotami** jsou ortogonální, právě když jsou nekorelované.
Obecně v $\mathcal{L}_2$
$\varrho(X,Y)$ je kosinus úhlu průmětů X, Y do $\mathcal{N}$,
$\operatorname{cov}(X,Y) = X \bullet Y - \mathrm{E}X\,\mathrm{E}Y$ 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 $\boldsymbol{X} = (X_1, \ldots, X_n)$ 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 aproximací náhodné veličiny Y ve smyslu kritéria

$$\Big|\Big|\sum_k c_k\,X_k - Y\Big|\Big|\,.$$

Řešení: K vektoru Y hledáme nejbližší 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, $j = 1, \ldots, n$,

$$\Big(\sum_k c_k\,X_k - Y\Big) \bullet X_j = 0\,,$$

$$\sum_i c_i \, (X_i \bullet X_j) = Y \bullet X_j \, .$$

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

Speciálně pro náhodné veličiny **s nulovými středními hodnotami**:

$$\sum_i c_i \, \mathrm{cov}(X_i, X_j) = \mathrm{cov}(Y, X_j) \, ,$$

takže matice soustavy je kovariančí matice $\Sigma_{\boldsymbol{X}}$.

8 Základní pojmy statistiky

8.1 K čemu potřebujeme 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, empirické rozdělení

Motivační úloha: Máme odhadnout rozdělení hmotnosti lidí a jeho parametry.
Náhodný pokus X: Ze **základního souboru (=populace)** náhodně vybereme jeden objekt (člověka), ω (s rovnoměrným rozdělením).
Určíme jeho hmotnost, $X(\omega)$, kterou prozatím považujeme za konstantní, tj. s Diracovým rozdělením a rovnou její realizaci, $x(\omega) \in \mathbb{R}$.
Za množinu všech elementárních jevů Ω můžeme vzít množinu objektů (lidí), z nichž vybíráme, za σ-algebru $\exp\Omega$, pravděpodobnostní míra je $M \mapsto \frac{|M|}{|\Omega|}$, $M \subseteq \Omega$.
Náhodná veličina $X\colon \Omega \to \mathbb{R}$ (závisí jen výběru člověka) má diskrétní rozdělení (směs Diracových),

$$\mathrm{E}X = \frac{1}{|\Omega|}\sum_{\omega\in\Omega} x(\omega)\,,$$
$$\mathrm{D}X = \frac{1}{|\Omega|}\sum_{\omega\in\Omega} (x(\omega) - \mathrm{E}X)^2\,.$$

Náhodný pokus $n\times$X: Náhodný pokus X nezávisle opakujeme n-krát.
Vybereme (s opakováním a s rovnoměrným rozdělením) n-tici lidí $(\omega_1,\dots,\omega_n) \in \Omega^n$. Výsledkem je n-tice hmotností

$$(x_1,\dots,x_n) := \big(x(\omega_1),\dots,x(\omega_n)\big)\,.$$

Za množinu všech elementárních jevů můžeme vzít

$$\Omega^n = \{(\omega_1,\dots,\omega_n) \mid \omega_1\in\Omega,\dots,\omega_n\in\Omega\}\,,$$

za σ-algebru $\exp(\Omega^n)$, pravděpodobnostní míra je $M \mapsto \frac{|M|}{|\Omega|^n}$, $M \subseteq \Omega^n$.
Při j-tém opakování pokusu realizujeme náhodnou veličinu $X_j\colon \Omega^n \to \mathbb{R}$,

$$X_j(\omega_1,\dots,\omega_n) \mapsto X(\omega_j) = x_j\,,$$

která závisí jen na j-té souřadnici a je konstantní na všech elementárních jevech (n-ticích lidí), které ji mají stejnou (ve kterých j-tý člověk je stejný).
Celý pokus charakterizuje náhodný vektor $\boldsymbol{X} := (X_1,\dots,X_n)\colon \Omega^n \to \mathbb{R}^n$,

$$(X_1,\dots,X_n)(\omega_1,\dots,\omega_n) \mapsto (x_1,\dots,x_n)\,.$$

Všechny jeho složky jsou **nezávislé** a mají **stejné rozdělení** jako původní náhodná veličina X.
Původní rozdělení náhodné veličiny X **neznáme**, neboť jsme ho **nezměřili** na celém základním souboru.
Posloupností n realizací jsme dostali **výběrový soubor**, který „reprezentuje“ základní soubor.
Výběrový soubor definuje obdobným způsobem tzv. **empirické rozdělení** $\mathrm{Emp}(\boldsymbol{x})$:

Náhodný pokus Y: Z **výběrového souboru** $\boldsymbol{x} := (x_1,\dots,x_n)$ náhodně vybereme jeden prvek, j-tý (s rovnoměrným rozdělením). Výsledkem je $x_j \in \mathbb{R}$.
Empirické rozdělení známe, neboť jsme ho **změřili**.

Za množinu všech elementárních jevů Ω_{Emp} můžeme vzít $\{1,\dots,n\}$, za σ-algebru $\exp\Omega_{\text{Emp}}$, pravděpodobnostní míra je $M \mapsto \frac{|M|}{n}$, $M \subseteq \Omega_{\text{Emp}}$,

$$\mathrm{E}\,\mathrm{Emp}(\boldsymbol{x}) = \frac{1}{n}\sum_{j=1}^{n} x_j =: \bar{\boldsymbol{x}}\,,$$
$$\mathrm{D}\,\mathrm{Emp}(\boldsymbol{x}) = \frac{1}{n}\sum_{j=1}^{n}(x_j - \mathrm{E}\,\mathrm{Emp}(\boldsymbol{x}))^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{\boldsymbol{x}})^2\,.$$

Tyto hodnoty mohou sloužit k odhadu neznámých parametrů původního rozdělení na základním souboru.
Vše funguje obdobně, pokud pozorovaná náhodná veličina X není pro každého jedince konstantní (závisí na dalších parametrech, „okolnostech“).
Stačí v modelu použít více elementárních jevů, kterými mohou být např. dvojice (člověk, okolnosti). Závislost na okolnostech může být i spojitá.
Empirické rozdělení je vždy diskrétní (směs Diracových, $\mathrm{Mix}_{(1/n,\dots,1/n)}(x_1,\dots,x_n)$), používá již změřené veličiny a jediný náhodný vliv je výběr prvku z posloupnosti.

Obvykle vylučujeme vícenásobný výběr stejného prvku (*výběr bez vracení*), čímž se dopouštíme chyby. Ta se obvykle zanedbává, neboť

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

Přesnost odhadu je dána velikostí výběrového souboru, nikoli populace.

Náhodný výběr $\boldsymbol{X} := (X_1,\dots,X_n)$ je vektor náhodných veličin, které jsou **nezávislé** a mají **stejné rozdělení**.
Náhodný pokus $n\times$X je jednou z možností, jak tyto podmínky zajistit.
(Vynecháváme indexy, např. F_X místo F_{X_k}.)
Provedením pokusu dostaneme **realizaci náhodného výběru**,
$\boldsymbol{x} := (x_1,\dots,x_n) \in \mathbb{R}^n$,
kde n je **rozsah výběru**.

funkce $f\colon D \to \mathbb{R}$	funkční hodnota $f(x) \in \mathbb{R}, \qquad x \in D$
náhodná veličina $X\colon \Omega \to \mathbb{R}$	realizace náhodné veličiny $x := X(\omega) \in \mathbb{R}, \qquad \omega \in \Omega$
náhodný vektor/výběr $\boldsymbol{X} = (X_1,\dots,X_n)\colon \Omega^n \to \mathbb{R}^n$	realizace náhodného vektoru/výběru $\boldsymbol{x} = (x_1,\dots,x_n) := \boldsymbol{X}(\boldsymbol{\omega}) \in \mathbb{R}^n, \qquad \boldsymbol{\omega} \in \Omega^n$

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

$$P(G(X_1,\dots,X_n) \le t) = F_{G(X_1,\dots,X_n)}(t)\,.$$

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

8.3 Obecné vlastnosti odhadů

Značení:
ϑ ... jakákoli hodnota parametru (reálné číslo),
ϑ^* ... skutečná (správná) hodnota parametru (reálné číslo),
$\widehat{\Theta}, \widehat{\Theta}_n$ ... odhad parametru založený na náhodném výběru rozsahu n (náhodná veličina)
$\widehat{\vartheta}, \widehat{\vartheta}_n$ ... realizace odhadu (reálné číslo)
Žádoucí vlastnosti odhadů:

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

8.4 Odhad střední hodnoty

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

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

Když totéž provedeme s náhodnými veličinami výběru, dostaneme náhodnou veličinu

$$\overline{\boldsymbol{X}} = \frac{1}{n}\sum_{j=1}^{n} X_j\,.$$

$\overline{\boldsymbol{X}}$ = **výběrový průměr**, $\bar{\boldsymbol{x}}$ = **realizace výběrového průměru**.
Alternativní značení: $\overline{\boldsymbol{X}}_n, \bar{\boldsymbol{x}}_n$ *(pokud potřebujeme zdůraznit rozsah výběru)*
Věta:

$$\mathrm{E}\overline{\boldsymbol{X}}_n = \frac{1}{n}\sum_{j=1}^{n} \mathrm{E}X = \mathrm{E}X\,,$$

$$\mathrm{D}\overline{\boldsymbol{X}}_n = \frac{1}{n^2}\sum_{j=1}^{n} \mathrm{D}X = \frac{1}{n}\,\mathrm{D}X\,,$$

$$\sigma_{\overline{\boldsymbol{X}}_n} = \sqrt{\frac{1}{n}\,\mathrm{D}X} = \frac{1}{\sqrt{n}}\,\sigma_X\,,$$

pokud existují. (Zde $\mathrm{E}X = \mathrm{E}X_j$ atd.) Výběrový průměr minimalizuje kritérium nejmenších čtverců

$$\ell_2(c) = \mathrm{E}(c - \mathrm{Emp}(\boldsymbol{x}))^2 = \frac{1}{n}\sum_{i=1}^{n}(c - x_i)^2\,.$$

Důsledek: Výběrový průměr je nestranný konzistentní odhad střední hodnoty.
(Nezávisle na typu rozdělení.)
Čebyševova nerovnost pro $\overline{\boldsymbol{X}}_n$ dává

$$P(|\overline{\boldsymbol{X}}_n - \mathrm{E}X| \geq \varepsilon) \leq \frac{\mathrm{D}\overline{\boldsymbol{X}}_n}{\varepsilon^2} = \frac{\mathrm{D}X}{n\,\varepsilon^2} \to 0 \qquad \text{pro } n \to \infty\,.$$

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 $n\,\mathrm{D}X \to \infty$. **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 $\mathrm{N}(\mu, \sigma^2)$ má normální rozdělení $\mathrm{N}(\mu, \frac{1}{n}\,\sigma^2)$ 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 $\mathrm{E}X$ a směrodatnou odchylkou $\sigma_X \neq 0$. Pak normované náhodné veličiny

$$Y_n = \operatorname{norm} \overline{\boldsymbol{X}}_n = \frac{\sqrt{n}}{\sigma_X} \left(\overline{\boldsymbol{X}}_n - \mathrm{E}X\right)$$

konvergují k normovanému normálnímu rozdělení v následujícím smyslu:

$$\forall t \in \mathbb{R} : \lim_{n\to\infty} F_{Y_n}(t) = \lim_{n\to\infty} F_{\operatorname{norm} \overline{\boldsymbol{X}}_n}(t) = \Phi(t)\,,$$

neboli

$$\forall t \in \mathbb{R} : \lim_{n\to\infty} |F_{Y_n}(t) - \Phi(t)| = 0\,,$$

Pokud 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 $1-\Phi(4) \doteq 1-0.9999683 = 0.0000317$ bude aspoň o $4\,\sigma_X$ větší, se stejnou pravděpodobností o $4\,\sigma_X$ menší, celkem

$$2\,(1-\Phi(4)) \doteq 2 \cdot 0.0000317 = 0.0000633\,,$$

v dobré shodě s přesnějším (a pracným) výsledkem 0.0000659.
Kdybychom zohlednili kvantizační chybu a uvažovali toleranci ± 199.5, dostali bychom ještě lepší odhad 0.0000661.

