# 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/MVT
http://cmp.felk.cvut.cz/˜navara/psi

7. února 2013

# Obsah

# 1 O čem to je?

1A. Jak vysoká by měla být pojistka auta proti krádeži (bez marže), je-li jeho cena 1 000 000 Kč a riziko ukradení během pojistného období 0.001?
$1\,000\,000 \cdot 0.001 = 1\,000$ Kč
1B. Jak vysoká by měla být pojistka auta pro případ havárie, při níž může být škoda různě velká?
⇒ **TEORIE PRAVDĚPODOBNOSTI**

2. Jak odhadnout pravděpodobnost krádeže auta nebo střední škodu při havárii a jak přesný bude odhad?
⇒ **STATISTIKA**
3. Jak označovat auta a jejich díly, abychom je jednoznačně určili?
⇒ **TEORIE INFORMACE A KÓDOVÁNÍ**
Určitě ne jako v rozvrhu na FEL: Cvičení AD0B01PSI bude v učebně KN:E-24.

## 1.1 Teorie pravděpodobnosti

je nástroj pro účelné rozhodování v systémech, kde **budoucí** pravdivost jevů závisí na okolnostech, které zcela neznáme.
Poskytuje model takových systémů a kvantifikaci výsledků.
Pravděpodobnostní **popis** ⇒ **chování** systému

## 1.2 Statistika

je nástroj pro hledání a ověřování pravděpodobnostního popisu reálných systémů na základě jejich pozorování.
**Chování** systému ⇒ pravděpodobnostní **popis**
Poskytuje daleko víc: nástroj pro zkoumání světa, pro hledání a ověřování závislostí, které nejsou zjevné.

# 2 Základní pojmy teorie pravděpodobnosti

## 2.1 Laplaceova (klasická) definice pravděpodobnosti

**Předpoklad:** Náhodný pokus s $n \in \mathbb{N}$ různými, vzájemně se vylučujícími výsledky, které jsou **stejně možné**.
Pravděpodobnost jevu, který nastává právě při $k$ z těchto výsledků, je $k/n$.
**1. problém:** „stejně možné“=„stejně pravděpodobné,“ ale co to znamená? (definice kruhem!)
**Elementární jevy** jsou všechny „stejně možné“ výsledky.
Množina všech elementárních jevů: $\Omega$
**Jev**: $A \subseteq \Omega$

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

### 2.1.1 Základní pojmy

**Jev jistý**: $\Omega$, **1**
**Jev nemožný**: $\emptyset$, **0**
**Konjunkce jevů** („**and**“): $A \cap B$
**Disjunkce jevů** („**or**“): $A \cup B$
**Jev opačný** k $A$: $\bar{A} = \Omega \setminus A$
$A \Rightarrow B$: $A \subseteq B$
**Jevy neslučitelné** (=**vzájemně se vylučující**): $A_1, \ldots, A_n : \bigcap_{i \leq n} A_i = \emptyset$
**Jevy po dvou neslučitelné**: $A_1, \ldots, A_n : \forall i, j \in \{1, \ldots, n\},\ i \neq j : A_i \cap A_j = \emptyset$
**Jevové pole**: všechny jevy pozorovatelné v náhodném pokusu, zde $\exp \Omega$ (=množina všech podmnožin množiny $\Omega$)

### 2.1.2 Pravděpodobnost

jevu $A$:

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

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

### 2.1.3 Náhodná veličina

je libovolná funkce $X : \Omega \to \mathbb{R}$
**Střední hodnota**:

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

kde $n = |\Omega|$.
**Příklad:** Elementární jevy jsou možné výsledky hry, náhodná veličina je výše výhry. Střední hodnota je spravedlivá cena za účast ve hře.

## 2.2 Vlastnosti pravděpodobnosti

$P(A) \in \langle 0, 1 \rangle$
$P(\mathbf{0}) = 0, \qquad P(\mathbf{1}) = 1$
$P(\bar{A}) = 1 - P(A)$
$A \subseteq B \Rightarrow P(A) \leq P(B)$
$A \subseteq B \Rightarrow P(B \setminus A) = P(B) - P(A)$
$A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$ (**aditivita**)
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

### 2.2.1 Úplný systém jevů

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

Speciální případ pro 2 jevy: $\{C, \bar{C}\}$
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:

$$P(A) = P(A \cap C) + P\left(A \cap \overline{C}\right) \ .$$

## 2.3 Problémy Laplaceovy definice pravděpodobnosti

**2. problém:** Nedovoluje nekonečné množiny jevů, geometrickou pravděpodobnost...
Nelze mít nekonečně mnoho stejně pravděpodobných výsledků.
**Příklad:** Podíl plochy pevniny k povrchu Země je pravděpodobnost, že náhodně vybraný bod na Zemi leží na pevnině (je-li výběr bodů prováděn „rovnoměrně“).
**Příklad (Buffonova úloha):** Na linkovaný papír hodíme jehlu, jejíž délka je rovna vzdálenosti mezi linkami. Jaká je pravděpodobnost, že jehla protne nějakou linku?
**3. problém:** Nedovoluje iracionální hodnoty pravděpodobnosti.

### 2.3.1 Rozšíření Laplaceova modelu pravděpodobnosti

**Příklad:** Místo hrací kostky házíme krabičkou od zápalek, jejíž strany jsou nestejně dlouhé. Jaká je pravděpodobnost možných výsledků?
Připustíme, že **elementární jevy nemusí být stejně pravděpodobné**.
Ztrácíme návod, jak pravděpodobnost stanovit. 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.
Tato nevýhoda je neodstranitelná a je důvodem pro vznik statistiky, která k danému opakovatelnému pokusu hledá pravděpodobnostní model.

## 2.4 Kombinatorické pojmy a vzorce

*(Dle [Zvára, Štěpán].)*
V urně je $n$ rozlišitelných objektů, postupně vytáhneme $k$.

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

Z této tabulky pouze **kombinace s opakováním** nejsou všechny stejně pravděpodobné (odpovídají různému počtu variací s opakováním) a nedovolují proto použití Laplaceova modelu pravděpodobnosti.
**Permutace (pořadí) bez opakování**: Tvoříme posloupnost z $n$ hodnot, přičemž každá se vyskytne právě jednou. Počet permutací je $n!$ (je to speciální případ variací bez opakování pro $n = k$).
**Permutace s opakováním**: Tvoříme posloupnost délky $k$ z $n$ hodnot, přičemž $j$-tá hodnota se opakuje $k_j$-krát, $\sum_{j=1}^{n} k_j = k$. Počet **různých** posloupností je

$$\frac{k!}{k_1! \cdot \ldots \cdot k_n!} \ .$$

Speciálně pro $n = 2$ dostáváme

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

což je počet kombinací bez opakování (ovšem $k_1$-prvkových z $k$ prvků).

| $n$ | 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 1.** *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.*

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

*Důkaz.*

$$\frac{n!}{(n-k)!\, n^k} = \frac{n\,(n-1)\cdots(n-(k-1))}{n^k} = $$
$$= 1\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{k-1}{n}\right) \to 1,$$
$$\frac{\binom{n}{k}}{\binom{n+k-1}{k}} = \frac{n\,(n-1)\cdots(n-(k-1))}{(n+(k-1))\cdots(n+1)\,n} =$$
$$= \frac{1\,\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)}{\left(1+\frac{k-1}{n}\right)\cdots\left(1+\frac{1}{n}\right)1} \to 1$$

(počet činitelů $k$ je konstantní). □

**Důsledek 1.** *Pro $n \gg k$ je počet variací (resp. kombinací) s opakováním přibližně*

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

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

## 2.5 Kolmogorovova definice pravděpodobnosti

Elementárních jevů (=prvků množiny $\Omega$) může být **nekonečně mnoho, nemusí být stejně pravděpodobné**.
**Jevy** jsou podmnožiny množiny $\Omega$, ale **ne nutně všechny**; tvoří podmnožinu $\mathcal{A} \subseteq \exp \Omega$, která splňuje následující podmínky:

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

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

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

Systém $\mathcal{A}$ podmnožin nějaké množiny $\Omega$, který splňuje podmínky (A1-3), se nazývá $\sigma$-**algebra**.
Důsledky: $\Omega = \overline{\emptyset} \in \mathcal{A}$,

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

Přirozený nápad $\mathcal{A} = \exp \Omega$ vede k nežádoucím paradoxům.
(A3) je uzavřenost na **spočetná** sjednocení.
Uzavřenost na **jakákoli** sjednocení se ukazuje jako příliš silný požadavek.
Uzavřenost na **konečná** sjednocení se ukazuje jako příliš slabý požadavek; nedovoluje např. vyjádřit kruh jako sjednocení obdélníků.
$\mathcal{A}$ nemusí ani obsahovat všechny jednobodové množiny, v tom případě **elementární jevy nemusí být jevy**!

### 2.5.1 Borelova $\sigma$-algebra

je nejmenší $\sigma$-algebra podmnožin $\mathbb{R}$, která obsahuje všechny intervaly.
Obsahuje všechny intervaly otevřené, uzavřené i polouzavřené, i jejich spočetná sjednocení, a některé další množiny, ale je menší než $\exp \mathbb{R}$. Její prvky nazýváme **borelovské množiny**.

### 2.5.2 Pravděpodobnost (=pravděpodobnostní míra)

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

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

(P2) $P\left(\bigcup\limits_{n \in \mathbb{N}} A_n\right) = \sum\limits_{n \in \mathbb{N}} P(A_n)$, pokud jsou množiny (=jevy) $A_n$, $n \in \mathbb{N}$, po dvou neslučitelné.
(**spočetná aditivita**)

**Pravděpodobnostní prostor** je trojice $(\Omega, \mathcal{A}, P)$, kde $\Omega$ je neprázdná množina, $\mathcal{A}$ je $\sigma$-algebra podmnožin množiny $\Omega$ a $P\colon \mathcal{A} \to \langle 0, 1 \rangle$ je pravděpodobnost.
Dříve uvedené vlastnosti pravděpodobnosti jsou důsledkem (P1), (P2).
(**Konečná**) **aditivita** by byla příliš slabá, nedovoluje např. přechod od obsahu obdélníka k obsahu kruhu.
**Příklad („nekonečná ruleta“):** Výsledkem může být libovolné přirozené číslo, každé má pravděpodobnost 0.