Buffonova úloha – přesnost odhadu
Binomické rozdělení $\mathrm{Bi}(n, \frac{2}{\pi})$ při n hodech se blíží normálnímu se
střední hodnotou $\mathrm{E}X = \frac{2}{\pi}\, n \doteq 0.6366\, n$,
směrodatnou odchylkou $\sigma_X = \sqrt{\frac{2}{\pi}\,(1-\frac{2}{\pi})\, n} \doteq 0.48\,\sqrt{n}$.
Kolik hodů n potřebujeme, abychom π odhadli na 99 % s přesností 1 %?
S pravděpodobností 99 % normovaná hodnota bude v absolutní hodnotě menší než kvantil $\Phi^{-1}(0.995) \doteq 2.576$.
To je tolerance měřená ve směrodatných odchylkách.
Má být nejvýše 1 % střední hodnoty,

$$\begin{aligned} 2.576\,\sigma_X &\leq 0.01\,\mathrm{E}X\,, \\ 2.576\,\sqrt{\tfrac{2}{\pi}\,(1-\tfrac{2}{\pi})\, n} &\leq 0.01\,\tfrac{2}{\pi}\, n\,, \\ \sqrt{n} &\geq 2.576 \cdot 100\,\sqrt{\tfrac{\pi}{2}-1}\,, \\ n &\geq 2.576^2 \cdot 10\,000\,(\tfrac{\pi}{2}-1) \doteq 37\,877\,. \end{aligned}$$

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

$$\sum_{k=23\,872}^{24\,354} p_{\mathrm{Bi}(37\,877,\,\frac{2}{\pi})}(k) = \sum_{k=23\,872}^{24\,354} \binom{37\,877}{k} (\tfrac{2}{\pi})^k\, (1-\tfrac{2}{\pi})^{37\,877-k} \doteq 0.00988\,.$$

8.5 Odhad k-tého obecného momentu $\mathrm{E}X^k$

pomocí k-tého obecného momentu empirického rozdělení:

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

Je realizací odhadu

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

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

$$\mathrm{E}M_{\boldsymbol{X}^k} = \mathrm{E}\Big(\frac{1}{n}\sum_{j=1}^{n} X_j^k\Big) = \frac{1}{n}\sum_{j=1}^{n}\mathrm{E}X_j^k = \mathrm{E}X^k\,,$$

$$\mathrm{D}M_{\boldsymbol{X}^k} = \frac{1}{n^2}\,n\,\mathrm{D}X^k = \frac{1}{n}\left(\mathrm{E}(X^k)^2 - (\mathrm{E}X^k)^2\right) = \frac{1}{n}\left(\mathrm{E}X^{2\,k} - (\mathrm{E}X^k)^2\right) \to 0\,.$$

8.6 Odhad rozptylu

8.6.1 Odhad rozptylu při známé střední hodnotě

$$\widehat{\theta} = \frac{1}{n}\sum_{j=1}^{n}(x_j - \mathrm{E}X)^2\,.$$

Je realizací odhadu

$$\widehat{\Theta} = \frac{1}{n}\sum_{j=1}^{n}(X_j - \mathrm{E}X)^2\,.$$

$$\mathrm{E}\widehat{\Theta} = \frac{1}{n}\sum_{j=1}^{n}\underbrace{\mathrm{E}(X_j - \mathrm{E}X)^2}_{\mathrm{D}X} = \mathrm{D}X\,.$$

- nestranný,
- konzistentní, pokud existuje 4. centrální moment

Rozdělení odhadu rozptylu pro výběr z normálního rozdělení $\mathrm{N}(\mu, \sigma^2)$

$$X_j - \mu \sim \mathrm{N}(0, \sigma^2) = \sigma\ \mathrm{N}(0,1)$$

$$U_j := \frac{X_j - \mu}{\sigma} \sim \mathrm{N}(0,1)$$

$$\sum_{j=1}^{n} U_j^2 = \frac{1}{\sigma^2}\sum_{j=1}^{n}(X_j - \mu)^2 = \frac{n\,\widehat{\Theta}}{\sigma^2} \sim$$

8.6.2 Rozdělení χ^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 $\mathrm{N}(0,1)$.

$$\mathrm{E}\chi^2(n) = \mathrm{E}\sum_{j=1}^{n} U_j^2 = \sum_{j=1}^{n}\underbrace{\mathrm{E}U_j^2}_{\mathrm{D}U_j=1} = n\,, \qquad \mathrm{E}\frac{\chi^2(n)}{n} = 1\,,$$

$$\mathrm{D}\chi^2(n) = \mathrm{D}\sum_{j=1}^{n} U_j^2 = \sum_{j=1}^{n}\underbrace{\mathrm{D}U_j^2}_{2} = 2\,n\,, \qquad \mathrm{D}\frac{\chi^2(n)}{n} = \frac{2}{n}\,.$$

Součet **nezávislých** náhodných veličin s rozděleními $\chi^2(k)$, $\chi^2(n)$ má rozdělení $\chi^2(k+n)$.

Hustota

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

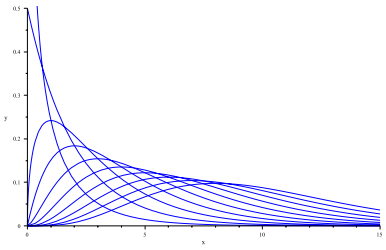
$$c(n) = \frac{1}{2^{\frac{n}{2}}\, \Gamma(\frac{n}{2})}\,,$$

$$\Gamma(z) = \int_0^\infty t^{z-1}\, \mathrm{e}^{-t}\, \mathrm{d}t\,,$$

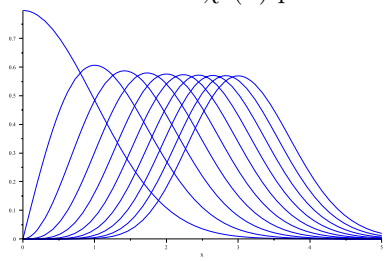
speciálně $\Gamma(m+1) = m!$ pro všechna $m \in \mathbb{N}$.

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

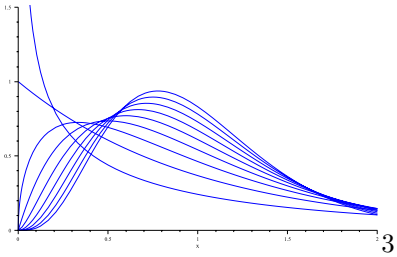
Důsledek: Součet m nezávislých náhodných veličin s exponenciálním rozdělením $\mathrm{Ex}(2) = \chi^2(2)$ má rozdělení $\chi^2(2m)$.



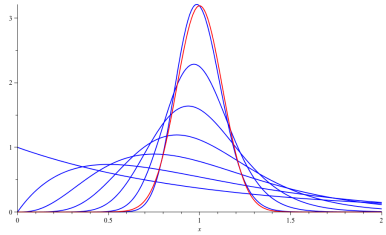
Hustota rozdělení $\chi^2(n)$ pro $n = 1, 2, \ldots, 10$ stupňů volnosti.



Hustota rozdělení $\sqrt{\chi^2(n)}$ („vzdálenosti od středu terče“) pro $n = 1, 2, \ldots, 10$ stupňů volnosti.



Hustota rozdělení $\chi^2(n)/n$ pro $n = 1, 2, \ldots, 10$ stupňů volnosti.



Hustota rozdělení $\chi^2(n)/n$ pro $n = 2, 4, 8, \ldots, 128$ stupňů volnosti.
Červeně náhrada posledního rozdělení normálním.

CLV pro $n \to \infty$

$\Rightarrow$ „$\chi^2(n) \rightsquigarrow \mathrm{N}(n, 2n)$“, „$\dfrac{\chi^2(n)}{n} \rightsquigarrow \mathrm{N}(1, \frac{2}{n})$“,

přesněji norm $\chi^2(n) = \dfrac{\chi^2(n)-n}{\sqrt{2n}} \rightsquigarrow \mathrm{N}(0,1)\,,$ $\qquad \sqrt{\dfrac{n}{2}}\left(\dfrac{\chi^2(n)}{n}-1\right) \rightsquigarrow \mathrm{N}(0,1)\,.$

Náhrada je „dobrá uprostřed“, kde je velká hustota.
Jenže my budeme potřebovat kvantily „na krajích“, a ty konvergují pomalu:

α-kvantily rozdělení $\chi^2(n)$ a jeho náhrady normálním rozdělením $\mathrm{N}(n, 2n)$

$n\backslash\alpha$	0.005	0.01	0.025	0.05	0.95	0.975	0.99	0.995	0.999	0.9995
100	67.33	70.06	74.22	77.93	124.3	129.6	135.8	140.2	149.4	153.2
	63.57	67.10	72.28	76.74	123.3	127.7	132.9	136.4	143.7	146.5
200	152.2	156.4	162.7	168.3	234.0	241.1	249.4	255.3	267.5	272.4
	148.5	153.5	160.8	167.1	232.9	239.2	246.5	251.5	261.8	265.8
300	240.7	246.0	253.9	260.9	341.4	349.9	359.9	366.8	381.4	387.2
	236.9	243.0	252.0	259.7	340.3	348.0	357.0	363.1	375.7	380.6
400	330.9	337.2	346.5	354.6	447.6	457.3	468.7	476.6	493.1	499.7
	327.1	334.2	344.6	353.5	446.5	455.4	465.8	472.9	487.4	493.1

α-kvantily rozdělení $\chi^2(n)/n$ **vyděleného počtem stupňů volnosti** a jeho náhrady normálním rozdělením $\mathrm{N}(1, 2/n)$

$n\backslash\alpha$	0.005	0.01	0.025	0.05	0.95	0.975	0.99	0.995	0.999	0.9995
100	0.673	0.701	0.742	0.779	1.24	1.30	1.36	1.40	1.49	1.53
	0.636	0.671	0.723	0.767	1.23	1.28	1.33	1.36	1.44	1.47
200	0.760	0.780	0.815	0.840	1.17	1.20	1.24	1.28	1.34	1.36
	0.742	0.767	0.804	0.836	1.16	1.20	1.23	1.26	1.31	1.33
300	0.803	0.820	0.847	0.870	1.14	1.17	1.20	1.22	1.27	1.29
	0.790	0.810	0.840	0.866	1.13	1.16	1.19	1.21	1.25	1.27
400	0.828	0.842	0.865	0.888	1.12	1.14	1.17	1.19	1.23	1.25
	0.818	0.836	0.861	0.884	1.12	1.14	1.16	1.18	1.22	1.23

8.6.3 Odhad rozptylu při neznámé střední hodnotě

pomocí rozptylu empirického rozdělení:

$$\widehat{\sigma^2_{\boldsymbol{x}}} := \mathrm{D}\,\mathrm{Emp}(\boldsymbol{x}) = \frac{1}{n}\sum_{j=1}^{n}(x_j-\bar{\boldsymbol{x}})^2\,.$$

Je realizací odhadu

$$\widehat{\sigma^2_{\boldsymbol{X}}} := \frac{1}{n}\sum_{j=1}^{n}(X_j-\overline{\boldsymbol{X}})^2 \neq \frac{1}{n}\sum_{j=1}^{n}(X_j-\mathrm{E}X)^2\,.$$

Vzorce jsou dvouprůchodové, ale lze je upravit na jednoprůchodové (numericky nevhodné):

$$\widehat{\sigma^2_{\boldsymbol{X}}} = \frac{1}{n}\Big(\sum_{j=1}^{n}X_j^2 - 2\,\overline{\boldsymbol{X}}\underbrace{\sum_{j=1}^{n}X_j}_{n\,\overline{\boldsymbol{X}}} + \underbrace{\sum_{j=1}^{n}\overline{\boldsymbol{X}}^2}_{n\,\overline{\boldsymbol{X}}^2}\Big) = \frac{1}{n}\sum_{j=1}^{n}X_j^2 - \overline{\boldsymbol{X}}^2\,,$$

$$\widehat{\sigma^2_{\boldsymbol{x}}} = \frac{1}{n}\sum_{j=1}^{n}x_j^2 - \bar{\boldsymbol{x}}^2 = \frac{1}{n}\sum_{j=1}^{n}x_j^2 - \Big(\frac{1}{n}\sum_{j=1}^{n}x_j\Big)^2\,.$$

Věta: $\widehat{\sigma^2_{\boldsymbol{X}}}$ je **vychýlený** konzistentní odhad rozptylu (pokud původní rozdělení má rozptyl a 4. centrální moment).

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

$$\mathrm{E}\widehat{\sigma^2_{\boldsymbol{X}}} = \mathrm{E}\Big(\frac{1}{n}\sum_{j=1}^{n}X_j^2 - \overline{\boldsymbol{X}}^2\Big) = \mathrm{E}X^2 - \mathrm{E}\overline{\boldsymbol{X}}^2 = (\mathrm{E}X)^2 + \mathrm{D}X - \underbrace{(\mathrm{E}\overline{\boldsymbol{X}})^2}_{\mathrm{E}X} - \underbrace{\mathrm{D}\overline{\boldsymbol{X}}}_{\frac{1}{n}\mathrm{D}X} =$$

$$= (1-\tfrac{1}{n})\,\mathrm{D}X = \frac{n-1}{n}\,\mathrm{D}X \to \mathrm{D}X \qquad \text{pro } n\to\infty\,.$$

$\Rightarrow$ **nestranný** odhad:

$$S^2_{\boldsymbol{X}} := \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{\boldsymbol{X}}_n)^2 = \frac{n}{n-1}\,\widehat{\sigma^2_X} \qquad (\textbf{výběrový rozptyl})\,,$$

$$s^2_{\boldsymbol{x}} := \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{\boldsymbol{x}}_n)^2 = \frac{n}{n-1}\,\widehat{\sigma^2_x} \qquad (\textbf{realizace výběrového rozptylu})\,.$$

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

$$S^2_{\boldsymbol{X}} = \frac{1}{n-1}\sum_{j=1}^{n}X_j^2 - \frac{n}{n-1}\,\overline{\boldsymbol{X}}^2_n = \frac{1}{n-1}\sum_{j=1}^{n}X_j^2 - \frac{1}{n\,(n-1)}\Big(\sum_{j=1}^{n}X_j\Big)^2\,,$$

$$s^2_{\boldsymbol{x}} = \frac{1}{n-1}\sum_{j=1}^{n}x_j^2 - \frac{n}{n-1}\,\bar{\boldsymbol{x}}^2_n = \frac{1}{n-1}\sum_{j=1}^{n}x_j^2 - \frac{1}{n\,(n-1)}\Big(\sum_{j=1}^{n}x_j\Big)^2\,.$$

Věta:

$$\mathrm{E}S^2_{\boldsymbol{X}} = \mathrm{D}X.$$

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

$$\sum_{j=1}^{n}(X_j - \overline{\boldsymbol{X}})^2$$

jsou veličiny $X_j - \overline{\boldsymbol{X}}$ závislé.

1. Pro $n = 2$ a X_1, X_2 s rozdělením $\mathrm{N}(0,1)$:

$$\overline{\boldsymbol{X}} = \frac{X_1 + X_2}{2}\,, \qquad X_1 - \overline{\boldsymbol{X}} = -(X_2 - \overline{\boldsymbol{X}}) = \frac{X_1 - X_2}{2} \text{ má rozdělení } \mathrm{N}(0, \tfrac{1}{2})\,,$$

$$S^2_{\boldsymbol{X}} = (X_1 - \overline{\boldsymbol{X}})^2 + (X_2 - \overline{\boldsymbol{X}})^2 = 2\left(\frac{X_1 - X_2}{2}\right)^2 = \left(\frac{X_1 - X_2}{\sqrt{2}}\right)^2 = U^2\,,$$

kde $U = \dfrac{X_1 - X_2}{\sqrt{2}}$ má rozdělení $\mathrm{N}(0,1)$, takže $S^2_{\boldsymbol{X}}$ má rozdělení $\chi^2(1)$.
2. Pro $X_1, \ldots, X_n$ s rozdělením $\mathrm{N}(0,1)$:
Náhodný vektor $\boldsymbol{X} = (X_1, \ldots, X_n)$ s nezávislými složkami má rozdělení sféricky symetrické (kolem počátku), které se nezmění rotací ani ortonormální transformací souřadnic, $\boldsymbol{U} = \boldsymbol{X}\,\boldsymbol{M}$, kde $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ je ortonormální matice. Použijeme ortogonální, resp. ortonormální matici

$$\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 & 1 \\ -1 & 1 & 1 & \ddots & 1 & 1 \\ 0 & -2 & 1 & \ddots & 1 & 1 \\ 0 & 0 & -3 & \ddots & 1 & 1 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -(n-1) & 1 \end{bmatrix}, \text{resp. } \boldsymbol{M} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{12}} & \cdots & \frac{1}{\sqrt{n(n+1)}} & \frac{1}{\sqrt{n}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{12}} & \ddots & \frac{1}{\sqrt{n(n+1)}} & \frac{1}{\sqrt{n}} \\ 0 & -\frac{2}{\sqrt{6}} & \frac{1}{\sqrt{12}} & \ddots & \frac{1}{\sqrt{n(n+1)}} & \frac{1}{\sqrt{n}} \\ 0 & 0 & -\frac{3}{\sqrt{12}} & \ddots & \frac{1}{\sqrt{n(n+1)}} & \frac{1}{\sqrt{n}} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -\frac{n-1}{\sqrt{n(n+1)}} & \frac{1}{\sqrt{n}} \end{bmatrix}.$$

Postup pro $n = 2$ byl speciálním případem pro

$$\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}, \text{ , resp. } \boldsymbol{M} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

Transformací se zachovala nezávislost $U_1, \ldots, U_n$, jejich rozdělení $\mathrm{N}(0,1)$ i součet

$$\sum_{j=1}^{n} X_j^2 = ||(X_1, \ldots, X_n)||^2 = ||(U_1, \ldots, U_n)||^2 = \sum_{j=1}^{n} U_j^2 ,$$

kde pro poslední člen platí $U_n = \sqrt{n}\,\overline{\boldsymbol{X}}, \quad U_n^2 = n\,\overline{\boldsymbol{X}}^2$,

$$\sum_{j=1}^{n-1} U_j^2 = \sum_{j=1}^{n} X_j^2 - U_n^2 = \sum_{j=1}^{n} X_j^2 - n\,\overline{\boldsymbol{X}}^2 = n\,\widehat{\sigma_{\boldsymbol{X}}^2} = (n-1)\,S_{\boldsymbol{X}}^2 ,$$

$n\,\widehat{\sigma_{\boldsymbol{X}}^2} = (n-1)\,S_{\boldsymbol{X}}^2$	má rozdělení	$\chi^2(n-1)$,
$S_{\boldsymbol{X}}^2$	má rozdělení	$\frac{\chi^2(n-1)}{n-1}$,
$\widehat{\sigma_{\boldsymbol{X}}^2}$	má rozdělení	$\frac{\chi^2(n-1)}{n}$.

3. Pro $X_1, \ldots, X_n$ s rozdělením $\mathrm{N}(\mu, \sigma^2)$:
μ nemá vliv, σ^2 se vše vynásobilo,

$S_{\boldsymbol{X}}^2$	$\sim$	$\frac{\chi^2(n-1)}{n-1}\,\sigma^2$,
$\frac{S_{\boldsymbol{X}}^2}{\sigma^2}$	$\sim$	$\frac{\chi^2(n-1)}{n-1}$,
$\widehat{\sigma_{\boldsymbol{X}}^2}$	$\sim$	$\frac{\chi^2(n-1)}{n}\,\sigma^2$,
$\frac{n\,\widehat{\sigma_{\boldsymbol{X}}^2}}{\sigma^2} = \frac{(n-1)\,S_{\boldsymbol{X}}^2}{\sigma^2}$	$\sim$	$\chi^2(n-1)$.

Důsledky:

$$\mathrm{E}S_{\boldsymbol{X}}^2 = \frac{n-1}{n-1}\,\mathrm{D}X = \mathrm{D}X \qquad \text{(to už víme i obecně)},$$
$$\mathrm{D}S_{\boldsymbol{X}}^2 = \frac{2(n-1)}{(n-1)^2}\,(\mathrm{D}X)^2 = \frac{2}{n-1}\,(\mathrm{D}X)^2 \to 0 \qquad \text{pro } n \to \infty .$$

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

8.6.4 Eficience odhadů rozptylu pro normální rozdělení

1. eficience odhadu $S_{\boldsymbol{X}}^2$ (z vlastností rozdělení χ^2):

$$\mathrm{E}(S_{\boldsymbol{X}}^2 - \mathrm{D}X)^2 = \mathrm{D}S_{\boldsymbol{X}}^2 = \frac{2}{n-1}\,(\mathrm{D}X)^2 .$$

2. eficience odhadu $\widehat{\sigma_{\boldsymbol{X}}^2}$ ($\mathrm{D}X$ je konstanta):

$$\begin{aligned} \mathrm{E}(\widehat{\sigma_{\boldsymbol{X}}^2} - \mathrm{D}X)^2 &= \mathrm{D}(\widehat{\sigma_{\boldsymbol{X}}^2} - \mathrm{D}X) + (\mathrm{E}(\widehat{\sigma_{\boldsymbol{X}}^2} - \mathrm{D}X))^2 = \\ &= \mathrm{D}(\widehat{\sigma_{\boldsymbol{X}}^2}) + (\frac{1}{n}\,\mathrm{D}X)^2 = \\ &= \left(\frac{n-1}{n}\right)^2 \frac{2}{n-1}\,(\mathrm{D}X)^2 + \frac{1}{n^2}\,(\mathrm{D}X)^2 = \frac{2n-1}{n^2}\,(\mathrm{D}X)^2 , \end{aligned}$$

a protože

$$\frac{2n-1}{n^2} < \frac{2}{n} < \frac{2}{n-1}\,,$$

je odhad $\widehat{\sigma^2_{\boldsymbol{X}}}$ více eficientní než $S^2_{\boldsymbol{X}}$ (který je nejlepší nestranný!).

8.7 Odhad směrodatné odchylky

pomocí směrodatné odchylky empirického rozdělení:

$$\widehat{\sigma_{\boldsymbol{x}}} := \sigma_{\mathrm{Emp}(\boldsymbol{x})} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{\boldsymbol{x}})^2}\,.$$

Je realizací odhadu

$$\widehat{\sigma_{\boldsymbol{X}}} := \sqrt{\frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{\boldsymbol{X}})^2}\,.$$

Je vychýlený. Alternativa:

$$S_{\boldsymbol{X}} = \sqrt{S^2_{\boldsymbol{X}}} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{\boldsymbol{X}}_n)^2} \qquad (\textbf{výběrová směrodatná odchylka})\,,$$

$$s_{\boldsymbol{x}} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{\boldsymbol{x}}_n)^2} \qquad (\textbf{realizace výběrové směrodatné odchylky})\,.$$

Alternativní značení: S, s

Věta:

$$\mathrm{E}S_{\boldsymbol{X}} \le \sigma_X\,.$$

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

Důkaz:

$$\mathrm{D}X = \mathrm{E}S^2_{\boldsymbol{X}} = (\mathrm{E}S_{\boldsymbol{X}})^2 + \underbrace{\mathrm{D}S_{\boldsymbol{X}}}_{\ge 0} \ge (\mathrm{E}S_{\boldsymbol{X}})^2\,,$$

$$\sigma_X \ge \mathrm{E}S_{\boldsymbol{X}}\,.$$

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

8.8 Histogram a popis empirického rozdělení

V realizaci náhodného výběru $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ 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 četností** nebo grafem zvaným **histogram**.

Normováním dostaneme **relativní četnosti** $r_t := \dfrac{n_t}{n} = p_{\mathrm{Emp}(\boldsymbol{x})}(t)$ (=hodnoty pravděpodobnostní funkce empirického rozdělení $\mathrm{Emp}(\boldsymbol{x})$), $t \in H$, kde $\sum_{t\in H} r_t = 1$.

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

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

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

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

$$= \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{\boldsymbol{x}})^2 = \widehat{\sigma^2_{\boldsymbol{x}}} = \frac{n-1}{n}\, s^2_{\boldsymbol{x}}\,.$$

8.9 Odhad mediánu

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

Výběrový medián minimalizuje kritérium

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

Navíc víme, jak se změní monotónní funkcí h: $q_{\mathrm{Emp}(h(\boldsymbol{x}))}(\frac{1}{2}) = h(q_{\mathrm{Emp}(\boldsymbol{x})}(\frac{1}{2}))\,.$
Proč se používá méně než výběrový průměr:

- Vyšší výpočetní náročnost: seřazení hodnot má pracnost úměrnou $n \ln n$, výběrový průměr n, a i když existují algoritmy počítající medián s lineární složitostí, jsou komplikované.
- Vyšší paměťová náročnost: úměrná n, u výběrového průměru stačí 2 registry.
- Obtížná decentralizace a paralelizace výpočtu.

Obecněji lze odhadnout α-kvantil $q_X(\alpha)$ pomocí α-kvantilu empirického rozdělení, $q_{\mathrm{Emp}(\boldsymbol{x})}(\alpha)$. Nesmíme však volit α blízké 0 nebo 1, nemůžeme např. na základě výběru rozsahu 1000 odhadovat kvantil $q_X(10^{-6})$.
Pokud známe model rozdělení, můžeme jeho parametry odhadnout pomocí **všech** hodnot ve výběrovém souboru a z parametrů pak odhadnout požadované kvantily.

8.10 Intervalové odhady

Dosud jsme skutečnou hodnotu parametru ϑ^* nahrazovali **bodovým odhadem** $\widehat{\Theta}$ (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,

$$I = (-\infty, q_{\widehat{\Theta}}(1-\alpha)\rangle,\ \text{resp. } I = \langle q_{\widehat{\Theta}}(\alpha), \infty)\,,$$

nebo (**symetrický**) **oboustranný** odhad,

$$I = \langle q_{\widehat{\Theta}}(\tfrac{\alpha}{2}), q_{\widehat{\Theta}}(1-\tfrac{\alpha}{2})\rangle\,.$$

K tomu potřebujeme znát rozdělení odhadu $\widehat{\Theta}$.

8.11 Intervalové odhady parametrů normálního rozdělení

Předpoklad: Náhodná veličina $X \sim \mathrm{N}(\mu, \sigma^2)$.