**Úplná aditivita** (pro jakékoli soubory po dvou neslučitelných jevů) by byla příliš silným požadavkem. Pak bychom nepřipouštěli ani rovnoměrné rozdělení na intervalu nebo na ploše. 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 |

# 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 jako $P(A \cap B) = 0$ neznamená, že jevy $A, B$ jsou neslučitelné.)
Pro nezávislé jevy $A, B$

$$P(A \cup B) = P(A) + P(B) - P(A) \cdot P(B) \,.$$

**Důkaz**:

$$P(A \cup B) = P(A) + P(B) - P(A) \cap P(B) = P(A) + P(B) - P(A) \cdot P(B) \,.$$

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

$$P(A \cap \bar{B}) = P(A) - P(A \cap B) = P(A) - P(A) \cdot P(B) =$$
$$= P(A) \cdot (1 - P(B)) = P(A) \cdot P(\bar{B}) \,.$$

Jevy $A_1, \ldots, A_n$ se nazývají **po dvou nezávislé**, jestliže každé dva z nich jsou nezávislé.
To je málo.
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 $\mathcal{K} \subseteq \mathcal{M}$.

## 3.2 Podmíněná pravděpodobnost

**Příklad:** Fotbalová družstva mohla mít před zápasem rovné šance na vítězství. Je-li však stav zápasu 5 minut před koncem $3:0$, pravděpodobnosti výhry jsou jiné.
Máme pravděpodobnostní popis systému. Dostaneme-li dodatečnou informaci, že nastal jev $B$, aktualizujeme naši znalost o pravděpodobnosti jevu $A$ na

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

což je **podmíněná pravděpodobnost** jevu $A$ za podmínky $B$. Je definována pouze pro $P(B) \neq 0$. (To předpokládáme i nadále.)
V novém modelu je $P(\bar{B}|B) = 0$, což odráží naši znalost, že jev $\bar{B}$ nenastal.
**Podmíněná pravděpodobnost** je chápána též jako funkce

$$P(.|B) \colon \mathcal{A} \to \langle 0, 1 \rangle\,, \qquad A \mapsto \frac{P(A \cap B)}{P(B)}$$

a je to pravděpodobnost v původním smyslu.
Vlastnosti podmíněné pravděpodobnosti:

- $P(\mathbf{1}|B) = 1, \qquad P(\mathbf{0}|B) = 0.$
- Jsou-li jevy $A_1, A_2, \ldots$ jsou po dvou neslučitelné, pak
$$P\Big(\bigcup_{n\in\mathbb{N}} A_n \Big| B\Big) = \sum_{n\in\mathbb{N}} P(A_n|B)\,.$$
- Je-li $P(A|B)$ definována, jsou jevy $A, B$ **nezávislé**, právě když $P(A|B) = P(A)$.
- $B \subseteq A \Rightarrow P(A|B) = 1, \qquad P(A \cap B) = 0 \Rightarrow P(A|B) = 0.$

**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:**

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

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

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

**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.
**Problém:** Ke stanovení **aposteriorní pravděpodobnosti** $P(B_i|A)$ potřebujeme znát i **apriorní pravděpodobnost** $P(B_i)$.
**Příklad:** Na vstupu informačního kanálu mohou být znaky $1, \dots, m$, výskyt znaku $j$ označujeme jako jev $B_j$. Na výstupu mohou být znaky $1, \dots, k$, výskyt znaku $i$ označujeme jako jev $A_i$. (Obykle $k = m$, ale není to nutné.) Obvykle lze odhadnout podmíněné pravděpoděpodobnosti $P(A_i|B_j)$, že znak $j$ bude přijat jako $i$. Pokud známe apriorní pravděpodobnosti (vyslání znaku $j$) $P(B_j)$, můžeme pravděpodobnosti příjmu znaků vypočítat maticovým násobením:

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

Všechny matice v tomto vzorci mají jednotkové součty řádků (takové matice nazýváme **stochastické**). Podmíněné rozdělení pravděpodobnosti, pokud byl přijat znak $i$, je

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

Rozdělení pravděpodobností vyslaných znaků je

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

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

### 3.2.1 Podmíněná nezávislost

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

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

Podobně definujeme podmíněnou nezávislost více jevů.

# 4 Náhodné veličiny a vektory

**Příklad:** Auto v ceně 10 000 EUR bude do roka ukradeno s pravděpodobností 1 : 1 000. Adekvátní cena ročního pojistného (bez zisku pojišťovny) je $10\,000/1\,000 = 10$ EUR.

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.

$\Rightarrow$ 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

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

definovanými pro libovolný interval $I$ (a tedy i pro libovolné sjednocení spočetně mnoha intervalů a pro libovolnou borelovskou množinu).
$P_X$ je **pravděpodobnostní míra** na Borelově $\sigma$-algebře určující **rozdělení náhodné veličiny** $X$.
K tomu, aby stačila znalost $P_X$ na intervalech, se potřebujeme omezit na tzv. *perfektní míry*; s jinými se v praxi nesetkáme.
Pravděpodobnostní míra $P_X$ splňuje podmínky:
$P_X(\mathbb{R}) = 1$,
$P_X\left(\bigcup_{n\in\mathbb{N}} I_n\right) = \sum_{n\in\mathbb{N}} P_X(I_n)$, pokud jsou množiny $I_n,\ n \in \mathbb{N}$, navzájem disjunktní.
Z toho vyplývá:
$P_X(\emptyset) = 0, \qquad P_X(\mathbb{R} \setminus I) = 1 - P_X(I)$,
jestliže $I \subseteq J$, pak $P_X(I) \le P_X(J)$ a $P_X(J \setminus I) = P_X(J) - P_X(I)$.
**Úspornější reprezentace:** omezíme se na intervaly tvaru $I = (-\infty, t\rangle$, $t \in \mathbb{R}$,

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

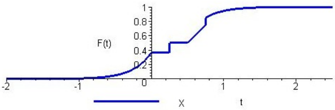
$F_X\colon \mathbb{R} \to \langle 0, 1\rangle$ je **distribuční funkce** náhodné veličiny $X$. Ta stačí, neboť

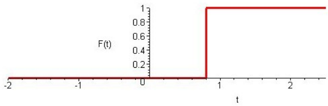
$$\begin{array}{ll}
(a,b\rangle = (-\infty,b\rangle \setminus (-\infty,a\rangle\,, & P_X((a,b\rangle) = P[a < X \le b] = F_X(b) - F_X(a)\,, \\
(a,\infty) = \mathbb{R} \setminus (-\infty,a\rangle\,, & P_X\left((a,\infty)\right) = 1 - F_X(a)\,, \\
(-\infty,a) = \bigcup\limits_{b\colon b<a} (-\infty,b\rangle\,, & P_X\left((-\infty,a)\right) = P\left[X < a\right] = \lim\limits_{b\to a-} F_X(b) = F_X(a-)\,, \\
\{a\} = (-\infty,a\rangle \setminus (-\infty,a)\,, & P_X(\{a\}) = P[X = a] = F_X(a) - F_X(a-)\,, \\
\ldots & \ldots
\end{array}$$

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 $r$:

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

($F_r$ je posunutá Heavisideova funkce.)
**Tvrzení:** $X \leq Y \Rightarrow F_X \geq F_Y$ .

## 4.2 $n$-rozměrný 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 **distribuční funkce** náhodného vektoru $\boldsymbol{X}$. Je

- neklesající (ve všech proměnných),
- zprava spojitá (ve všech proměnných),

- $\lim_{t_1 \to \infty, \ldots, t_n \to \infty} F_{\boldsymbol{X}}(t_1, \ldots, t_n) = 1$,
- $\forall k \in \{1, \ldots, n\}\ \forall t_1, \ldots, t_{k-1}, t_{k+1}, \ldots, t_n : \lim_{t_k \to -\infty} F_{\boldsymbol{X}}(t_1, \ldots, t_n) = 0$.

**Věta:** Tyto podmínky jsou **nutné, nikoli postačující**.
Nestačí znát **marginální** rozdělení náhodných veličin $X_1, \ldots, X_n$, neboť ta neobsahují informace o závislosti.

## 4.3 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.

$$P\left[X_1 \in I_1, X_2 \in I_2\right] = P\left[X_1 \in I_1\right] \cdot P\left[X_2 \in I_2\right] .$$

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

$$P\left[X_1 \leq t_1, X_2 \leq t_2\right] = P\left[X_1 \leq t_1\right] \cdot P\left[X_2 \leq t_2\right] ,$$

neboli

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

pro všechna $t_1, t_2 \in \mathbb{R}$.
Náhodné veličiny $X_1, \ldots, X_n$ jsou **nezávislé**, pokud pro libovolné intevaly $I_1, \ldots, I_n$ platí

$$P\left[X_1 \in I_1, \ldots, X_n \in I_n\right] = \prod_{i=1}^{n} P\left[X_i \in I_i\right] .$$

*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 $I_i = \mathbb{R}$. Pak $P\left[X_i \in I_i\right] = 1$ a v součinu se tento činitel neprojeví.*
Ekvivalentně stačí požadovat

$$P\left[X_1 \leq t_1, \ldots, X_n \leq t_n\right] = \prod_{i=1}^{n} P\left[X_i \leq t_i\right]$$

pro všechna $t_1, \ldots, t_n \in \mathbb{R}$, což pro sdruženou distribuční funkci **nezávislých** náhodných veličin znamená

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

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 $X_1, \ldots, X_n$.

## 4.4 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.)

## 4.5 Směs náhodných veličin

**Příklad:** Náhodné veličiny $U, V$ 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 $U$, resp. $V$ 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 $X\colon \Omega \to \mathbb{R}$:

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

$X$ je náhodná veličina na $(\Omega, \mathcal{A}, P)$.
$X$ nazýváme **směs náhodných veličin** $U, V$ **s koeficientem** $c$ (angl. *mixture*), značíme $\mathrm{Mix}_c(U, V)$. Má pravděpodobnostní míru

$$P_X = c\, P_U + (1 - c)\, P_V$$

a distribuční funkci

$$F_X = c\, F_U + (1 - c)\, F_V\,.$$

Podobně definujeme obecněji **směs náhodných veličin** $U_1, \ldots, U_n$ **s koeficienty** $c_1, \ldots, c_n \in \langle 0, 1\rangle$, $\sum\limits_{i=1}^{n} c_i = 1$, značíme $\mathrm{Mix}_{(c_1,\ldots,\, c_n)}(U_1, \ldots, U_n) = \mathrm{Mix}_{\boldsymbol{c}}(U_1, \ldots, U_n)$, kde $\boldsymbol{c} = (c_1, \ldots,\, c_n)$.
Má pravděpodobnostní míru $\sum\limits_{i=1}^{n} c_i\, P_{U_i}$ a distribuční funkci $\sum\limits_{i=1}^{n} c_i\, F_{U_i}$. (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 $c_n = 1 - \sum\limits_{i=1}^{n-1} c_i$, poslední koeficient někdy vynecháváme.
Speciálně pro dvě náhodné veličiny $\mathrm{Mix}_{(c,1-c)}(U, V) = \mathrm{Mix}_c(U, V)$ (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 $X = \mathrm{Mix}_{(c_1, \ldots, c_n)}(r_1, \ldots, r_n)$,

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

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.