8.11.1 Intervalový odhad střední hodnoty při známém rozptylu σ^2

σ^2 známe; X odhadneme z realizace x; μ neznáme
S pravděpodobností $1-\alpha$ je $q_{\mathrm{N}(\mu,\sigma^2)}(\frac{\alpha}{2}) \leq X \leq q_{\mathrm{N}(\mu,\sigma^2)}(1-\frac{\alpha}{2})\,,$
po znormování

$$-\Phi^{-1}(1-\tfrac{\alpha}{2}) = \Phi^{-1}(\tfrac{\alpha}{2}) \leq \operatorname{norm} X = \frac{X-\mu}{\sigma} \leq \Phi^{-1}(1-\tfrac{\alpha}{2})\,,$$

$$X - \sigma\,\Phi^{-1}(1-\tfrac{\alpha}{2}) \leq \mu \leq X + \sigma\,\Phi^{-1}(1-\tfrac{\alpha}{2})\,.$$

Obvykle místo jedné realizace použijeme realizaci výběrového průměru $\overline{\boldsymbol{X}}_n \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$; po normalizaci norm $\overline{\boldsymbol{X}}_n = \frac{\sqrt{n}}{\sigma}(\overline{\boldsymbol{X}}_n - \mu) \sim \mathrm{N}(0,1)$.
S pravděpodobností $1-\alpha$ je

$$\overline{\boldsymbol{X}}_n - \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\tfrac{\alpha}{2}) \le \mu \le \overline{\boldsymbol{X}}_n + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\tfrac{\alpha}{2}),$$

pro jednostranné odhady

$$\mu \le \overline{\boldsymbol{X}}_n + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\alpha),$$
$$\overline{\boldsymbol{X}}_n - \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\alpha) \le \mu,$$

Dostali jsme intervalové odhady pro μ

$$\left\langle \overline{\boldsymbol{X}} - \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\tfrac{\alpha}{2}), \overline{\boldsymbol{X}} + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\tfrac{\alpha}{2}) \right\rangle,$$
$$\left(-\infty, \overline{\boldsymbol{X}} + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\alpha)\right\rangle,$$
$$\left\langle \overline{\boldsymbol{X}} - \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\alpha), \infty\right).$$

Při výpočtu nahradíme výběrový průměr $\overline{\boldsymbol{X}}_n$ jeho realizací $\bar{\boldsymbol{x}}_n$.
Díky centrální limitní větě je odhad použitelný i pro výběr z jiného než normálního rozdělení, pokud má (nenulový) rozptyl a rozsah výběru je velký.

8.11.2 Intervalový odhad střední hodnoty při neznámém rozptylu

σ^2 neznáme, ale odhadneme z realizace výběrového rozptylu $S^2_{\boldsymbol{X}} \sim \frac{\sigma^2}{n-1}\chi^2(n-1)$;
$\overline{\boldsymbol{X}}_n \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$ odhadneme z realizace $\bar{\boldsymbol{x}}_n$;
μ neznáme

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

8.11.3 Studentovo t-rozdělení (autor: Gossett)

s n stupni volnosti je rozdělení náhodné veličiny $\frac{U}{\sqrt{\frac{V}{n}}}$ $\left(= \frac{U}{\sqrt{W}}\right)$,

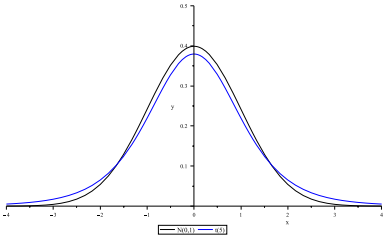
kde $U \sim \mathrm{N}(0,1)$,
$V \sim \chi^2(n)$ $\left(\frac{V}{n} = W \sim \frac{\chi^2(n)}{n}\right)$,
U, V jsou nezávislé (U, W jsou nezávislé).
Značení: $\mathrm{t}(n)$.
Hustota:

$$f_{\mathrm{t}(n)}(x) = c(n)\left(1 + \frac{x^2}{n}\right)^{-\frac{1+n}{2}}, \qquad c(n) = \frac{\Gamma(\frac{1+n}{2})}{\sqrt{n\,\pi}\,\Gamma(\frac{n}{2})}.$$

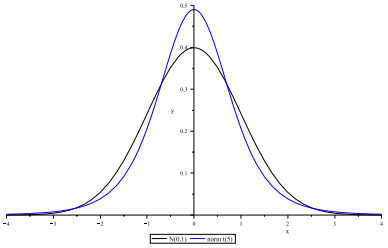
Symetrie kolem nuly $\Rightarrow q_{\mathrm{t}(n)}(1-\alpha) = -q_{\mathrm{t}(n)}(\alpha)$.
$\mathrm{t}(1)$ je Cauchyho rozdělení, které nemá střední hodnotu,

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

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 **normovaného** Studentova rozdělení s 5 stupni volnosti.

8.11.4 Intervalový odhad střední hodnoty při neznámém rozptylu 2

V našem případě:

$$U = \frac{\sqrt{n}}{\sigma}\,(\overline{\boldsymbol{X}}_n - \mu) \sim \mathrm{N}(0,1)\,,$$
$$V = \frac{(n-1)\,S^2_{\boldsymbol{X}}}{\sigma^2} \sim \chi^2(n-1)\,,$$
$$\frac{U}{\sqrt{\frac{V}{n-1}}} = \frac{\frac{\sqrt{n}}{\sigma}\,(\overline{\boldsymbol{X}}_n - \mu)}{\sqrt{\frac{S^2_{\boldsymbol{X}}}{\sigma^2}}} = \frac{\sqrt{n}}{S_{\boldsymbol{X}}}\,(\overline{\boldsymbol{X}}_n - \mu) \sim \mathrm{t}(n-1)\,.$$

Z toho vyplývají intervalové odhady

$$\left\langle \overline{\boldsymbol{X}}_n - \frac{S_{\boldsymbol{X}}}{\sqrt{n}}\,q_{\mathrm{t}(n-1)}(1-\tfrac{\alpha}{2}), \overline{\boldsymbol{X}}_n + \frac{S_{\boldsymbol{X}}}{\sqrt{n}}\,q_{\mathrm{t}(n-1)}(1-\tfrac{\alpha}{2}) \right\rangle ,$$
$$\left(-\infty, \overline{\boldsymbol{X}}_n + \frac{S_{\boldsymbol{X}}}{\sqrt{n}}\,q_{\mathrm{t}(n-1)}(1-\alpha) \right\rangle ,$$
$$\left\langle \overline{\boldsymbol{X}}_n - \frac{S_{\boldsymbol{X}}}{\sqrt{n}}\,q_{\mathrm{t}(n-1)}(1-\alpha), \infty \right) .$$

Při výpočtu nahradíme výběrový průměr $\overline{\boldsymbol{X}}_n$ jeho realizací $\bar{\boldsymbol{x}}_n$ a výběrovou směrodatnou odchylku $S_{\boldsymbol{X}}$ její realizací $s_{\boldsymbol{x}}$.
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).

8.11.5 Intervalový odhad rozptylu

μ nás nezajímá;

σ^2 odhadneme výběrovým rozptylem $S^2_{\boldsymbol{X}} \sim \dfrac{\chi^2(n-1)}{n-1}\,\sigma^2$;

$\frac{(n-1)\,S^2_{\boldsymbol{X}}}{\sigma^2} \sim \chi^2(n-1)$.
S pravděpodobností $1-\alpha$ je

$$q_{\chi^2(n-1)}(\tfrac{\alpha}{2}) \le \frac{(n-1)\,S^2_{\boldsymbol{X}}}{\sigma^2} \le q_{\chi^2(n-1)}(1-\tfrac{\alpha}{2})\,,$$
$$\frac{(n-1)\,S^2_{\boldsymbol{X}}}{q_{\chi^2(n-1)}(1-\frac{\alpha}{2})} \le \sigma^2 \le \frac{(n-1)\,S^2_{\boldsymbol{X}}}{q_{\chi^2(n-1)}(\frac{\alpha}{2})}\,,$$

jednostranné odhady

$$\sigma^2 \leq \frac{(n-1)\,S_{\boldsymbol{X}}^2}{q_{\chi^2(n-1)}(\alpha)}\,,$$
$$\frac{(n-1)\,S_{\boldsymbol{X}}^2}{q_{\chi^2(n-1)}(1-\alpha)} \leq \sigma^2\,.$$

Alternativní formulace:
σ^2 odhadneme výběrovým rozptylem $S_{\boldsymbol{X}}^2 \sim \dfrac{\chi^2(n-1)}{n-1}\,\sigma^2$;
$\dfrac{S_{\boldsymbol{X}}^2}{\sigma^2} \sim \dfrac{\chi^2(n-1)}{n-1}$.
S pravděpodobností $1-\alpha$ je

$$q_{\frac{\chi^2(n-1)}{n-1}}\left(\tfrac{\alpha}{2}\right) \leq \frac{S_{\boldsymbol{X}}^2}{\sigma^2} \leq q_{\frac{\chi^2(n-1)}{n-1}}\left(1-\tfrac{\alpha}{2}\right),$$
$$\frac{S_{\boldsymbol{X}}^2}{q_{\frac{\chi^2(n-1)}{n-1}}\left(1-\frac{\alpha}{2}\right)} \leq \sigma^2 \leq \frac{S_{\boldsymbol{X}}^2}{q_{\frac{\chi^2(n-1)}{n-1}}\left(\frac{\alpha}{2}\right)}\,,$$

jednostranné odhady

$$\sigma^2 \leq \frac{S_{\boldsymbol{X}}^2}{q_{\frac{\chi^2(n-1)}{n-1}}(\alpha)}\,,$$
$$\frac{S_{\boldsymbol{X}}^2}{q_{\frac{\chi^2(n-1)}{n-1}}(1-\alpha)} \leq \sigma^2\,.$$

Dostali jsme intervalové odhady pro σ^2

$$\left\langle \frac{(n-1)\,S_{\boldsymbol{X}}^2}{q_{\chi^2(n-1)}(1-\frac{\alpha}{2})}, \frac{(n-1)\,S_{\boldsymbol{X}}^2}{q_{\chi^2(n-1)}(\frac{\alpha}{2})} \right\rangle, \qquad \left\langle \frac{S_{\boldsymbol{X}}^2}{q_{\frac{\chi^2(n-1)}{n-1}}(1-\frac{\alpha}{2})}, \frac{S_{\boldsymbol{X}}^2}{q_{\frac{\chi^2(n-1)}{n-1}}(\frac{\alpha}{2})} \right\rangle,$$
$$\left(-\infty, \frac{(n-1)\,S_{\boldsymbol{X}}^2}{q_{\chi^2(n-1)}(\alpha)}\right\rangle, \qquad \left(-\infty, \frac{S_{\boldsymbol{X}}^2}{q_{\frac{\chi^2(n-1)}{n-1}}(\alpha)}\right\rangle,$$
$$\left\langle \frac{(n-1)\,S_{\boldsymbol{X}}^2}{q_{\chi^2(n-1)}(1-\alpha)}, \infty\right), \qquad \left\langle \frac{S_{\boldsymbol{X}}^2}{q_{\frac{\chi^2(n-1)}{n-1}}(1-\alpha)}, \infty\right).$$

Při výpočtu nahradíme výběrový rozptyl $S_{\boldsymbol{X}}^2$ jeho realizací $s_{\boldsymbol{x}}^2$.
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 rozdělení $\chi^2(n-1)$ použít normální $\mathrm{N}(n-1, 2(n-1))$).

8.11.6 Intervalové odhady spojitých rozdělení, která nejsou normální

převádíme obvykle na normální rozdělení nelineární neklesající transformací

$$h(\mathrm{t}) = \Phi^{-1}(F_X(\mathrm{t}))$$

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

8.12 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 $\langle 0, 1\rangle$, jejichž součet je 1.
Motivační příklad (směs normálních rozdělení):
Spojité rozdělení, jehož hustota má více maxim, aproximujeme směsí normálních,

$$\mathrm{Mix}_{(c_1,\dots,c_m)}(\mathrm{N}(\mu_1,\sigma_1^2),\dots,\mathrm{N}(\mu_m,\sigma_m^2))\,,$$

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 $(c_1,\dots,c_m;\mu_1,\dots,\mu_m;\sigma_1,\dots,\sigma_m)$ s omezujícími podmínkami

$$\begin{aligned} \sigma_i &\geq 0\,, & i &= 1,\dots,m\,,\\ 0 \leq c_i &\leq 1\,, & i &= 1,\dots,m\,,\\ \sum_{i=1}^{m} c_i &= 1\,. \end{aligned}$$

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

8.12.1 Metoda momentů

(angl. *moment matching*)
Pro $k=1,2,\dots$ lze k-tý obecný moment vypočítat z modelu jako funkci $\boldsymbol{\vartheta}$,

$$\mathrm{E}X^k(\boldsymbol{\vartheta}) = \mathrm{E}X^k(\vartheta_1,\dots,\vartheta_i)$$

a současně odhadnout pomocí výběrového k-tého obecného momentu

$$m_{\boldsymbol{x}^k} = \frac{1}{n}\sum_{j=1}^{n} x_j^k\,.$$

Metoda momentů doporučuje realizaci odhadu $\widehat{\boldsymbol{\vartheta}}=(\widehat{\vartheta}_1,\dots,\widehat{\vartheta}_i)$ takovou, že

$$\mathrm{E}X^k(\widehat{\vartheta}_1,\dots,\widehat{\vartheta}_i) = \frac{1}{n}\sum_{j=1}^{n} x_j^k\,.$$

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

Použitelnost metody momentů

Možné problémy:

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

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

3. Některé z rovnic nemusí být použitelné, protože nezávisí na parametrech nebo je nelze splnit. (Např. střední hodnota vychází teoreticky nulová, její odhad nulový být nemusí.)
   Takové rovnice vyloučíme.

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

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

6. Může být obtížné řešení nalézt.

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

8. Můžeme dospět k řešení, které **nesplňuje předpoklady**, $\widehat{\boldsymbol{\vartheta}} \notin \Pi$ (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.

8.12.2 Metoda maximální věrohodnosti *(angl. likelihood)*

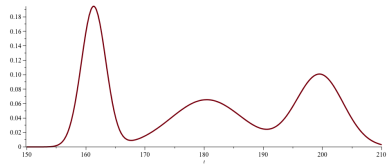
Motivace: Occamova břitva
„Přikloňme se k nejjednodušššímu a nejpřirozenějšímu vysvětlení“

Příklad. 1000 *respondentů vyjádřilo preference* 3 *politickým stranám následovně:*

A	B	C
200	300	500

Strana A by měla uznat, že její preference jsou malé, místo aby výsledek přičítala statistické chybě nebo zfalšování průzkumu.

Příklad. *V populaci nám vyšlo následující rozdělení parametru (např. tělesná výška):*



Rozumný závěr je, že populaci tvoří 3 *skupiny, v nichž je rozdělení podobné normálnímu (s různými středními hodnotami). Není vhodné usuzovat, že celé rozdělení je normální, nebo že ho tvoří* 42 *takových skupin.*

Problém: Pro globální oteplování jsou podobné argumenty, ale ne tak jednoznačné.

Pravidlo: Z navržených modelů vybíráme ten, pro který jsou naměřené hodnoty „nejméně překvapivé.“

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

$$L(\boldsymbol{\vartheta}) := p_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\vartheta}) = P(X_1 = x_1 \wedge \ldots \wedge X_n = x_n;\boldsymbol{\vartheta}) = \\ = \prod_{j=1}^{n} P(X_j = x_j;\boldsymbol{\vartheta}) = \prod_{j=1}^{n} p_X(x_j;\boldsymbol{\vartheta})\,.$$

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

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

(Nutno vyloučit případ $p_X(x_j; \boldsymbol{\vartheta}) = 0$, 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**

$$\Lambda(\boldsymbol{\vartheta}) := f_{\boldsymbol{X}}(\boldsymbol{x}; \boldsymbol{\vartheta}) = \prod_{j=1}^{n} f_X(x_j; \boldsymbol{\vartheta}) \,.$$

Nicméně i tato funkce $\Lambda \colon \Pi \to \langle 0, \infty)$, $\Pi \subseteq \mathbb{R}^i$, 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.

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

(Nutno vyloučit případ $f_X(x_j; \boldsymbol{\vartheta}) = 0$, 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 souvislý.)
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.

8.12.3 Příklady na odhady parametrů

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

výsledek s	1	2	3	$\sum$
pravděpodobnost	a	b	c	1
četnost n_s	10	12	10	32

Řešení. ***Metoda momentů:***

$$a+b+c=1\,,$$
$$\mathrm{E}X=\sum_s s\,p_s=a+2\,b+3\,c=\frac{1}{n}\sum_s s\,n_s=\frac{10+2\cdot 12+3\cdot 10}{32}=2\,,$$
$$\mathrm{E}X^2=\sum_s s^2\,p_s=a+4\,b+9\,c=\frac{1}{n}\sum_s s^2\,n_s=\frac{10+4\cdot 12+9\cdot 10}{32}=\frac{37}{8}\,,$$
$$a=c=\frac{5}{16}\,,\qquad b=\frac{6}{16}\,.$$

Metoda maximální věrohodnosti:

$$L(a,b)=a^{10}\cdot b^{12}\cdot\underbrace{(1-a-b)}_{c}{}^{10}\,,$$
$$\ell(b)=\ln\ell(a,b)=10\,\ln a+12\,\ln b+10\,\ln(1-a-b)\,,$$
$$0=\frac{\partial}{\partial a}\ell(a,b)=\frac{10}{a}-\frac{10}{1-a-b}=\frac{10}{a}-\frac{10}{c}\,,$$
$$0=\frac{\partial}{\partial b}\ell(a,b)=\frac{12}{b}-\frac{10}{1-a-b}=\frac{12}{b}-\frac{10}{c}\,,$$
$$a=c=\frac{5}{6}\,b\,,$$

a z jednotkového součtu pravděpodobností opět

$$a=c=\frac{5}{16}\,,\qquad b=\frac{6}{16}\,.$$

□

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

Věta. *Pokud nějaká hodnota vektoru parametrů odpovídá empirickému rozdělení, pak je odhadem podle metody momentů i maximálně věrohodným odhadem.*

Optimálních hodnot parametrů může být víc (a mohou vést na stejné rozdělení).

Důkaz. Označme $u_1,\dots,u_i$ $(i\le n)$ všechny **různé** hodnoty, které se vyskytly v realizaci $\boldsymbol{x}$, n_s četnost a $r_s=n_s/n$ relativní četnost hodnoty u_s. Máme odhadnout pravděpodobnosti

$$q_s=p_X(u_s)\,,\qquad s=1,\dots,i\,,\qquad \sum_{s=1}^{i}q_s=1\,.$$

Připomeňme, že empirické rozdělení $\mathrm{Emp}(\boldsymbol{x})$ nabývá hodnot $u_1,\dots,u_i$ s pravděpodobnostmi po řadě $r_1,\dots,r_i$.

Metoda maximální věrohodnosti:

$$L(q_1,\dots,q_i)=\prod_{j=1}^{n}p_X(x_j)=\prod_{s=1}^{i}(p_X(u_s))^{n_s}=\prod_{s=1}^{i}q_s^{n_s}\,,$$
$$\ell(q_1,\dots,q_i)=\ln L(q_1,\dots,q_i)=\sum_{s=1}^{i}n_s\,\ln q_s\,.$$

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

$$h(c, q_1, \dots, q_i) = \sum_{s=1}^{i} n_s \, \ln q_s + c \left(1 - \sum_{s=1}^{i} q_s\right),$$
$$0 = \frac{\partial}{\partial q_s} h(c, q_1, \dots, q_i) = \frac{n_s}{q_s} - c\,.$$

Hodnota $\frac{n_s}{q_s} = c$ je nezávislá na $s \in \{1, \dots, i\}$. 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:

$$\mathrm{E}X^k = \sum_{s=1}^{i} q_s \, u_s^k = \frac{1}{n} \sum_{j=1}^{n} x_j^k = \frac{1}{n} \sum_{s=1}^{i} n_s \, u_s^k = \sum_{s=1}^{i} r_s \, u_s^k\,.$$

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

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

(tzv. Vandermondova matice) je regulární, právě když čísla $u_1, \dots, u_i$ jsou navzájem různá. □

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:***

$$\mathrm{E}X = \frac{a+b}{2} = \frac{1}{n} \sum_{s=1}^{n} x_s\,,$$
$$\mathrm{E}X^2 = (\mathrm{E}X)^2 + \mathrm{D}X = \left(\frac{a+b}{2}\right)^2 + \underbrace{\frac{(b-a)^2}{12}}_{\mathrm{D}X} = \frac{a^2 + a\,b + b^2}{3} = \frac{1}{n} \sum_{s=1}^{n} x_s^2\,.$$

1. Soustava

$$\mathrm{E}X = \frac{a+b}{2} = \frac{24}{5}\,,$$
$$\mathrm{E}X^2 = \frac{a^2 + a\,b + b^2}{3} = \frac{148}{5}\,,$$

má 2 řešení

$$a \doteq 0.36\,, \qquad b \doteq 9.24\,,$$
$$a \doteq 9.24\,, \qquad b \doteq 0.36$$

První řešení je jediné správné.

2. Soustava

$$\mathrm{E}X = \frac{a+b}{2} = 1\,,$$
$$\mathrm{E}X^2 = \frac{a^2+a\,b+b^2}{3} = 5\,,$$

má 2 řešení

$$a \doteq -2.5\,, \qquad b \doteq 4.5\,,$$
$$a \doteq 4.5\,, \qquad b \doteq -2.5\,.$$

Žádné není správné, neboť $5 \notin \langle a, b\rangle$.

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

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

pokud $x_j \in \langle a, b\rangle$ *pro všechna* j*; jinak je nulová. Věrohodnost je maximální, pokud* $b-a$ *je minimální, tj.*

$$a = \min_j x_j, \qquad b = \max_j x_j\,.$$

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

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

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

$$\mathrm{E}X = \mu\,, \qquad \mathrm{E}X^2 = (\mathrm{E}X)^2 + \mathrm{D}X = \mu^2 + \sigma^2 = \mu^2 + r\,.$$

Pro odhady $\widehat{\mu}, \widehat{r}$ máme soustavu rovnic

$$\widehat{\mu} = \frac{1}{n}\sum_{j=1}^{n} x_j\,,$$
$$\widehat{\mu}^2 + \widehat{r} = \frac{1}{n}\sum_{j=1}^{n} x_j^2\,.$$

Řešení:

$$\widehat{\mu} = \bar{\boldsymbol{x}}\,,$$
$$\widehat{r} = \frac{1}{n}\sum_{j=1}^{n} x_j^2 - \widehat{\mu}^2 = \frac{1}{n}\sum_{j=1}^{n} x_j^2 - \bar{\boldsymbol{x}}^2 = \widehat{\sigma_X^2} = \mathrm{D}\,\mathrm{Emp}(\boldsymbol{x})\,.$$

Metoda maximální věrohodnosti:

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

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

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

$$\Rightarrow \widehat{\mu} = \bar{\boldsymbol{x}}\,,$$

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

$$\Rightarrow \widehat{\widehat{r}} = \frac{1}{n} \sum_{j=1}^{n} (x_j - \bar{\boldsymbol{x}})^2 = \widehat{\sigma_X^2} = \mathrm{D}\,\mathrm{Emp}(\boldsymbol{x})\,.$$

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

$$\mathrm{Mix}_{(c_1,\ldots,\,c_m)}(\mathrm{N}(\mu_1, \sigma_1^2), \ldots, \mathrm{N}(\mu_m, \sigma_m^2))\,,$$

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

$$\begin{aligned} \sigma_i &\geq 0\,, & i &= 1, \ldots, m\,, \\ 0 \leq c_i &\leq 1\,, & i &= 1, \ldots, m\,, \\ \sum_{i=1}^{m} c_i &= 1\,. \end{aligned}$$

Pokus o řešení:

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

$$\begin{aligned} \Lambda(\mu, \mathbf{c}) &= \prod_{j=1}^{n} f_X(x_j) = \prod_{j=1}^{n} \sum_{i=1}^{m} c_i\, f_{\mathrm{N}(\mu_i, \sigma^2)}(x_j) = \\ &= \prod_{j=1}^{n} \left(\frac{1}{\sqrt{2\,\pi}\,\sigma} \sum_{i=1}^{m} c_i \exp\left(\frac{-(x_j - \mu_i)^2}{2\,\sigma^2} \right) \right) = \\ &= \left(\frac{1}{\sqrt{2\,\pi}\,\sigma} \right)^n \prod_{j=1}^{n} \sum_{i=1}^{m} c_i \exp\left(\frac{-(x_j - \mu_i)^2}{2\,\sigma^2} \right), \end{aligned}$$

$$\begin{aligned} \lambda(\mu, \mathbf{c}) &= \sum_{j=1}^{n} \ln \sum_{i=1}^{m} c_i\, f_{\mathrm{N}(\mu_i, \sigma^2)}(x_j) = \\ &= -n \ln(\sqrt{2\,\pi}\,\sigma) + \sum_{j=1}^{n} \ln \sum_{i=1}^{m} c_i \exp\left(\frac{-(x_j - \mu_i)^2}{2\,\sigma^2} \right). \end{aligned}$$

Věrohodnost se těžko maximalizuje přímo, používá se iterační metoda:

EM algoritmus

EM (Expectation-Maximization) [Dempster, Laird, and Rubin 1977, M.I. Schlesinger 1968, US Army ∼1950].
Stupeň příslušnosti x_j ke i-té složce směsi popíšeme koeficientem $\alpha_{j,i} \in \langle 0, 1 \rangle$, 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 μ_i a nenulové koeficienty c_i, $i = 1, ..., m$, splňující $\sum_{i=1}^{m} c_i = 1$.

E. Stanovíme stupně příslušnosti

$$\alpha_{j,i} := \frac{c_i \, f_{\mathrm{N}(\mu_i,\sigma^2)}(x_j)}{\sum\limits_{i'=1}^{m} c_{i'} \, f_{\mathrm{N}(\mu_{i'},\sigma^2)}(x_j)} = \frac{c_i \, \exp\left(\frac{-(x_j-\mu_i)^2}{2\,\sigma^2}\right)}{\sum\limits_{i'=1}^{m} \left(c_{i'} \, \exp\left(\frac{-(x_j-\mu_{i'})^2}{2\,\sigma^2}\right)\right)}$$

(jmenovatel je normalizační faktor).