## 4.6 Druhy náhodných veličin

1. **Diskrétní**: (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 $\sum\limits_{t \in \mathbb{R}} p_X(t) = 1$.

2. **Spojitá**: Má spojitou distribuční funkci.

3. **Smíšená**: Směs předchozích dvou případů;

   $\Omega_X \neq \emptyset$, $P_X(\mathbb{R} \setminus \Omega_X) = P\,[X \notin \Omega_X] \neq 0$.

## 4.7 Popis spojité náhodné veličiny

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 $\int\limits_{-\infty}^{\infty} f_X(u)\,\mathrm{d}u = 1$.

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.

## 4.8 Popis smíšené náhodné veličiny

Náhodnou veličinu $X$ se smíšeným rozdělením nelze popsat ani pravděpodobnostní funkcí (existuje, ale neurčuje celé rozdělení) ani hustotou (neexistuje, nevychází konečná), ale lze ji **jednoznačně** vyjádřit ve tvaru $X = \mathrm{Mix}_c(U, V)$, kde $U$ je diskrétní, $V$ je spojitá a $c \in (0,1)$:

$$
\begin{aligned}
c &= P_X(\Omega_X) = P_X(\{t \in \mathbb{R} : P_X(\{t\}) \neq 0\})\,, \\
c\,P_U(\{t\}) + (1-c)\,\underbrace{P_V(\{t\})}_{0} &= c\,P_U(\{t\}) = P_X(\{t\})\,, \\
p_U(t) = P_U(\{t\}) &= \frac{P_X(\{t\})}{c}\,, \\
\Omega_U &= \Omega_X\,, \\
c\,P_U(I) + (1-c)\,P_V(I) &= P_X(I)\,, \\
P_V(I) &= \frac{P_X(I) - c\,P_U(I)}{1-c}\,, \\
F_V(t) &= \frac{F_X(t) - c\,F_U(t)}{1-c}\,.
\end{aligned}
$$

Alternativa bez použití pravděpodobnostní míry:

$$
\begin{aligned}
p_X(t) = P[X = t] &= \lim_{u \to t+} F_X(t) - \lim_{u \to t-} F_X(t)\,, \\
c &= \sum_{t \in \mathbb{R}} p_X(t)\,, \\
c\,p_U(t) &= p_X(t)\,, \\
p_U(t) &= \frac{p_X(t)}{c}\,, \\
c\,F_U(t) + (1-c)\,F_V(t) &= F_X(t)\,, \\
F_V(t) &= \frac{F_X(t) - c\,F_U(t)}{1-c}\,.
\end{aligned}
$$

(Lze ještě pokračovat rozkladem diskrétní části na směs Diracových rozdělení.)

## 4.9 Kvantilová funkce náhodné veličiny

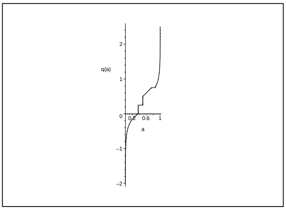
**Příklad 1.** *Pokud absolvent školy říká, že patří mezi* 5 % *nejlepších, pak tvrdí, že distribuční funkce prospěchu (náhodně vybraného absolventa) má u jeho prospěchu hodnotu nejvýše* 0.05. *(Předpokládáme, že lepšímu prospěchu odpovídá nižší průměr známek.)*
*Neostrá nerovnost v definici znamená, 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 $F_X(t) = \alpha$. Máme však zaručeno pouze, že

$$\exists t \in \mathbb{R} : P[X < t] \leq \alpha \leq P[X \leq t] = F_X(t)\,.$$

Všechna taková $t$ tvoří omezený interval a vezmeme z něj *(obvykle)* střed, přesněji tedy

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




Číslo $q_X(\alpha)$ se nazývá $\alpha$-**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,
- $q_X(\alpha) = \frac{1}{2}\,(q_X(\alpha-) + q_X(\alpha+))$.

**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).

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

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

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

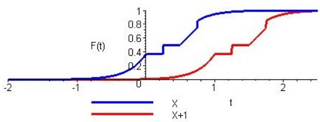
Mnohé obecnější případy se snažíme vyjádřit alespoň jako směsi náhodných veličin s rozděleními známého typu, abychom vystačili s konečně mnoha parametry.

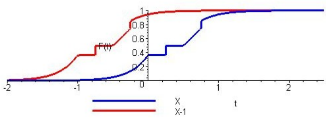
3. **Smíšená**: Jako u spojité náhodné veličiny. Tento popis je však pro diskrétní část zbytečně nepřesný.
Můžeme použít rozklad na diskrétní a spojitou část.

## 4.11 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:

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



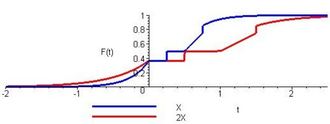



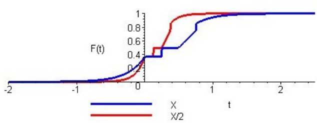

**Vynásobení nenulovou konstantou** $r$ odpovídá podobnost ve směru vodorovné osy:

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

Pro distribuční funkci musíme rozlišit případy:

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

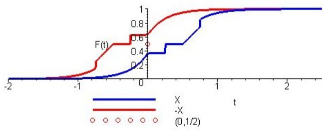



- $r = -1$: $\quad F_{-X}(-t) = P_{-X}((-\infty, -t\rangle) = P_X(\langle t, \infty)) = 1 - P_X((-\infty, t))$, **v bodech spojitosti** distribuční funkce $\quad F_{-X}(-t) = 1 - P_X((-\infty, t)) = 1 - P[X < t] = 1 - P[X \le t] = 1 - P_X((-\infty, t\rangle) = 1 - F_X(t)$,

  $F_{-X}(u) = 1 - F_X(-u)$, $\quad$ v bodech nespojitosti limita zprava (středová symetrie grafu podle bodu $\left(0, \frac{1}{2}\right)$ s opravou na spojitost zprava),

  $q_{-X}(\alpha) = -q_X(1-\alpha)$,




- $r < 0$: kombinace předchozích případů.

**Zobrazení spojitou rostoucí funkcí $h$:**
$P_{h(X)}(h(I)) = P_X(I)$, $F_{h(X)}(h(t)) = F_X(t)$, $F_{h(X)}(u) = F_X(h^{-1}(u))$,
$q_{h(X)}(\alpha) = h(q_X(\alpha))$ **v bodech spojitosti kvantilové funkce**.
**Zobrazení neklesající, zleva spojitou funkcí $h$:**
$F_{h(X)}(u) = \sup\{F_X(t) \mid h(t) \le u\}$.
**Zobrazení po částech monotonní, zleva spojitou funkcí $h$:**
Můžeme vyjádřit $h = h_+ - h_-$, kde $h_+, h_-$ jsou neklesající.
$X$ vyjádříme jako směs $X = \mathrm{Mix}_c(U, V)$, kde $U$ nabývá pouze hodnot, v nichž je $h$ neklesající, $V$ pouze hodnot, v nichž je $h$ nerostoucí. Výsledek dostaneme jako směs dvou náhodných veličin, vzniklých zobrazením funkcemi $h_+, h_-$. Funkci $h$ lze aplikovat na směs „po složkách“, tj. $h(\mathrm{Mix}_c(U, V)) = \mathrm{Mix}_c(h(U), h(V))$.
**Součet náhodných veličin** není jednoznačně určen, jedině za předpokladu **nezávislosti**. Ani pak není vztah jednoduchý.
**Směs náhodných veličin** viz výše. *Na rozdíl od součtu je plně určena (marginálními) rozděleními vstupních náhodných veličin a koeficienty směsi.*

## 4.12 Jak realizovat náhodnou veličinu na počítači

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

2. Náhodná veličina $q_Y(X)$ má stejné rozdělení jako $Y$. (Stačí tedy na každou realizaci náhodné veličiny $X$ aplikovat funkci $q_Y$.)

Všechna rozdělení **spojitých** náhodných veličin jsou stejná až na (nelineární) změnu měřítka.

## 4.13 Střední hodnota

Značení: E. nebo $\mu_.$
Je definována zvlášť pro

- **diskrétní** náhodnou veličinu $U$:

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

- **spojitou** náhodnou veličinu $V$:

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

- **směs** náhodných veličin $X = \mathrm{Mix}_c(U, V)$, kde $U$ je diskrétní, $V$ je spojitá:

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

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

Lze vyjít z definice pro diskrétní náhodnou veličinu a ostatní případy dostat jako limitu pro aproximaci jiných rozdělení diskrétním.
Všechny tři případy pokrývá univerzální vzorec s použitím kvantilové funkce

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

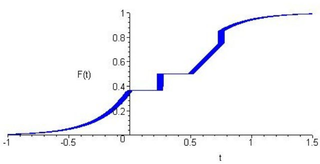
Ten lze navíc jednoduše zobecnit na střední hodnotu jakékoli funkce náhodné veličiny:

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

Speciálně pro **diskrétní** náhodnou veličinu

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

pro spojitou náhodnou veličinu by obdobný vzorec platil jen za omezujících předpokladů, protože spojitost náhodné veličiny se nemusí zachovávat.
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 (což není vždy splněno).

### 4.13.1 Vlastnosti střední hodnoty

$$\mathrm{E}r = r\,, \qquad \text{spec.} \qquad \mathrm{E}(\mathrm{E}X) = \mathrm{E}X\,,$$
$$\mathrm{E}\,(X+Y) = \mathrm{E}X + \mathrm{E}Y\,, \qquad \text{spec.} \qquad \mathrm{E}\,(X+r) = \mathrm{E}X + r\,,$$
$$\mathrm{E}\,(X-Y) = \mathrm{E}X - \mathrm{E}Y\,,$$
$$\mathrm{E}\,(r\,X) = r\,\mathrm{E}X\,, \qquad \text{obecněji} \qquad \mathrm{E}\,(r\,X + s\,Y) = r\,\mathrm{E}X + s\,\mathrm{E}Y\,.$$

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

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

*(To **není** linearita střední hodnoty.)*
Pouze pro **nezávislé** náhodné veličiny

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

## 4.14 Rozptyl (disperze)

Značení: $\sigma^2_{\cdot}$, D., var .

$$\mathrm{D}X = \mathrm{E}\left((X-\mathrm{E}X)^2\right) = \mathrm{E}\left(X^2\right) - (\mathrm{E}X)^2 \ ,$$

$$\mathrm{E}\left(X^2\right) = (\mathrm{E}X)^2 + \mathrm{D}X \,. \tag{1}$$

**Vlastnosti:**

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

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

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

Pouze pro **nezávislé** náhodné veličiny

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

## 4.15 Směrodatná odchylka

Značení: $\sigma_{\cdot}$

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

Na rozdíl od rozptylu má **stejný fyzikální rozměr** jako původní náhodná veličina.