M. Aktualizujeme koeficienty složek směsi

$$c_i := \frac{\sum\limits_{j=1}^{n} \alpha_{j,i}}{\sum\limits_{i'=1}^{m} \sum\limits_{j=1}^{n} \alpha_{j,i'}} = \frac{1}{n} \sum_{j=1}^{n} \alpha_{j,i}$$

a střední hodnoty složek jako těžiště hodnot realizace vážených stupni příslušnosti,

$$\mu_i := \frac{\sum\limits_{j=1}^{n} \alpha_{j,i} \, x_j}{\sum\limits_{j=1}^{n} \alpha_{j,i}} = \frac{\sum\limits_{j=1}^{n} \alpha_{j,i} \, x_j}{n \, c_i} \, .$$

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 směsí rozdělení je typické, ne však jediné možné.
Problém: Uvíznutí v lokálním extrému.
EM algoritmus rozšiřuje možnosti použití metody maximální věrohodnosti.

Použití empirického rozdělení $\mathrm{Emp}(\boldsymbol{x})$ **v odhadech**

veličina	realizace odhadu	nestranný
$\mathrm{E}X$	$\mathrm{E}\,\mathrm{Emp}(\boldsymbol{x}) = \frac{1}{n} \sum\limits_i x_i = \bar{\boldsymbol{x}}$	+
$\mathrm{E}X^k$	$\mathrm{E}\big(\mathrm{Emp}(\boldsymbol{x})^k\big) = \frac{1}{n} \sum\limits_i x_i^k = m_k$	+
$\mathrm{D}X$	$\mathrm{D}\,\mathrm{Emp}(\boldsymbol{x}) = \frac{1}{n} \sum\limits_i (x_i - \bar{\boldsymbol{x}})^2$	−
	$\frac{n}{n-1}\,\mathrm{D}\,\mathrm{Emp}(\boldsymbol{x}) = \frac{1}{n-1} \sum\limits_i (x_i - \bar{\boldsymbol{x}})^2 = s_{\boldsymbol{x}}^2$	+
σ_X	$\sigma_{\mathrm{Emp}(\boldsymbol{x})} = \sqrt{\frac{1}{n} \sum\limits_i (x_i - \bar{\boldsymbol{x}})^2}$	−
	$\sqrt{\frac{n}{n-1}}\,\sigma_{\mathrm{Emp}(\boldsymbol{x})} = \sqrt{\frac{1}{n-1} \sum\limits_i (x_i - \bar{\boldsymbol{x}})^2} = s_{\boldsymbol{x}}$	−
$q_X(\frac{1}{2})$	$q_{\mathrm{Emp}(\boldsymbol{x})}(\frac{1}{2})$	?
$q_X(\alpha)$	$q_{\mathrm{Emp}(\boldsymbol{x})}(\alpha)$	?
p_X bez omezení	$p_{\mathrm{Emp}(\boldsymbol{x})}$	+
p_X s omezením	?	?
f_X	–	?

? = záleží na okolnostech

9 Testování hypotéz

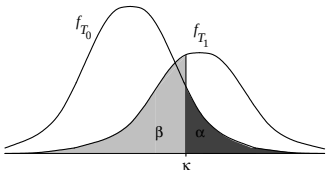
9.1 Základní pojmy a principy testování hypotéz

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

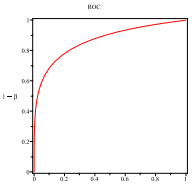
Předpoklad: Parametr ϑ nabývá pouze 2 hodnot, 0 pro „normální“ populaci, 1 pro „anomální“ prvky. O prvku máme rozhodnout, ke které skupině patří (tj. odhadnout ϑ). K tomu použijeme testovací statistiku T (resp. její realizaci t). Ta závisí na ϑ. Předpokládejme, že obě skupiny mají známá rozdělení statistiky T, která pro anomální skupinu nabývá „větších“ 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:

$$\begin{aligned} &\text{pro} \quad T \leq \kappa \quad \text{normální,} \\ &\text{pro} \quad T > \kappa \quad \text{anomální.} \end{aligned}$$

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): Zamítneme nulovou hypotézu, která platí. Normální je klasifikován jako anomální s pravděpodobností $\alpha(\kappa)$ (nerostoucí funkce κ).
Chyba 2. druhu (osvobodíme vinného): Nezamítneme nulovou hypotézu, která neplatí. Anomální je klasifikován jako normální s pravděpodobností $\beta(\kappa)$ (neklesající funkce κ).



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



Typický průběh **ROC** křivky

Možná kritéria pro volbu prahu κ:

- $\alpha(\kappa) = \beta(\kappa)$,
- $\min\limits_{\kappa}(\alpha(\kappa) + \beta(\kappa))$,
- $\min\limits_{\kappa} e(\alpha(\kappa), \beta(\kappa))$, např. $\min\limits_{\kappa}(a\,\alpha(\kappa) + b\,\beta(\kappa))$, tj. minimalizace **výplatní funkce**,
- $\alpha(\kappa) =$ 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 κ stanovíme tak, aby chyba 1. druhu nastávala s danou pravděpodobností α 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 nepotvrzujeme**, neboť tím bychom se mohli dopustit chyby 2. druhu s blíže neurčenou pravděpodobností β.
Slovníček pojmů (pro porozumění jiným textům, zde se téměř nepoužijí)

skutečnost / test	anomální H_0 neplatí	normální H_0 platí	celkem
pozitivní, H_0 zamítnuta	TP	FP	P'
negativní, H_0 nezamítnuta	FN	TN	N'
celkem	P	N	

(Položky v tabulce mohou být pravděpodobnosti empirického nebo skutečného rozdělení nebo empirické četnosti.)
TP skutečně 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*)
$\alpha = \dfrac{FP}{N} = \dfrac{FP}{TN+FP}$ = pravděpodobnost chyby 1. druhu
$\beta = \dfrac{FN}{P} = \dfrac{FN}{TP+FN}$ = pravděpodobnost chyby 2. druhu

$\dfrac{TP}{TP+FN} = \dfrac{TP}{P} = 1-\beta$ — senzitivita, síla, míra skutečně pozitivních *sensitivity, recall, true positive rate*

$\dfrac{TN}{TN+FP} = \dfrac{TN}{N} = 1-\alpha$ — specificita, míra skutečně negativních *specificity, true negative rate*

$\dfrac{FP}{TN+FP} = \dfrac{FP}{N} = \alpha$ — míra falešně pozitivních *false positive rate*

$\dfrac{TP+TN}{TP+TN+FP+FN} = \dfrac{TP+TN}{P+N}$ — nesprávně přesnost *accuracy*

$\dfrac{TP+FN}{TP+TN+FP+FN} = \dfrac{P}{P+N}$ — prevalence

$\dfrac{TP}{TP+FP} = \dfrac{TP}{P'}$ — přesnost, prediktivní hodnota pozitivního testu *precision, positive predictive value*

$\dfrac{TN}{TN+FN} = \dfrac{TN}{N'}$ — prediktivní hodnota negativního testu *negative predictive value*

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 obykle odpovídá případu $\vartheta = c$, takže postup je stejný.)

U složené hypotézy požadujeme, aby pravděpodobnost chyby 1. druhu byla nejvýše α 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 mezí 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 ϑ roste. Kvantily tohoto rozdělení určují intervalový odhad s koeficientem spolehlivosti $1-\alpha$. Nulovou hypotézu zamítneme, pokud realizace t statistiky T padne mimo tento interval:

H_0	H_1	zamítáme pro	dosažená významnost
$\vartheta \le c$	$\vartheta > c$	$t > q_T(1-\alpha)$	$1 - F_T(t)$
$\vartheta \ge c$	$\vartheta < c$	$t < q_T(\alpha)$	$F_T(t)$
$\vartheta = c$	$\vartheta \ne c$	$t > q_T(1-\frac{\alpha}{2})$ nebo $t < q_T(\frac{\alpha}{2})$	$2\,\min(F_T(t), 1-F_T(t))$

V literatuře se setkáme i s následujícími případy hypotéz, které se však řeší stejně:

H_0	H_1
$\vartheta = c$	$\vartheta > c$
$\vartheta = c$	$\vartheta < c$

nahradíme

H_0	H_1
$\vartheta \le c$	$\vartheta > c$
$\vartheta \ge c$	$\vartheta < c$

9.2 Testy střední hodnoty normálního rozdělení

9.2.1 Při známém rozptylu σ^2

Výběrový průměr $\overline{\boldsymbol{X}}_n$, který má rozdělení $\mathrm{N}(\mu, \frac{\sigma^2}{n})$. S pravděpodobností $1-\alpha$ je

$$q_{\mathrm{N}(\mu,\sigma^2/n)}(\frac{\alpha}{2}) \le \overline{\boldsymbol{X}}_n \le q_{\mathrm{N}(\mu,\sigma^2/n)}(1-\frac{\alpha}{2})\,,$$

po znormování

$$\Phi^{-1}(\frac{\alpha}{2}) = -\Phi^{-1}(1-\frac{\alpha}{2}) \le \operatorname{norm}\overline{\boldsymbol{X}}_n = \underbrace{\frac{\overline{\boldsymbol{X}}_n - \mu}{\sigma}\sqrt{n}}_{T} \le \Phi^{-1}(1-\frac{\alpha}{2})\,,$$

V nulové hypotéze daná hodnota c nahrazuje neznámou střední hodnotu μ.
Realizaci t testové statistiky T,

$$t = \frac{\bar{\boldsymbol{x}} - c}{\sigma}\sqrt{n}\,,$$

porovnáváme s kvantily **normovaného normálního rozdělení**:

H_0	zamítáme pro	dosažená významnost
$\mu = c$	$\lvert t\rvert > \Phi^{-1}(1-\frac{\alpha}{2})$	$2\,(1-\Phi(\lvert t\rvert))$
$\mu \le c$	$t > \Phi^{-1}(1-\alpha)$	$1-\Phi(t)$
$\mu \ge c$	$t < -\Phi^{-1}(1-\alpha)$	$\Phi(t)$

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

9.2.2 Při neznámém rozptylu

$$t = \frac{\bar{\boldsymbol{x}} - c}{s_{\boldsymbol{x}}} \sqrt{n}$$

porovnáváme s kvantily **Studentova rozdělení** s $n-1$ stupni volnosti:

H_0	zamítáme pro	dosažená významnost
$\mu = c$	$\lvert t\rvert > q_{\mathrm{t}(n-1)}(1-\frac{\alpha}{2})$	$2\,(1 - F_{\mathrm{t}(n-1)}(\lvert t\rvert))$
$\mu \le c$	$t > q_{\mathrm{t}(n-1)}(1-\alpha)$	$1 - F_{\mathrm{t}(n-1)}(t)$
$\mu \ge c$	$t < -q_{\mathrm{t}(n-1)}(1-\alpha)$	$F_{\mathrm{t}(n-1)}(t)$

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).

9.3 Testy rozptylu normálního rozdělení

S pravděpodobností $1-\alpha$ je

$$q_{\chi^2(n-1)}(\frac{\alpha}{2}) \le \underbrace{\frac{(n-1)\,S_{\boldsymbol{X}}^2}{\sigma^2}}_{T} \le q_{\chi^2(n-1)}(1-\frac{\alpha}{2})\,,$$

V nulové hypotéze daná hodnota c nahrazuje neznámý rozptyl σ^2.
Realizaci t testové statistiky T,

$$t = \frac{(n-1)\,s_{\boldsymbol{x}}^2}{c}\,,$$

porovnáváme s kvantily **χ^2-rozdělení** s $n-1$ stupni volnosti:

H_0	zamítáme pro	dosažená významnost
$\sigma^2 = c$	$t < q_{\chi^2(n-1)}(\frac{\alpha}{2})$ nebo $t > q_{\chi^2(n-1)}(1-\frac{\alpha}{2})$	$2\,\min(F_{\chi^2(n-1)}(t), 1 - F_{\chi^2(n-1)}(t))$
$\sigma^2 \le c$	$t > q_{\chi^2(n-1)}(1-\alpha)$	$1 - F_{\chi^2(n-1)}(t)$
$\sigma^2 \ge c$	$t < q_{\chi^2(n-1)}(\alpha)$	$F_{\chi^2(n-1)}(t)$

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 χ^2-rozdělení použít normální).