**Vlastnosti:**

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

$$\sigma_X \geq 0\,,$$
$$\sigma_r = 0\,,$$
$$\sigma_{X+r} = \sigma_X\,,$$
$$\sigma_{r\,X} = |r|\;\sigma_X\,.$$

Pouze pro **nezávislé** náhodné veličiny

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

## 4.16 Obecné a centrální momenty

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

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

Pomocí kvantilové funkce:

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

## 4.17 Normovaná náhodná veličina

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

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

(pokud má vzorec smysl). Zpětná transformace je

$$X = \mathrm{E}X + \sigma_X \;\operatorname{norm} X\,. \tag{2}$$

## 4.18 Základní typy diskrétních rozdělení

### 4.18.1 Diracovo

Je jediný možný výsledek $r \in \mathbb{R}$.

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

Všechna diskrétní rozdělení jsou směsi Diracových rozdělení.

### 4.18.2 Rovnoměrné

Je $m$ možných výsledků stejně pravděpodobných.
Speciálně pro obor hodnot $\{1, 2, \ldots, m\}$ dostáváme

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

### 4.18.3 Alternativní (Bernoulliovo)

Jsou 2 možné výsledky. (Směs dvou Diracových rozdělení.)
Pokud výsledky jsou $0, 1$, kde $1$ má pravděpodobnost $q \in (0, 1)$, dostáváme

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

### 4.18.4 Binomické $\mathrm{Bi}(m, q)$

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

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

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

### 4.18.5 Poissonovo $\mathrm{Po}(\lambda)$

Limitní případ binomického rozdělení pro $m \to \infty$ při konstantním
$m\,q = \lambda > 0$ (tedy $q \to 0$).

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

*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).*

**Poissonovo rozdělení jako limitní případ binomického** Pro $m \to \infty$ při konstantním $m\,q = \lambda$, tj. $q = \frac{\lambda}{m}$:

$$\begin{aligned} p_X(k) &= \tbinom{m}{k}\, q^k\, (1-q)^{m-k} = \\ &= \frac{m\,(m-1)\dots(m-(k-1))}{k!} \left(\frac{\lambda}{m}\right)^k \left(1-\frac{\lambda}{m}\right)^{m-k} = \\ &= \frac{\lambda^k}{k!} \underbrace{1\left(1-\frac{1}{m}\right)\cdots\left(1-\frac{k-1}{m}\right)}_{\to 1} \underbrace{\left(1-\frac{\lambda}{m}\right)^{-k}}_{\to 1} \underbrace{\left(1-\frac{\lambda}{m}\right)^{m}}_{\to \mathrm{e}^{-\lambda}} \to \\ &\to \frac{\lambda^k}{k!}\,\mathrm{e}^{-\lambda}\,. \end{aligned}$$

### 4.18.6 Geometrické

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

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

### 4.18.7 Hypergeometrické

Počet výskytů v $m$ vzorcích, vybraných z $M$ objektů, v nichž je $K$ výskytů ($1 \le m \le K \le M$).

$$\begin{gathered} p_X(k) = \frac{\binom{K}{k}\binom{M-K}{m-k}}{\binom{M}{m}}\,, \qquad k \in \{0,1,2,\dots,m\}\,, \\ \mathrm{E}X = \frac{m\,K}{M}\,, \qquad \mathrm{D}X = \frac{m\,K\,(M-K)\,(M-m)}{M^2\,(M-1)}\,. \end{gathered}$$

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

**Binomické rozdělení jako limitní případ hypergeometrického**
**Lemma:** Pro $m, M \in \mathbb{N}$, $m < M$, je

$$\lim_{M\to\infty} \tbinom{M}{m}\,\frac{m!}{M^m} = 1\,.$$

**Důkaz:**

$$\tbinom{M}{m}\,\frac{m!}{M^m} = \frac{M\,(M-1)\cdots(M-(m-1))}{M^m} = 1\left(1-\frac{1}{M}\right)\cdots\left(1-\frac{m-1}{M}\right) \to 1\,.$$

**Důsledek:** Pro $M \gg m$ můžeme $\binom{M}{m}$ počítat přibližně jako $\frac{M^m}{m!}$.

Hypergeometrické rozdělení pro $M \to \infty$ při konstantním $\frac{K}{M} = q$, tj. $\frac{M-K}{M} = 1 - q$ (s využitím předchozího lemmatu):

$$p_X(k) = \frac{\binom{K}{k}\binom{M-K}{m-k}}{\binom{M}{m}} \to \frac{\frac{K^k}{k!} \cdot \frac{(M-K)^{m-k}}{(m-k)!}}{\frac{M^m}{m!}} =$$

$$= \frac{m!}{k!\,(m-k)!} \cdot \frac{K^k}{M^k} \cdot \frac{(M-K)^{m-k}}{M^{m-k}} = \binom{m}{k} q^k \,(1-q)^{m-k} \,.$$

## 4.19 Základní typy spojitých rozdělení

### 4.19.1 Rovnoměrné $\mathrm{R}(a, b)$

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

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

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

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

### 4.19.2 Normální (Gaussovo) $\mathrm{N}(\mu, \sigma^2)$

A. Normované $\mathrm{N}(0,1)$:

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

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

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

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

B. Obecné $\mathrm{N}(\mu, \sigma^2)$:

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

### 4.19.3 Logaritmickonormální $\mathrm{LN}(\mu, \sigma^2)$

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

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

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

$$\mathrm{E}X = \exp\left(\mu + \frac{\sigma^2}{2}\right), \qquad \mathrm{D}X = \left(\exp\left(2\,\mu + \sigma^2\right)\right)\,\left(\exp\left(\sigma^2\right) - 1\right) \,.$$

### 4.19.4 Exponenciální $\mathrm{Ex}(\tau)$

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

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

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

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

$$\mathrm{E}X = \tau\,, \qquad \mathrm{D}X = \tau^2\,, \qquad \sigma_X = \tau\,.$$

## 4.20 Náhodné vektory 2

Náhodný vektor $\boldsymbol{X} = (X_1, \ldots, X_n)$ je popsaný sdruženou distribuční funkcí $F_{\boldsymbol{X}}\colon \mathbb{R}^n \to \langle 0, 1\rangle$

$$F_{\boldsymbol{X}}(t_1, \ldots, t_n) = P[X_1 \le t_1, \ldots, X_n \le t_n]\,.$$

### 4.20.1 Diskrétní náhodný vektor

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

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

která je nenulová jen ve spočetně mnoha bodech.
**Diskrétní** náhodné veličiny $X_1, \ldots, X_n$ jsou **nezávislé**, právě když

$$P\left[X_1 = t_1, \ldots, X_n = t_n\right] = \prod_{i=1}^{n} P\left[X_i = t_i\right]$$

pro všechna $t_1, \ldots, t_n \in \mathbb{R}$. Ekvivalentní formulace:

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

### 4.20.2 Spojitý náhodný vektor

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

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

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

$$f_{\boldsymbol{X}}(u_1, \ldots, u_n) = \frac{\partial}{\partial t_1}\frac{\partial}{\partial t_2}\cdots\frac{\partial}{\partial t_n}F_{\boldsymbol{X}}(t_1, \ldots, t_n) = \mathrm{D}_1\mathrm{D}_2\ldots\mathrm{D}_n F_{\boldsymbol{X}}(t_1, \ldots, t_n)$$

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

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

**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 $t_1, \ldots, t_n \in \mathbb{R}$.

## 4.21 Číselné charakteristiky náhodného vektoru

**Střední hodnota**

- náhodného vektoru $\boldsymbol{X} = (X_1, \ldots, X_n)$: $\quad \mathrm{E}\boldsymbol{X} := (\mathrm{E}X_1, \ldots, \mathrm{E}X_n)$
- 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 $\boldsymbol{X} = (X_1, \ldots, X_n)$: $\quad \mathrm{D}\boldsymbol{X} := (\mathrm{D}X_1, \ldots, \mathrm{D}X_n)$

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.

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

kde $\mathrm{cov}(X,Y) := \mathrm{E}\,(X\,Y) - \mathrm{E}X\,\mathrm{E}Y$ je **kovariance** náhodných veličin $X, Y$. Ekvivalentně ji lze definovat

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

neboť

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

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

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

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

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

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

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

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

kde $\boldsymbol{t} = (t_1,\dots,t_n) \in \mathbb{R}^n$,
$\mu = (\mu_1,\dots,\mu_n) \in \mathbb{R}^n$,
$\boldsymbol{T} \in \mathbb{R}^{n\times n}$ je matice, BÚNO symetrická.
Parametry rozdělení:
$\mu = (\mu_1,\dots,\mu_n) \in \mathbb{R}^n$ je střední hodnota náhodného vektoru,
$\Sigma := \boldsymbol{T}^{-1}$ je kovarianční matice, speciálně její hlavní diagonála
$(\Sigma_{11},\Sigma_{22},\dots,\Sigma_{nn}) \in \mathbb{R}^n$ je rozptyl náhodného vektoru,
marginální rozdělení $i$-té složky je $\mathrm{N}(\mu_i,\Sigma_{ii})$;
pomocí těchto parametrů píšeme

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

## 4.22 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**,

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

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)

$$X \bullet X = ||X||^2 = ||X - \mathrm{E}X||^2 + ||\mathrm{E}X||^2 \,,$$
$$\mathrm{E}\left(X^2\right) = \mathrm{D}X + (\mathrm{E}X)^2 \,.$$

### 4.22.1 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 $\mathcal{N}$.

### 4.22.2 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

$$\left|\left|\sum_k c_k\, X_k - Y\right|\right|.$$

**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 $X_j$, $j = 1, \ldots, n$,

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

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

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

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

$$\sum_i c_i\; \operatorname{cov}\left(X_i, X_j\right) = \operatorname{cov}\left(Y, X_j\right) \,,$$

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

## 4.23 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).

## 4.24 Čebyševova nerovnost

**Věta:**

$$\forall \delta > 0 : P\left[|\operatorname{norm} X| < \delta\right] \geq 1 - \frac{1}{\delta^2}\,,$$

kde $\operatorname{norm} X = \frac{X - \mathrm{E}X}{\sigma_X}$ (pokud má výraz smysl).
**Důkaz pomocí kvantilové funkce:**

$$\underbrace{\mathrm{D}\left(\operatorname{norm} X\right)}_{1} = \mathrm{E}\left(\left(\operatorname{norm} X\right)^2\right) - \underbrace{\left(\mathrm{E}\left(\operatorname{norm} X\right)\right)^2}_{0},$$

$$1 = \mathrm{E}\left(\left(\operatorname{norm} X\right)^2\right) = \mathrm{E}Y\,,$$

kde $Y = \left(\operatorname{norm} X\right)^2$. Odhad pravděpodobnosti $\beta = P\left[|\operatorname{norm} X| < \delta\right] = P[Y < \delta^2] = F_Y(\delta^2-)$:

$$1 = \mathrm{E}Y = \int\limits_0^1 q_Y(\alpha)\,\mathrm{d}\alpha = \int\limits_0^\beta \underbrace{q_Y(\alpha)}_{\geq 0}\,\mathrm{d}\alpha + \int\limits_\beta^1 \underbrace{q_Y(\alpha)}_{\geq \delta^2}\,\mathrm{d}\alpha \geq (1-\beta)\,\delta^2\,,$$

$$\beta \geq 1 - \frac{1}{\delta^2}\,.$$

**Důkaz pomocí směsi:** Vyjádříme $Y = \left(\operatorname{norm} X\right)^2 = \mathrm{Mix}_\beta(L, U)$, kde
$L$ nabývá pouze hodnot z $\langle 0, \delta^2)$,
$U$ nabývá pouze hodnot z $\langle \delta^2, \infty)$, takže $\mathrm{E}U \geq \delta^2$,
$\beta = F_Y(\delta^2)$.

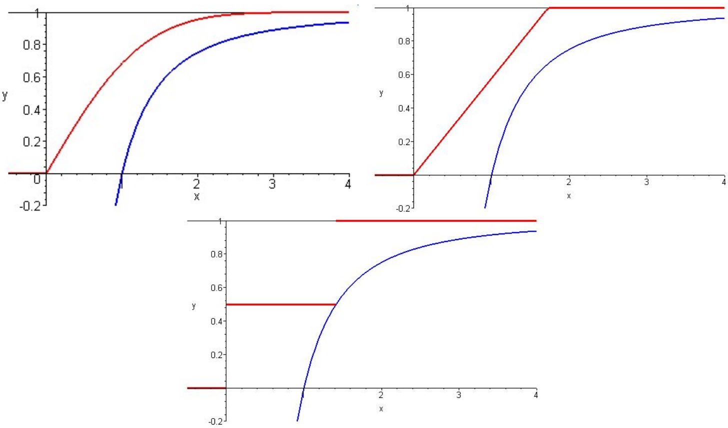
$$1 = \mathrm{E}Y = \beta\,\underbrace{\mathrm{E}L}_{\geq 0} + (1-\beta)\,\underbrace{\mathrm{E}U}_{\geq \delta^2} \geq (1-\beta)\,\delta^2\,.$$

Rovnost nastává pro $U = \delta^2$, $L = 0$, tj. pro diskrétní rozdělení
$\{(\mathrm{E}X - \delta\,\sigma_X, \frac{1-\beta}{2}), (\mathrm{E}X, \beta), (\mathrm{E}X + \delta\,\sigma_X, \frac{1-\beta}{2})\}$.

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

$$\forall \delta > 0 : P\left[\left|\frac{X - \mathrm{E}X}{\sigma_X}\right| \geq \delta\right] \leq \frac{1}{\delta^2}\,,$$

$$\forall \varepsilon > 0 : P\left[|X - \mathrm{E}X| \geq \varepsilon\right] \leq \frac{\sigma_X^2}{\varepsilon^2} = \frac{\mathrm{D}X}{\varepsilon^2}\,.$$




# 5 Základní pojmy statistiky

## 5.1 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].

## 5.2 Pojem náhodného výběru, odhady

**Soubor**

- **základní (=populace)**
- **výběrový**

Náhodný výběr jednoho prvku **základního souboru** (s rovnoměrným rozdělením) a stanovení určitého parametru tohoto prvku určuje rozdělení náhodné veličiny.
Opakovaným výběrem dostaneme náhodný vektor, jehož složky mají stejné rozdělení a jsou nezávislé.
Takto vytvoříme **výběrový soubor rozsahu** $n$, obvykle však vyloučíme vícenásobný výběr stejného prvku (*výběr bez vracení*). Jeho rozdělení se může poněkud lišit od původního. Tento rozdíl se obvykle zanedbává, neboť

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

Přesnost odhadu je dána velikostí výběrového souboru, nikoli populace.

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

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

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

$$P[G(X_1, \ldots, X_n) \leq t] = F_{G(X_1,\ldots,X_n)}(t)\,.$$

Statistika jako funkce náhodných veličin je rovněž náhodná veličina.
Obvykle se používá jako **odhad parametrů rozdělení** (které nám zůstávají skryté).
Značení:
$\vartheta$ ... skutečný parametr (reálné číslo),
$\widehat{\Theta}, \widehat{\Theta}_n$ ... jeho odhad založený na náhodném výběru rozsahu $n$ (náhodná veličina)
$\widehat{\vartheta}, \widehat{\vartheta}_n$ ... realizace odhadu (obvykle reálné číslo)
Žádoucí vlastnosti odhadů:

- $\mathrm{E}\widehat{\Theta}_n = \vartheta$ **nestranný** (opak: **vychýlený**)
- $\lim\limits_{n\to\infty} \mathrm{E}\widehat{\Theta}_n = \vartheta$ **asymptoticky nestranný**
- **eficientní** = s malým rozptylem, což posuzujeme podle
  $\mathrm{E}\left((\widehat{\Theta}_n - \vartheta)^2\right) = \mathrm{D}\widehat{\Theta}_n + \left(\mathrm{E}(\widehat{\Theta}_n - \vartheta)\right)^2$, 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)
- $\lim\limits_{n\to\infty} \mathrm{E}\widehat{\Theta}_n = \vartheta,\ \lim\limits_{n\to\infty} \sigma_{\widehat{\Theta}_n} = 0$ **konzistentní**
- **robustní**, tj. odolný vůči šumu („i při zašuměných datech dostáváme dobrý výsledek“) – zde už přesné kritérium chybí, zato je to velmi praktická vlastnost

## 5.3 Výběrový průměr

z náhodného výběru $\boldsymbol{X} = (X_1, \ldots, X_n)$ je

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

Alternativní značení: $\bar{\boldsymbol{X}}_n$ *(pokud potřebujeme zdůraznit rozsah výběru)*
Jeho realizaci značíme malým písmenem:

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

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

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

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 $\mathrm{N}\left(\mu, \frac{1}{n}\sigma^2\right)$ 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:

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

## 5.4 Výběrový rozptyl

náhodného výběru $\boldsymbol{X} = (X_1, \ldots, X_n)$ je statistika

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

Alternativní značení: $S^2$ *(Dvojka v horním indexu zde neznamená kvadrát!)*
Jeho realizaci značíme malým písmenem:

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

Praktičtější jednoprůchodový vzorec:

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

**Věta:**

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

**Důkaz:** Z jednoprůchodového vzorce pro $S_{\boldsymbol{X}}^2$ dostáváme

$$\begin{aligned} \mathrm{E} S_{\boldsymbol{X}}^2 = \frac{n}{n-1} \mathrm{E} X^2 - \frac{n}{n-1} \mathrm{E} \overline{\boldsymbol{X}}_n^2 &= \frac{n}{n-1} \left( \mathrm{D} X + (\mathrm{E} X)^2 - \mathrm{D} \overline{\boldsymbol{X}}_n - \big(\mathrm{E} \overline{\boldsymbol{X}}_n\big)^2 \right) = \\ &= \frac{n}{n-1} \left( \mathrm{D} X + (\mathrm{E} X)^2 - \frac{1}{n} \mathrm{D} X - (\mathrm{E} X)^2 \right) = \mathrm{D} X . \end{aligned}$$

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

Rozdělení výběrového rozptylu může být podstatně složitější.

Speciálně pro rozdělení $\mathrm{N}(0,1)$ a $n = 2$:

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

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

kde $U = \frac{X_1 - X_2}{\sqrt{2}}$ má rozdělení $\mathrm{N}(0,1)$. Tomu říkáme:

### 5.4.1 Rozdělení $\chi^2$ s 1 stupněm volnosti

= rozdělení náhodné veličiny $V = U^2$, kde $U$ má **normované normální** rozdělení $\mathrm{N}(0,1)$. Značení: $\chi^2(1)$. *(Toto rozdělení není zvykem normovat.)*

$$\begin{aligned} \mathrm{E} V &= \mathrm{E} U^2 = \underbrace{\mathrm{D} U}_{1} + (\underbrace{\mathrm{E} U}_{0})^2 = 1 , \\ \mathrm{D} V &= 2 . \qquad \text{(bez důkazu)} \end{aligned}$$

Pro $t > 0$ vychází distribuční funkce

$$\begin{aligned} F_V(t) &= P[V \le t] = P[-\sqrt{t} \le U \le \sqrt{t}] = 2\,P[0 \le U \le \sqrt{t}] = \\ &= 2\left( \Phi(\sqrt{t}) - \Phi(0) \right) = 2 \int_0^{\sqrt{t}} \mathrm{e}^{-\frac{u^2}{2}} \,\mathrm{d}u , \end{aligned}$$

hustota

$$f_V(t) = F_V'(t) = \left( 2\,\Phi(\sqrt{t}) \right)' = 2\,(\sqrt{t})'\,\Phi'(\sqrt{t}) = \frac{1}{\sqrt{t}}\,\varphi(\sqrt{t}) = \frac{1}{\sqrt{2\,\pi\,t}}\,\mathrm{e}^{\frac{-t}{2}} .$$

Zobecnění:

### 5.4.2 Rozdělení $\chi^2$ s $\eta$ stupni volnosti

= rozdělení náhodné veličiny $Y = \sum_{j=1}^{\eta} V_j$, kde $V_j$ jsou **nezávislé** náhodné veličiny s rozdělením $\chi^2(1)$

= rozdělení náhodné veličiny $Y = \sum_{j=1}^{\eta} U_j^2$, kde $U_j$ jsou **nezávislé** náhodné veličiny s **normovaným normálním** rozdělením $\mathrm{N}(0,1)$.

Značení: $\chi^2(\eta)$.

$$\mathrm{E}Y = \mathrm{E}\sum_{j=1}^{\eta} V_j = \sum_{j=1}^{\eta} \underbrace{\mathrm{E}V_j}_{1} = \eta\,,$$

$$\mathrm{D}Y = \mathrm{D}\sum_{j=1}^{\eta} V_j = \sum_{j=1}^{\eta} \underbrace{\mathrm{D}V_j}_{2} = 2\,\eta\,.$$

**Věta:** Nechť $X, Y$ jsou **nezávislé** náhodné veličiny s rozdělením $\chi^2(\xi)$, resp. $\chi^2(\eta)$. Pak $X+Y$ má rozdělení $\chi^2(\xi+\eta)$.