9.4 Porovnání dvou normálních rozdělení

Předpoklad: Nezávislé výběry

$$(X_1, \ldots, X_m) \text{ z rozdělení } \mathrm{N}(\mathrm{E}X, \mathrm{D}X),$$
$$(Y_1, \ldots, Y_n) \text{ z rozdělení } \mathrm{N}(\mathrm{E}Y, \mathrm{D}Y).$$

9.4.1 Testy rozptylu dvou normálních rozdělení [Fisher]

Je-li $\mathrm{D}X = \mathrm{D}Y$, pak $S_{\boldsymbol{X}}^2 \doteq S_{\boldsymbol{Y}}^2$. Testovací statistikou je

$$T = \frac{S_{\boldsymbol{X}}^2}{S_{\boldsymbol{Y}}^2}\,.$$

F-rozdělení (Fisherovo-Snedecorovo rozdělení) s ξ a η stupni volnosti je rozdělení náhodné veličiny

$$F = \frac{\frac{U}{\xi}}{\frac{V}{\eta}}\,,$$

kde U, V jsou **nezávislé** náhodné veličiny s rozdělením $\chi^2(\xi)$, resp. $\chi^2(\eta)$.
Značení: $\mathrm{F}(\xi, \eta)$

Hustota pro $x > 0$:

$$f_{\mathrm{F}(\xi,\eta)}(x) = c(\xi,\eta)\, x^{\frac{\xi}{2}-1}\,(1+\frac{\xi}{\eta}\,x)^{-\frac{\xi+\eta}{2}}\,,$$

$$c(\xi,\eta) = \frac{\Gamma(\frac{\xi+\eta}{2})}{\Gamma(\frac{\xi}{2})\,\Gamma(\frac{\eta}{2})}\,(\frac{\xi}{\eta})^{\frac{\xi}{2}}$$

Je-li $\mathrm{D}X = \mathrm{D}Y = \sigma^2$, pak dosadíme

$$U := \frac{(m-1)\,S_{\boldsymbol{X}}^2}{\sigma^2} \text{ má } \chi^2(m-1)\,,$$
$$V := \frac{(n-1)\,S_{\boldsymbol{Y}}^2}{\sigma^2} \text{ má } \chi^2(n-1)\,,$$
$$\xi := m-1,\ \eta := n-1\,,$$
$$F = \frac{\frac{U}{\xi}}{\frac{V}{\eta}} = \frac{\frac{(m-1)\,S_{\boldsymbol{X}}^2}{(m-1)\,\sigma^2}}{\frac{(n-1)\,S_{\boldsymbol{Y}}^2}{(n-1)\,\sigma^2}} = \frac{S_{\boldsymbol{X}}^2}{S_{\boldsymbol{Y}}^2} = T\,.$$

Testujeme realizaci

$$t = \frac{s_{\boldsymbol{x}}^2}{s_{\boldsymbol{y}}^2}$$

na rozdělení $\mathrm{F}(m-1, n-1)$:

H_0	zamítáme pro	dosažená významnost
$\mathrm{D}X \le \mathrm{D}Y$	$t > q_{\mathrm{F}(m-1,n-1)}(1-\alpha)$	$1 - F_{\mathrm{F}(m-1,n-1)}(t)$
$\mathrm{D}X \ge \mathrm{D}Y$	$t < q_{\mathrm{F}(m-1,n-1)}(\alpha)$	$F_{\mathrm{F}(m-1,n-1)}(t)$
$\mathrm{D}X = \mathrm{D}Y$	$t < q_{\mathrm{F}(m-1,n-1)}(\frac{\alpha}{2})$ nebo $t > q_{\mathrm{F}(m-1,n-1)}(1-\frac{\alpha}{2})$	$2\,\min(F_{\mathrm{F}(m-1,n-1)}(t),$ $1 - F_{\mathrm{F}(m-1,n-1)}(t))$

Pro každou hladinu významnosti potřebujeme dvoudimenzionální tabulku kvantilů indexovanou ξ, η; obvykle je tabelována jen polovina, druhou je třeba dopočítat podle vzorce

$$q_{\mathrm{F}(\xi,\eta)}(\beta) = \frac{1}{q_{\mathrm{F}(\eta,\xi)}(1-\beta)}\,.$$

(Pozor na opačné pořadí indexů!)
Lépe je uvažovat $\frac{S_{\boldsymbol{Y}}^2}{S_{\boldsymbol{X}}^2}$ místo $\frac{S_{\boldsymbol{X}}^2}{S_{\boldsymbol{Y}}^2}$, takže rozlišíme 2 případy:
1. Pro $s_{\boldsymbol{x}}^2 \ge s_{\boldsymbol{y}}^2$ testujeme

$$t = \frac{s_{\boldsymbol{x}}^2}{s_{\boldsymbol{y}}^2} \ge 1$$

na rozdělení $\mathrm{F}(m-1, n-1)$:

H_0	zamítáme pro	dosažená významnost
$\mathrm{D}X \le \mathrm{D}Y$	$t > q_{\mathrm{F}(m-1,n-1)}(1-\alpha)$	$1 - F_{\mathrm{F}(m-1,n-1)}(t)$
$\mathrm{D}X \ge \mathrm{D}Y$	nezamítáme	žádná
$\mathrm{D}X = \mathrm{D}Y$	$t > q_{\mathrm{F}(m-1,n-1)}(1-\frac{\alpha}{2})$	$2\,(1 - F_{\mathrm{F}(m-1,n-1)}(t))$

2. Pro $s_{\boldsymbol{x}}^2 \le s_{\boldsymbol{y}}^2$ testujeme

$$t' = \frac{1}{t} = \frac{s_{\boldsymbol{y}}^2}{s_{\boldsymbol{x}}^2} \ge 1$$

na rozdělení $\mathrm{F}(n-1, m-1)$ (pozor na pořadí počtů stupňů volnosti!):

H_0	zamítáme pro	dosažená významnost
$\mathrm{D}X \le \mathrm{D}Y$	nezamítáme	žádná
$\mathrm{D}X \ge \mathrm{D}Y$	$t' > q_{\mathrm{F}(n-1,m-1)}(1-\alpha)$	$1 - F_{\mathrm{F}(n-1,m-1)}(t')$
$\mathrm{D}X = \mathrm{D}Y$	$t' > q_{\mathrm{F}(n-1,m-1)}(1-\frac{\alpha}{2})$	$2\,(1 - F_{\mathrm{F}(n-1,m-1)}(t'))$

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 σ^2

$$\overline{\boldsymbol{X}}_m \text{ má } \mathrm{N}(\mathrm{E}X, \frac{\sigma^2}{m}),$$
$$\overline{\boldsymbol{Y}}_n \text{ má } \mathrm{N}(\mathrm{E}Y, \frac{\sigma^2}{n}),$$
$$\overline{\boldsymbol{X}}_m - \overline{\boldsymbol{Y}}_n \text{ má } \mathrm{N}(\mathrm{E}X - \mathrm{E}Y, \sigma^2(\frac{1}{m} + \frac{1}{n})).$$

Za předpokladu $\mathrm{E}X = \mathrm{E}Y$:

$$T := \frac{\overline{\boldsymbol{X}}_m - \overline{\boldsymbol{Y}}_n}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}} \text{ má } \mathrm{N}(0,1).$$

Testujeme realizaci t na $\mathrm{N}(0,1)$ (viz kapitola 9.2.1).

9.4.3 Testy středních hodnot dvou normálních rozdělení s různými známými rozptyly σ_X^2, σ_Y^2

$$\overline{\boldsymbol{X}}_m \text{ má } \mathrm{N}(\mathrm{E}X, \frac{\sigma_X^2}{m}),$$
$$\overline{\boldsymbol{Y}}_n \text{ má } \mathrm{N}(\mathrm{E}Y, \frac{\sigma_Y^2}{n}),$$
$$\overline{\boldsymbol{X}}_m - \overline{\boldsymbol{Y}}_n \text{ má } \mathrm{N}(\mathrm{E}X - \mathrm{E}Y, (\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n})).$$

Za předpokladu $\mathrm{E}X = \mathrm{E}Y$:

$$T := \frac{\overline{\boldsymbol{X}}_m - \overline{\boldsymbol{Y}}_n}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \text{ má } \mathrm{N}(0,1).$$

Testujeme realizaci t na $\mathrm{N}(0,1)$ (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 σ^2

Nejprve ověříme předpoklad $\mathrm{D}X = \mathrm{D}Y = \sigma^2$ (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.].)*

$$\overline{\boldsymbol{X}}_m \text{ má } \mathrm{N}(\mathrm{E}X, \frac{\sigma^2}{m}),$$
$$\overline{\boldsymbol{Y}}_n \text{ má } \mathrm{N}(\mathrm{E}Y, \frac{\sigma^2}{n}),$$
$$\overline{\boldsymbol{X}}_m - \overline{\boldsymbol{Y}}_n \text{ má } \mathrm{N}(\mathrm{E}X - \mathrm{E}Y, \sigma^2(\frac{1}{m} + \frac{1}{n})).$$

Za předpokladu $\mathrm{E}X = \mathrm{E}Y$:

$$\frac{\overline{\boldsymbol{X}}_m - \overline{\boldsymbol{Y}}_n}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}} \text{ má } \mathrm{N}(0,1).$$

Pro měření významnosti potřebujeme odhad rozptylu. Máme dva odhady $S_{\boldsymbol{X}}^2, S_{\boldsymbol{Y}}^2$ téže hodnoty σ^2; použijeme jejich vážený průměr takový, abychom znali i jeho rozdělení.

$$\frac{(m-1)\,S_{\boldsymbol{X}}^2}{\sigma^2} \text{ má } \chi^2(m-1),$$
$$\frac{(n-1)\,S_{\boldsymbol{Y}}^2}{\sigma^2} \text{ má } \chi^2(n-1),$$
$$\frac{(m-1)\,S_{\boldsymbol{X}}^2 + (n-1)\,S_{\boldsymbol{Y}}^2}{\sigma^2} \text{ má } \chi^2(m+n-2)$$

se střední hodnotou $m+n-2$,

$$\frac{(m-1)\,S_{\boldsymbol{X}}^2+(n-1)\,S_{\boldsymbol{Y}}^2}{(m+n-2)\,\sigma^2}=\frac{S^2}{\sigma^2}$$

má střední hodnotu 1 a

$$S^2:=\frac{(m-1)\,S_{\boldsymbol{X}}^2+(n-1)\,S_{\boldsymbol{Y}}^2}{m+n-2}$$

je nestranný odhad σ^2, vedoucí na (vychýlený) odhad směrodatné odchylky

$$S:=\sqrt{\frac{(m-1)\,S_{\boldsymbol{X}}^2+(n-1)\,S_{\boldsymbol{Y}}^2}{m+n-2}}\,.$$

Ten použijeme místo neznámé směrodatné odchylky σ a výsledné kritérium

$$T:=\frac{\overline{\boldsymbol{X}}_m-\overline{\boldsymbol{Y}}_n}{S\sqrt{\frac{1}{m}+\frac{1}{n}}}=\frac{\frac{\overline{\boldsymbol{X}}_m-\overline{\boldsymbol{Y}}_n}{\sigma\sqrt{\frac{1}{m}+\frac{1}{n}}}}{\sqrt{\frac{S^2}{\sigma^2}}}\ \text{má}\ \mathrm{t}(m+n-2)\,.$$

Testujeme realizaci t na rozdělení $\mathrm{t}(m+n-2)$ (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:

$$\begin{aligned}X_j&=Z_j+U_j\,,\\Y_j&=Z_j+V_j-c\,,\end{aligned}$$

kde náhodné veličiny $U_1\ldots,U_n,V_1\ldots,V_n$ jsou nezávislé, $U_1\ldots,U_n$ mají rozdělení $\mathrm{N}(0,\sigma_U^2)$, $V_1\ldots,V_n$ mají rozdělení $\mathrm{N}(0,\sigma_V^2)$ a $c\in\mathbb{R}$. Testujeme hypotézu o hodnotě c (nejčastěji testujeme předpoklad $c=0$).
Náhodné veličiny $\Delta_j:=X_j-Y_j=U_j-V_j+c\quad(j=1,\ldots,n)$ s rozdělením $\mathrm{N}(c,\sigma_\Delta^2)$ (kde $\sigma_\Delta^2:=\sigma_U^2+\sigma_V^2$) jsou nezávislé. Obecněji nám stačí:
Předpoklad: Náhodné veličiny $\Delta_j:=X_j-Y_j$ jsou nezávislé a mají rozdělení $\mathrm{N}(c,\sigma_\Delta^2)$.

Pokud rozptyl σ_Δ^2 známe, testujeme

$$T:=\frac{\overline{\boldsymbol{\Delta}}-c}{\sigma_\Delta}\sqrt{n}=\frac{\overline{\boldsymbol{X}}-\overline{\boldsymbol{Y}}-c}{\sigma_\Delta}\sqrt{n}$$

na rozdělení $\mathrm{N}(0,1)$ dle kapitoly 9.2.1.