Hustota

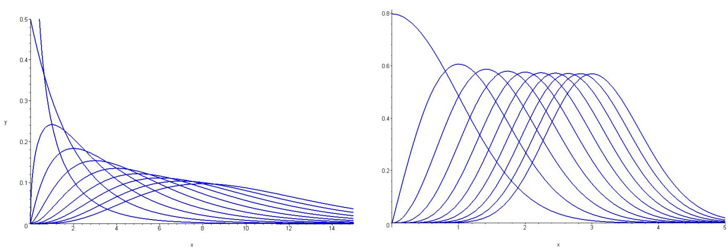
$$f_Y(y) = \begin{cases} c(\eta)\, y^{\frac{\eta}{2}-1}\, \mathrm{e}^{\frac{-y}{2}} & \text{pro } y > 0\,, \\ 0 & \text{jinak}\,, \end{cases}$$

$$c(\eta) = \frac{1}{2^{\frac{\eta}{2}}\,\Gamma\left(\frac{\eta}{2}\right)}\,,$$

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

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

Speciálně pro $\eta = 2$ je $c(\eta) = 1/2$ a dostáváme exponenciální rozdělení.

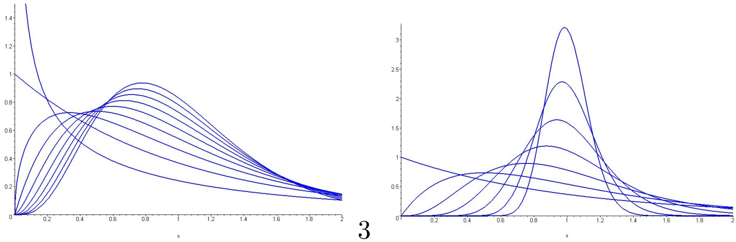


Hustoty rozdělení $\chi^2$ s $1, 2, \ldots, 10$ stupni volnosti a jeho odmocniny („vzdálenost od středu terče“).

### 5.4.3 Výběrový rozptyl

z **normálního** rozdělení $\mathrm{N}(\mathrm{E}X, \mathrm{D}X)$ splňuje:

$$\frac{(n-1)\, S_{\mathbf{X}}^2}{\mathrm{D}X} \text{ má rozdělení } \chi^2(n-1)\,.$$



3

Rozdělení odhadu rozptylu pomocí výběrového rozptylu $S^2_{\boldsymbol{X}}$ pro rozsah výběru $2, 3, \dots, 10$ a $3 = 2^1 + 1, 2^2 + 1, \dots, 2^7 + 1 = 129$.

**Důsledek:** Rozptyl výběrového rozptylu z normálního rozdělení $\mathrm{N}(\mathrm{E}X, \mathrm{D}X)$ je

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

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

Existuje však vychýlený odhad rozptylu, který je eficientnější:

### 5.4.4 Alternativní odhad rozptylu

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

**Věta:** $\widehat{\mathrm{D}X}$ je vychýlený konzistentní odhad rozptylu.
**Důkaz:**

$$\mathrm{E}\widehat{\mathrm{D}X} = \frac{n-1}{n}\,\mathrm{D}X \to \mathrm{D}X\,,$$

$\widehat{\mathrm{D}X}$ má rozptyl menší než $S^2_{\boldsymbol{X}}$, a to v poměru $\left(\frac{n-1}{n}\right)^2$.

Eficienci nemůžeme porovnat obecně; aspoň pro **normální** rozdělení:

1. eficience odhadu $S^2_{\boldsymbol{X}}$:

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

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

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

a protože

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

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

## 5.5 Výběrová směrodatná odchylka

náhodného výběru $\boldsymbol{X} = (X_1, \ldots, X_n)$ je statistika

$$S_{\boldsymbol{X}} = \sqrt{S^2_{\boldsymbol{X}}} = \sqrt{\frac{1}{n-1} \sum_{j=1}^{n} (X_j - \bar{\boldsymbol{X}}_n)^2}.$$

Alternativní značení: $S$
Její realizaci značíme malým písmenem:

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

**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 konzistentní odhad směrodatné odchylky (pokud původní rozdělení má rozptyl a 4. centrální moment).

## 5.6 Výběrový $k$-tý obecný moment

náhodného výběru $\boldsymbol{X} = (X_1, \ldots, X_n)$ je statistika

$$M_{X^k} = \frac{1}{n} \sum_{j=1}^{n} X_j^k\,.$$

Alternativní značení: $M_k$
Jeho realizaci značíme malým písmenem:

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

**Věta:**

$$\mathrm{E}M_{X^k} = \mathrm{E}X^k\,.$$

(Tj. je to **nestranný** odhad $k$-tého obecného momentu.)
**Věta:** Výběrový $k$-tý obecný moment je konzistentní odhad $k$-tého obecného momentu (pokud $X$ má $k$-tý a $2\,k$-tý obecný moment).
**Důkaz:**

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

## 5.7 Histogram a empirické rozdělení

V (nenáhodném) vektoru $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ (získaném např. jako realizace náhodného výběru) nezáleží na pořadí složek (ale záleží na jejich opakování). Úsporněji je popsán množinou hodnot $H = \{x_1, \ldots, x_n\}$ (ta má nejvýše $n$ prvků, obvykle méně) a jejich **četnostmi** $n_t$, $t \in H$. Tato data obvykle znázorňujeme **tabulkou četností** nebo grafem zvaným **histogram**.
Normováním dostaneme **relativní četnosti** $r_t = \frac{n_t}{n}$, $t \in H$. Jelikož $\sum_{t \in H} r_t = 1$, definují relativní četnosti pravděpodobnostní funkci $p_{\mathrm{Emp}(\boldsymbol{x})}(t) = r_t$ tzv. **empirického rozdělení** $\mathrm{Emp}(\boldsymbol{x})$. Je to diskrétní rozdělení s nejvýše $n$ hodnotami charakterizující vektor $\boldsymbol{x}$.

### 5.7.1 Vlastnosti empirického rozdělení

*(Indexem* $\mathrm{Emp}(\boldsymbol{x})$ *označujeme parametry jakékoli náhodné veličiny, která má toto rozdělení.)*

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

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

$$\begin{aligned} \mathrm{D}\,\mathrm{Emp}(\boldsymbol{x}) &= \sum_{t \in H} (t - \mathrm{E}\,\mathrm{Emp}(\boldsymbol{x}))^2\; r_t = \frac{1}{n} \sum_{t \in H} (t - \bar{\boldsymbol{x}})^2\; n_t \\ &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{\boldsymbol{x}})^2 = \frac{n-1}{n}\, s_{\boldsymbol{x}}^2\,. \end{aligned}$$

Obecné momenty empirického rozdělení se rovnají výběrovým momentům původního rozdělení.
Výpočet z histogramu (z empirického rozdělení) může být jednodušší než z původní realizace náhodného výběru (pokud se opakují stejné hodnoty).
Rozptyl empirického rozdělení odpovídá odhadu $\widehat{\mathrm{D}X} = \frac{n-1}{n}\, S_{\boldsymbol{X}}^2$ rozptylu původního rozdělení, odlišnému od $S_{\boldsymbol{X}}^2$.

## 5.8 Výběrový medián

je medián empirického rozdělení, $q_{\mathrm{Emp}(\boldsymbol{x})}(\frac{1}{2})$. Poskytuje jinou informaci než výběrový průměr, mnohdy užitečnější (mj. **robustnější** – odolnější vůči vlivu vychýlených hodnot, outliers). Navíc víme, jak se změní monotonní funkcí.
Proč se používá méně než výběrový průměr:

- Výpočetní náročnost je vyšší; seřazení hodnot má pracnost úměrnou $n \ln n$, zatímco výběrový průměr $n$.
- Paměťová náročnost je vyšší – potřebujeme zapamatovat všechna data, u výběrového průměru stačí 2 registry.
- Možnosti decentralizace a paralelizace výpočtu výběrového mediánu jsou velmi omezené.

## 5.9 Intervalové odhady

Dosud jsme skutečnou hodnotu parametru $\vartheta$ 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 $\alpha \in (0, \frac{1}{2})$ je pravděpodobnost, že meze intervalu $I$ budou překročeny; $1 - \alpha$ je **koeficient spolehlivosti.** Obvykle hledáme **horní**, resp. **dolní jednostranný** odhad, kdy

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

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

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

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

## 5.10 Intervalové odhady parametrů normálního rozdělení $\mathrm{N}(\mu, \sigma^2)$

### 5.10.1 Odhad střední hodnoty při známém rozptylu $\sigma^2$

$\mu$ odhadneme výběrovým průměrem $\overline{\boldsymbol{X}}$ s rozdělením $\mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$.
Normovaná náhodná veličina $\operatorname{norm} \overline{\boldsymbol{X}} = \frac{\sqrt{n}}{\sigma}(\overline{\boldsymbol{X}} - \mu)$, stejně jako $-\operatorname{norm} \overline{\boldsymbol{X}} = \frac{\sqrt{n}}{\sigma}(\mu - \overline{\boldsymbol{X}})$ má rozdělení $\mathrm{N}(0, 1)$;

$$\begin{aligned}
&P\left[\frac{\sqrt{n}}{\sigma}(\mu - \overline{\boldsymbol{X}}) \in (-\infty, \Phi^{-1}(1-\alpha)\rangle\right] \\
&= 1 - \alpha \\
&= P\left[\frac{\sqrt{n}}{\sigma}(\mu - \overline{\boldsymbol{X}}) \leq \Phi^{-1}(1-\alpha)\right] \\
&= P\left[\mu \leq \overline{\boldsymbol{X}} + \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(1-\alpha)\right] \\
&= P\left[\mu \in \left(-\infty, \overline{\boldsymbol{X}} + \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(1-\alpha)\right\rangle\right]\,.
\end{aligned}$$

Obdobně dostaneme i další intervalové odhady

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

kde $\overline{\boldsymbol{X}} - \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(1-\alpha) = \overline{\boldsymbol{X}} + \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(\alpha)$
($\Phi^{-1}(\alpha) = -\Phi^{-1}(1-\alpha)$ ovšem nebývá v tabulkách).
Při výpočtu nahradíme výběrový průměr $\overline{\boldsymbol{X}}$ jeho realizací $\bar{\boldsymbol{x}}$.

### 5.10.2 Odhad střední hodnoty při neznámém rozptylu

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

$\sigma^2$ odhadneme výběrovým rozptylem $S^2_{\boldsymbol{X}}$; $\frac{(n-1)\,S^2_{\boldsymbol{X}}}{\sigma^2}$ má rozdělení $\chi^2(n-1)$.