Pokud rozptyl neznáme, testujeme

$$T:=\frac{\overline{\boldsymbol{\Delta}}-c}{S_\Delta}\sqrt{n}$$

na rozdělení $\mathrm{t}(n-1)$ 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 $((x_1, y_1), \dots, (x_n, y_n))$ můžeme korelaci

$$\varrho(X, Y) = \frac{\mathrm{E}((X - \mathrm{E}X)\,(Y - \mathrm{E}Y))}{\sigma_X\,\sigma_Y} \in \langle -1, 1\rangle$$

odhadnout pomocí korelace empirického rozdělení neboli **realizace výběrového koeficientu korelace**

$$r_{\boldsymbol{x},\boldsymbol{y}} = \varrho(\mathrm{Emp}(\boldsymbol{x}, \boldsymbol{y})) = \frac{\sum_{j=1}^{n} (x_j - \bar{\boldsymbol{x}})\,(y_j - \bar{\boldsymbol{y}})}{\sqrt{(\sum_{j=1}^{n} (x_j - \bar{\boldsymbol{x}})^2)\,(\sum_{j=1}^{n} (y_j - \bar{\boldsymbol{y}})^2)}} \in \langle -1, 1\rangle\,,$$

což je kosinus úhlu vektorů

$$(x_1 - \bar{\boldsymbol{x}}, \dots, x_n - \bar{\boldsymbol{x}}), (y_1 - \bar{\boldsymbol{y}}, \dots, y_n - \bar{\boldsymbol{y}}) \in \mathbb{R}^n\,.$$

Jednoprůchodový vzorec:

$$\begin{aligned} r_{\boldsymbol{x},\boldsymbol{y}} &= \frac{n \sum_{j=1}^{n} x_j\,y_j - \left(\sum_{j=1}^{n} x_j\right)\left(\sum_{j=1}^{n} y_j\right)}{\sqrt{\left(n \sum_{j=1}^{n} x_j^2 - \left(\sum_{j=1}^{n} x_j\right)^2\right)\left(n \sum_{j=1}^{n} y_j^2 - \left(\sum_{j=1}^{n} y_j\right)^2\right)}} = \\ &= \frac{n}{n-1}\,\frac{\frac{1}{n}\sum_{j=1}^{n} x_j\,y_j - \bar{\boldsymbol{x}}\,\bar{\boldsymbol{y}}}{s_{\boldsymbol{x}}\,s_{\boldsymbol{y}}}\,. \end{aligned} \tag{2}$$

Výběrový koeficient korelace je odpovídající odhad

$$R_{\boldsymbol{X},\boldsymbol{Y}} = \frac{\sum_{j=1}^{n} (X_j - \overline{\boldsymbol{X}})\,(Y_j - \overline{\boldsymbol{Y}})}{\sqrt{(\sum_{j=1}^{n} (X_j - \overline{\boldsymbol{X}})^2)\,(\sum_{j=1}^{n} (Y_j - \overline{\boldsymbol{Y}})^2)}}\,.$$

9.5.1 Test nekorelovanosti dvou normálních rozdělení

Předpoklad: Dvojrozměrná náhodná veličina (X, Y) má (dvojrozměrné) normální rozdělení, $n \geq 3$.
Testovací statistikou je

$$T = \frac{R_{\boldsymbol{X},\boldsymbol{Y}}\,\sqrt{n-2}}{\sqrt{1 - R^2_{\boldsymbol{X},\boldsymbol{Y}}}}\,,$$

za předpokladu nekorelovanosti má rozdělení $\mathrm{t}(n-2)$, 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 χ^2-test dobré shody

9.6.1 Základní podoba testu

Slouží k testování hypotézy, že náhodná veličina má předpokládané rozdělení. Protože umíme hypotézy jen zamítat, 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 nenulovými pravděpodobnostmi $p_1, \dots, p_k$.

Testujeme pomocí realizace náhodného výběru rozsahu n. Není důležité pořadí výsledků, pouze jejich **četnosti** N_i, resp. **realizace četností** n_i (**empirické četnosti**) nebo **realizace relativních četností** $\frac{n_i}{n}$ $(i = 1, \dots, k)$. Porovnáváme je s **teoretickými četnostmi** $n\,p_i$.
Speciální případ: Pro $k = 2$ mají N_1, N_2 binomická rozdělení $\mathrm{Bi}(n, p_1)$, $\mathrm{Bi}(n, p_2)$, která lze pro velká n přibližně nahradit normálními,

$$\mathrm{N}(n\,p_1, n\,p_1\,(1-p_1)) = \mathrm{N}(n\,p_1, n\,p_1\,p_2)\,,$$
$$\mathrm{N}(n\,p_2, n\,p_2\,(1-p_2)) = \mathrm{N}(n\,p_2, n\,p_1\,p_2)\,.$$

Kvadráty normovaných veličin

$$(\mathrm{norm}\,N_1)^2 = \frac{(N_1 - n\,p_1)^2}{n\,p_1\,p_2}\,, \qquad (\mathrm{norm}\,N_2)^2 = \frac{(N_2 - n\,p_2)^2}{n\,p_1\,p_2}$$

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

$$\begin{aligned}(\mathrm{norm}\,N_1)^2 &= p_2\,(\mathrm{norm}\,N_1)^2 + p_1\,(\mathrm{norm}\,N_1)^2\\ &= p_2\,(\mathrm{norm}\,N_1)^2 + p_1\,(\mathrm{norm}\,N_2)^2\\ &= \frac{(N_1 - n\,p_1)^2}{n\,p_1} + \frac{(N_2 - n\,p_2)^2}{n\,p_2} = \sum_{i=1}^{2} \frac{(N_i - n\,p_i)^2}{n\,p_i}\,.\end{aligned}$$

Obecně pro libovolné k je testovací statistikou

$$T := \sum_{i=1}^{k} \frac{(N_i - n\,p_i)^2}{n\,p_i}\,,$$

jejíž rozdělení se pro $n \to \infty$ blíží $\chi^2(k-1)$. Její realizace

$$t := \sum_{i=1}^{k} \frac{(n_i - n\,p_i)^2}{n\,p_i}\,.$$

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.*

relativní preference	0.376	0.344	0.136	0.077	0.067
počet poslanců	81	74	26	13	6

Řešení. *Doplníme tabulku (poslední sloupec uvádí celkový údaj):*

relativní preference	0.376	0.344	0.136	0.077	0.067	1
počet poslanců	81	74	26	13	6	200
teor. četnost	75.2	68.8	27.2	15.4	13.4	200
příspěvek k χ^2	0.447	0.393	0.052	0.374	4.086	*5.353*

Hodnotu kritéria 5.353 *porovnáme s kvantilem* $q_{\chi^2(4)}(0.95) \doteq 9.4877$ *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 skutečné jen limitně blíží. Tím se dopouštíme blíže neurčené dodatečné chyby. Teoretické četnosti tříd nesmí být příliš malé (aspoň 5), aby náš 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 $\frac{m^2}{m} = m\times$. 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 napřed diskretizujeme, tj. všechny možné výsledky rozdělíme do k disjunktních tříd. Prvky v jedné třídě si mají být „blízké“, jinak snižujeme sílu testu. Všechny teoretické četnosti musí být dostatečně velké a nejlépe zhruba stejné.

Poznámka: Zásadně musíme pracovat s jednotkami (objekty), z nichž každá zvlášť (a nezávisle) je zařazena do nějaké třídy. Nelze počítat s tisíci, procenty, spojitým množstvím atd.

9.6.3 χ^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 $p_1, \dots, p_k$,
Y nabývá m hodnot s pravděpodobnostmi $q_1, \dots, q_m$.
Realizace dvojrozměrného náhodného výběru $((x_1, y_1), \dots, (x_n, y_n))$ obsahuje dvojice realizací náhodných veličin X, Y; potřebujeme pouze četnosti N_{ij}, resp. jejich realizace n_{ij} $(i = 1, \dots, k;\ j = 1, \dots, m)$. Ty bývají uspořádány do tzv. **kontingenční tabulky**. Počet tříd je $k\,m$.

Za předpokladu nezávislosti jsou pravděpodobnosti výsledků $p_i\,q_j$ $(i = 1, \dots, k;\ j = 1, \dots, m)$,

$$T := \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(N_{ij} - n\,p_i\,q_j)^2}{n\,p_i\,q_j} \text{ se blíží } \chi^2(k\,m-1)\,.$$

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 neboli parametry empirického rozdělení,

$$p_i = \frac{1}{n} \sum_{j=1}^{m} n_{ij}\,, \qquad q_j = \frac{1}{n} \sum_{i=1}^{k} n_{ij}\,.$$

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

$$k\,m - 1 - (k-1) - (m-1) = (k-1)\,(m-1)$$

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 χ^2-test dobré shody dvou rozdělení

H_0 : Dva náhodné výběry pocházejí ze stejného diskrétního rozdělení.
Rozsahy výběrů jsou m, n, četnosti výsledků m_i, n_i $(i = 1, \ldots, k)$.

Sjednocení obou výběrů považujeme za $m + n$ realizací 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

$$(4, 5, 6, 4)\,,$$
$$(6, 6, 6, 5, 5, 4)$$

vytvoříme dvojrozměrný výběr

$$\big((4,1),(5,1),(6,1),(4,1),(6,2),(6,2),(6,2),(5,2),(5,2),(4,2)\big)\,.$$

Za předpokladu platnosti nulové hypotézy je X nezávislé 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 znaménko odchylky od zvolené hodnoty c. Tím ztrácíme kvantativní 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 $\mathrm{Bin}(n, \frac{1}{2})$. Nulovou hypotézu zamítáme pro

$$t < q_{\mathrm{Bin}(n,\frac{1}{2})}(\frac{\alpha}{2}) \text{ nebo } t > q_{\mathrm{Bin}(n,\frac{1}{2})}(1-\frac{\alpha}{2})\,.$$

(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 $\mathrm{N}(0, 1)$.
Lze použít i k porovnání dvou mediánů u párového testu.
Příklad použití: Odhad smrtelné 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-tému 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

$$T_2 := \min\Big(\sum_{j:z_j>0} r_j, \sum_{j:z_j<0} r_j\Big),$$

porovnáme s tabulkou kritických hodnot pro tento test.

Literatura

[Navara: PMS] Navara, M.: *Pravděpodobnost a matematická statistika.* Skriptum ČVUT, Praha, 2007.

[Rogalewicz] Rogalewicz, V.: *Pravděpodobnost a statistika pro inženýry.* 2. přepracované vydání, Skriptum FBMI ČVUT, Praha, 2007.

[Zvára, Štěpán] Zvára, K., Štěpán, J.: *Pravděpodobnost a matematická statistika* (2. vydání). Matfyzpress, MFF UK, Praha, 2002.

[Kalina, Bacigál, Schiesslová] Kalina, M., Bacigál, T., Schiesslová, A.: Základy pravdepodobnosti a matematickej štatistiky. STU Bratislava, 2010.

[Kalina, Minarechová] Kalina, M., Minarechová, Z.: Applied Mathematics For Civil Engineers. STU Bratislava, 2015.

[Anděl: Statistické metody] Anděl, J.: *Statistické metody.* 2. vyd., Matfyzpress, Praha, 1998.

[Anděl: Matematická statistika] Anděl, J.: *Matematická statistika.* SNTL/Alfa, Praha, 1978.

[Disman] Disman, M.: *Jak se vyrábí sociologická znalost.* Karolinum, UK, Praha, 2005.

[Jaroš a kol.] Jaroš, F. a kol.: *Pravděpodobnost a statistika.* Skriptum VŠCHT, 2. vydání, Praha, 1998.

[Likeš, Machek] Likeš, J., Machek, J.: *Matematická statistika.* 2. vydání, SNTL, Praha, 1988.

[Nagy] Nagy, I.: *Pravděpodobnost a matematická statistika.* Cvičení. Skriptum FD ČVUT, Praha, 2002.

[Něničková] Něničková, A.: *Matematická statistika — cvičení.* Skriptum ČVUT, Praha, 1990.

[Riečanová a kol.] Riečanová, Z. a kol.: *Numerické metódy a matematická štatistika.* Alfa/SNTL, Bratislava, 1987.

[Riečan a kol.] Riečan, B., Lamoš, F., Lenárt, C.: *Pravdepodobnosť a matematická štatistika.* Alfa/SNTL, Bratislava, 1984.

[SH10] Schlesinger, M.I., Hlaváč, V.: *Deset přednášek z teorie statistického a strukturního rozpoznávání.* ČVUT, Praha, 1999.

[Swoboda] Swoboda, H.: *Moderní statistika.* Svoboda, Praha, 1977.

[Chatfield] Chatfield, C.: *Statistics for Technology.* 3rd ed., Chapman & Hall, London, 1992.

[Gonick & Smith] Gonick, L., Smith, W.: The Cartoon Guide to Statistics. HarperPerennial 1993, HarperCollins Publishers, New York, 2005.

[Hsu] Hsu, H.P.: *Probability, Random Variables, and Random Processes.* McGraw-Hill, 1996.

[Mood a kol.] Mood, A.M., Graybill, F.A., Boes, D.C.: *Introduction to the Theory of Statistics.* 3rd ed., McGraw-Hill, 1974.

[Papoulis] Papoulis, A.: *Probability and Statistics.* Prentice-Hall, 1990.

[Papoulis, Pillai] Papoulis, A., Pillai, S.U.: *Probability, Random Variables, and Stochastic Processes.* 4th ed., McGraw-Hill, Boston, USA, 2002.

[Spiegel et al. 2000] Spiegel, M.R., Schiller, J.J., Srinivasan, R.A.: *Probability and Statistics.* McGraw-Hill, 2000.

[Wasserman] Wasserman, L.: *All of Statistics. A Concise Course in Statistical Inference.* Springer, 2004.