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

### 5.10.3 Studentovo t-rozdělení (autor: Gossett)

s $\eta$ stupni volnosti je rozdělení náhodné veličiny

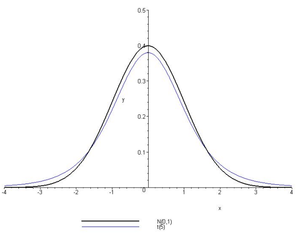
$$\frac{U}{\sqrt{\frac{V}{\eta}}}\,,$$

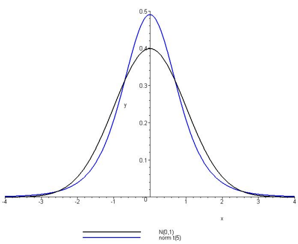
kde $U$ má rozdělení $\mathrm{N}(0,1)$,
$V$ má rozdělení $\chi^2(\eta)$,
$U, V$ jsou nezávislé.
Značení: $\mathrm{t}(\eta)$.
Hustota:

$$f_{\mathrm{t}(\eta)}\,(x) = c\,(\eta)\,\left(1+\frac{x^2}{\eta}\right)^{-\frac{1+\eta}{2}}\,,$$

$$c\,(\eta) = \frac{\Gamma\left(\frac{1+\eta}{2}\right)}{\sqrt{\eta\,\pi}\,\Gamma\left(\frac{\eta}{2}\right)}\,.$$

Symetrie kolem nuly $\Rightarrow q_{\mathrm{t}(\eta)}(1-\alpha) = -q_{\mathrm{t}(\eta)}(\alpha)$.
Pro velký počet stupňů volnosti se nahrazuje normálním rozdělením.




Hustota normovaného normálního rozdělení a Studentova rozdělení s 5 stupni volnosti (původního a normovaného).

### 5.10.4 Odhad střední hodnoty při neznámém rozptylu II

V našem případě:

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

Z toho vyplývají intervalové odhady

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

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

### 5.10.5 Odhad rozptylu

$\sigma^2$ odhadneme výběrovým rozptylem $S^2_{\boldsymbol{X}}$; $\frac{(n-1)S^2_{\boldsymbol{X}}}{\sigma^2}$ má rozdělení $\chi^2(n-1)$;

$$\begin{aligned}
&P\left[\frac{(n-1)\,S^2_{\boldsymbol{X}}}{\sigma^2} \in (-\infty, q_{\chi^2(n-1)}(1-\alpha)\rangle\right] \\
&= 1-\alpha \\
&= P\left[\frac{(n-1)\,S^2_{\boldsymbol{X}}}{\sigma^2} \le q_{\chi^2(n-1)}(1-\alpha)\right] \\
&= P\left[\frac{(n-1)\,S^2_{\boldsymbol{X}}}{q_{\chi^2(n-1)}(1-\alpha)} \le \sigma^2\right] \\
&= P\left[\sigma^2 \in \left\langle \frac{(n-1)\,S^2_{\boldsymbol{X}}}{q_{\chi^2(n-1)}(1-\alpha)}, \infty\right)\right]\,.
\end{aligned}$$

Dostali jsme **dolní** odhad.
Obdobně dostaneme i další intervalové odhady

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

Při výpočtu nahradíme výběrový rozptyl $S^2_{\boldsymbol{X}}$ jeho realizací $s^2_{\boldsymbol{x}}$.

### 5.10.6 Intervalové odhady spojitých rozdělení, která nejsou normální

převádíme obvykle na normální rozdělení nelineární 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

$$h^{-1}(u) = q_X^{-1}(\Phi(u)) \,.$$

## 5.11 Obecné odhady parametrů

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

### 5.11.1 Metoda momentů

Pro $k = 1, 2, \ldots$ je $k$-tý obecný moment funkcí $\boldsymbol{\vartheta}$,

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

(závislost na parametrech lze stanovit dle pravděpodobnostního modelu).
Lze jej též odhadnout pomocí výběrového $k$-tého obecného momentu $m_{X^k}$.
Metoda momentů doporučuje realizaci odhadu $\widehat{\boldsymbol{\vartheta}} = (\widehat{\vartheta}_1, \ldots, \widehat{\vartheta}_i)$ takovou, že

$$\mathrm{E}X^k(\widehat{\vartheta}_1, \ldots, \widehat{\vartheta}_i) = m_{X^k} = \frac{1}{n} \sum_{j=1}^{n} x_j^k, \qquad k = 1, 2, \ldots \,.$$

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

**Použitelnost metody momentů**
**Možné problémy:**

1. Řešení neexistuje $\Rightarrow$ zkusme ubrat rovnice.
2. Je nekonečně mnoho řešení $\Rightarrow$ zkusme přibrat další rovnice.
3. Je více než jedno řešení (např. soustavy kvadratických rovnic).
4. Je jediné řešení, ale je obtížné je nalézt.
5. Soustava je špatně podmíněná (typicky pro velký počet parametrů).
6. Našli jsme jediné řešení, které však **nesplňuje předpoklady**, $\widehat{\boldsymbol{\vartheta}} \notin \Pi$ (např. parametry nemohou být libovolná čísla) $\Rightarrow$ NELZE! **Vždy kontrolujte řešení!**

7. Všem rovnicím je přikládána stejná důležitost, což bývá nežádoucí (typicky pro velký počet parametrů).

8. Nelze použít pro nenumerická data (pokud je nelze smysluplně očíslovat).

**Výhoda:**

1. Lze použít pro diskrétní, spojité i **smíšené** rozdělení beze změn.

### 5.11.2 Metoda maximální věrohodnosti (likelihood)

**Pro diskrétní rozdělení** Pravděpodobnost realizace,

$$\begin{aligned} p_{\mathbf{X}}(\boldsymbol{x};\boldsymbol{\vartheta}) &= P\left[X_1 = x_1 \wedge \ldots \wedge X_n = x_n;\boldsymbol{\vartheta}\right] \\ &= \prod_{j=1}^{n} P\left[X_j = x_j;\boldsymbol{\vartheta}\right] = \prod_{j=1}^{n} p_X(x_j;\boldsymbol{\vartheta}) = L(\boldsymbol{\vartheta})\,, \end{aligned}$$

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**.
Řešením jsou takové hodnoty $\widehat{\boldsymbol{\vartheta}} = (\widehat{\vartheta}_1,\ldots,\widehat{\vartheta}_i)$, které maximalizují věrohodnost.
Maximalizujeme buď věrohodnost, nebo její logaritmus (*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.)

**Příklad:** Empirické rozdělení je maximálně věrohodný odhad diskrétního rozdělení (pokud na rozdělení nejsou kladeny další podmínky).
**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.

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

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

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. Je více než jedno řešení. (Může se stát, že různé hodnoty parametrů popisují totéž rozdělení – vadí to?)
2. Řešení neexistuje (to se může stát jedině když věrohodnostní funkce je nespojitá nebo parametrický prostor neuzavřený).
3. Je jediné řešení, ale je obtížné je nalézt. (Lokální extrémy nemusí být globální.)
4. Soustava je špatně podmíněná.
5. Hodnoty věrohodnosti mohou být velmi malé.
6. **Nelze použít pro smíšené rozdělení!**

**Výhody:**

1. Hledání optima je o něco snazší než řešení soustavy rovnic.
2. Různým datům je dán společný (srovnatelný) význam.
3. Lze použít i na nenumerická data.

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

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

$$\widehat{r} = \frac{1}{n} \sum_{j=1}^{n} x_j^2 - \widehat{\mu}^2 = \frac{1}{n} \sum_{j=1}^{n} x_j^2 - \bar{\boldsymbol{x}}^2\,,$$

což je alternativní (vychýlený konzistentní) odhad rozptylu

$$\widehat{\mathrm{D}X} = \frac{1}{n}\sum_{j=1}^{n}(x_j-\bar{\boldsymbol{x}})^2 = \frac{1}{n}\sum_{j=1}^{n}x_j^2 - \frac{2}{n}\,\bar{\boldsymbol{x}}\sum_{j=1}^{n}x_j + \frac{1}{n}\sum_{j=1}^{n}\bar{\boldsymbol{x}}^2 =$$

$$= \frac{1}{n}\sum_{j=1}^{n}x_j^2 - 2\,\bar{\boldsymbol{x}}^2 + \bar{\boldsymbol{x}}^2 = \frac{1}{n}\sum_{j=1}^{n}x_j^2 - \bar{\boldsymbol{x}}^2 = \widehat{r}\,.$$

**Metoda maximální věrohodnosti:**

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

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

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

$$\frac{\partial}{\partial r}\lambda\left(\mu,r\right) = \frac{1}{2\,r^2}\sum_{j=1}^{n}(x_j-\mu)^2 - \frac{n}{2\,r} = \frac{n}{2\,r^2}\left(\frac{1}{n}\sum_{j=1}^{n}(x_j-\mu)^2 - r\right).$$

Maximum opět nastává pro

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

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

**Odhad parametrů směsi normálních rozdělení**

**Úloha:** Z realizace náhodného výběru $(x_1,\dots,x_n)$ určete maximálně věrohodný odhad směsi normálních rozdělení se středními hodnotami $\mu_k$, $k=1,\dots,K$, stejným **známým** rozptylem $\sigma^2$ a koeficienty směsi (váhami) $c_k$, $k=1,\dots,K$.

**Pokus o řešení:**

$$f_X(t) = \sum_k c_k\,f_{\mathrm{N}(\mu_k,\sigma^2)}(t) = \sum_k c_k\,\frac{1}{\sqrt{2\,\pi}\,\sigma}\;\exp\left(\frac{-(t-\mu_k)^2}{2\,\sigma^2}\right)\,,$$

$$\Lambda(\mu,\mathbf{c}) = \prod_j f_X(x_j) = \prod_j\sum_k c_k\,f_{\mathrm{N}(\mu_k,\sigma^2)}(x_j) =$$

$$= \prod_j\left(\frac{1}{\sqrt{2\,\pi}\,\sigma}\sum_k c_k\,\exp\left(\frac{-(x_j-\mu_k)^2}{2\,\sigma^2}\right)\right) =$$

$$= \left(\frac{1}{\sqrt{2\,\pi}\,\sigma}\right)^n\prod_j\sum_k c_k\,\exp\left(\frac{-(x_j-\mu_k)^2}{2\,\sigma^2}\right)\,,$$

$$\lambda(\mu,\mathbf{c}) = \sum_j\ln\sum_k c_k\,f_{\mathrm{N}(\mu_k,\sigma^2)}(x_j) =$$

$$= -n\ln\left(\sqrt{2\,\pi}\,\sigma\right) + \sum_j\ln\sum_k c_k\,\exp\left(\frac{-(x_j-\mu_k)^2}{2\,\sigma^2}\right)\,.$$

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 $k$-té složce směsi popíšeme koeficientem $\alpha_{j,k} \in \langle 0, 1 \rangle$, přičemž

$$\sum_{k=1}^{K} \alpha_{j,k} = 1\,, \quad \sum_{j=1}^{n} \alpha_{j,k} > 0\,.$$

1. Zvolíme náhodně různé střední hodnoty složek směsi $\mu_k$ a nenulové koeficienty $c_k$, $k = 1, ..., K$, splňující $\sum_k c_k = 1$.

E. Stanovíme stupně příslušnosti

$$\alpha_{j,k} := \frac{c_k\, f_{\mathrm{N}(\mu_k,\sigma^2)}(x_j)}{\sum\limits_{k'=1}^{K} c_{k'}\, f_{\mathrm{N}(\mu_{k'},\sigma^2)}(x_j)} = \frac{c_k\, \exp\left(\frac{-(x_j-\mu_k)^2}{2\,\sigma^2}\right)}{\sum\limits_{k'=1}^{K} \left(c_{k'}\, \exp\left(\frac{-(x_j-\mu_{k'})^2}{2\,\sigma^2}\right)\right)}$$

(jmenovatel je normalizační faktor).

M. Aktualizujeme koeficienty složek směsi

$$c_k := \frac{\sum\limits_{j=1}^{n} \alpha_{j,k}}{\sum\limits_{k'=1}^{K} \sum\limits_{j=1}^{n} \alpha_{j,k'}} = \frac{1}{n} \sum_{j=1}^{n} \alpha_{j,k}$$

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

$$\mu_k := \frac{\sum\limits_{j=1}^{n} \alpha_{j,k}\, x_j}{\sum\limits_{j=1}^{n} \alpha_{j,k}} = \frac{\sum\limits_{j=1}^{n} \alpha_{j,k}\, x_j}{n\, c_k}\,.$$

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.

# 6 Testování hypotéz

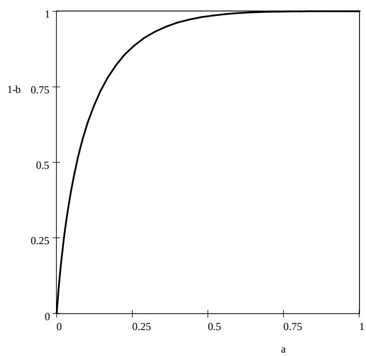
## 6.1 Základní pojmy a principy testování hypotéz

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

**Předpoklad:** Parametr $\vartheta$ nabývá pouze 2 hodnot, 0 pro „normální“ populaci, 1 pro „anomální“ prvky. O prvku máme rozhodnout, ke které skupině patří (tj. odhadnout $\vartheta$). K tomu použijeme testovací statistiku $T$ (resp. její realizaci $t$). Ta závisí na $\vartheta$. 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:

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

**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 $\kappa$).
**Chyba 2. druhu** (osvobodíme vinného): Nezamítneme nulovou hypotézu, která neplatí. Anomální je klasifikován jako normální s pravděpodobností $\beta(\kappa)$ (neklesající funkce $\kappa$).
**ROC křivka** (angl. **ROC curve**, **receiver operating characteristic**) vyjadřuje závislost pravděpodobnosti chyby prvního druhu $\alpha$ (vodorovně) a síly testu $1-\beta$ (svisle), parametrem křivky je kritická hodnota $\kappa$. Volbou kritické hodnoty se chceme co nejvíce přiblížit bodu $(0, 1)$, tj. bezchybné klasifikaci. Nicméně vybereme bod, v němž se pravděpodobnost chyby prvního druhu rovná zvolenému číslu $\alpha$ (tj. s danou vodorovnou souřadnicí).




Obrázek 1: Typický průběh **ROC** křivky

Možná kritéria pro volbu prahu $\kappa$:

- $\alpha(\kappa) = \beta(\kappa)$,
- $\min_{\kappa}(\alpha(\kappa) + \beta(\kappa))$,
- $\min_{\kappa} e(\alpha(\kappa), \beta(\kappa))$, např. $\min_{\kappa}(a\,\alpha(\kappa) + b\,\beta(\kappa))$, tj. minimalizace **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 $\kappa$ stanovíme tak, aby chyba 1. druhu nastávala s danou pravděpodobností $\alpha$ zvanou **hladina významnosti** (nebo s menší pravděpodobností, nelze-li dosáhnout rovnost).

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

Hodnoty kritéria, která přesahují kritickou hodnotu (odpovídají výsledkům málo pravděpodobným při platnosti nulové hypotézy) považujeme za **statisticky významné** a v tom případě **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 $\beta$.

**Síla testu** $1 - \beta$.

Rozlišuje se

- **jednoduchá hypotéza**: nulové hypotéze odpovídá jediná hodnota parametru,
- **složená hypotéza**: nulové hypotéze odpovídá více hodnot parametru,

a dále

- **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 $H_0 : \vartheta \le c$. (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 $\alpha$ pro všechny hodnoty parametru vyhovující nulové hypotéze.

*(Statistická významnost neznamená významnost praktickou.)*

**Řešení:** Nulovou hypotézu zamítneme, právě když hodnota kritéria získaná z realizace nepadne do intervalu spolehlivosti pro koeficient spolehlivosti $1-\alpha$, tj. kritická hodnota je 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:** Testovací statistiku $T$, která roste s parametrem $\vartheta$ a má známé rozdělení, (přesněji její realizaci $t$) porovnáváme s kvantily příslušného rozdělení a zamítneme při extrémních hodnotách (nepravděpodobných při platnosti nulové hypotézy):

| $H_0$ | $H_1$ | zamítáme pro | dosažená významnost |
|---|---|---|---|
| $\vartheta \leq c$ | $\vartheta > c$ | $t > q_T(1-\alpha)$ | $1 - F_T(t)$ |
| $\vartheta \geq c$ | $\vartheta < c$ | $t < q_T(\alpha)$ | $F_T(t)$ |
| $\vartheta = c$ | $\vartheta \neq c$ | $t > q_T(1-\frac{\alpha}{2})$ nebo $t < q_T(\frac{\alpha}{2})$ | $2\min\left(F_T(t), 1-F_T(t)\right)$ |

V literatuře se setkáme i s následujícími případy hypotéz, které se však řeší stejně jako první dva výše uvedené:

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

## 6.2 Testy střední hodnoty normálního rozdělení

### 6.2.1 Při známém rozptylu $\sigma^2$

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

porovnáváme s kvantily **normovaného normálního rozdělení**:

| $H_0$ | zamítáme pro | dosažená významnost |
|---|---|---|
| $\mu \leq c$ | $t > \Phi^{-1}(1-\alpha)$ | $1-\Phi(t)$ |
| $\mu \geq c$ | $t < -\Phi^{-1}(1-\alpha) = \Phi^{-1}(\alpha)$ | $\Phi(t)$ |
| $\mu = c$ | $\lvert t\rvert > \Phi^{-1}(1-\frac{\alpha}{2})$ | $2\left(1-\Phi(\lvert t\rvert)\right)$ |

### 6.2.2 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 |
|---|---|---|
| $\mu \leq c$ | $t > q_{\mathrm{t}(n-1)}(1-\alpha)$ | $1 - F_{\mathrm{t}(n-1)}(t)$ |
| $\mu \geq c$ | $t < -q_{\mathrm{t}(n-1)}(1-\alpha)$ | $F_{\mathrm{t}(n-1)}(t)$ |
| $\mu = c$ | $\lvert t\rvert > q_{\mathrm{t}(n-1)}(1-\frac{\alpha}{2})$ | $2\left(1 - F_{\mathrm{t}(n-1)}(\lvert t\rvert)\right)$ |

## 6.3 Testy rozptylu normálního rozdělení

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

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

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

## 6.4 Porovnání dvou normálních rozdělení

**Předpoklad: Nezávislé** výběry

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

### 6.4.1 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 $\xi$ a $\eta$ 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 $x > 0$:

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

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

Je-li $\mathrm{D}X = \mathrm{D}Y = \sigma^2$, 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 |
|---|---|---|
| $\mathrm{D}X \leq \mathrm{D}Y$ | $t > q_{\mathrm{F}(m-1,n-1)}(1-\alpha)$ | $1 - F_{\mathrm{F}(m-1,n-1)}(t)$ |
| $\mathrm{D}X \geq \mathrm{D}Y$ | $t < q_{\mathrm{F}(m-1,n-1)}(\alpha)$ | $F_{\mathrm{F}(m-1,n-1)}(t)$ |
| $\mathrm{D}X = \mathrm{D}Y$ | $t < q_{\mathrm{F}(m-1,n-1)}(\frac{\alpha}{2})$ 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 $\xi, \eta$; 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 $s_{\boldsymbol{x}}^2 \geq s_{\boldsymbol{y}}^2$ testujeme

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

na rozdělení $\mathrm{F}(m-1, n-1)$:

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

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

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

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

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

### 6.4.2 Testy středních hodnot dvou normálních rozdělení se známým rozptylem $\sigma^2$

$$\overline{\boldsymbol{X}}_m \text{ má } \mathrm{N}\left(\mathrm{E}X, \frac{\sigma^2}{m}\right),$$

$$\overline{\boldsymbol{Y}}_n \text{ má } \mathrm{N}\left(\mathrm{E}Y, \frac{\sigma^2}{n}\right),$$

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

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 6.2.1).

### 6.4.3 Testy středních hodnot dvou normálních rozdělení se (stejným) neznámým rozptylem

**Předpoklad:** $\mathrm{D}X = \mathrm{D}Y = \sigma^2$

Nejprve ověříme tento předpoklad (viz kapitola 6.4.1).

*(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.].)*

Máme dva odhady $S^2_{\boldsymbol{X}}, S^2_{\boldsymbol{Y}}$ stejné hodnoty $\sigma^2$; použijeme jejich průměr vážený rozsahy výběrů ($-1$ kvůli výpočtu výběrového průměru):

$$\frac{(m-1)\, S^2_{\boldsymbol{X}}}{\sigma^2} \text{ má } \chi^2(m-1)\,,$$

$$\frac{(n-1)\, S^2_{\boldsymbol{Y}}}{\sigma^2} \text{ má } \chi^2(n-1)\,,$$

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

se střední hodnotou $m+n-2$,

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

má střední hodnotu 1 a

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

je nestranný odhad $\sigma^2$,

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

$$\overline{\boldsymbol{X}}_m \text{ má } \mathrm{N}\left(\mathrm{E}X, \frac{\sigma^2}{m}\right)\ ,$$

$$\overline{\boldsymbol{Y}}_n \text{ má } \mathrm{N}\left(\mathrm{E}Y, \frac{\sigma^2}{n}\right)\ ,$$

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

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

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

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

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

Testujeme realizaci $t$ na rozdělení $\mathrm{t}(m+n-2)$ (viz kapitola 6.2.2).

## 6.5 Testy středních hodnot dvou normálních rozdělení - párový pokus

*(dle [SH10])*
**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 a projevuje se proto synchronně v obou výběrech; proto výběry **nejsou navzájem nezávislé**. Měříme vždy obě veličiny současně.
**Předpoklad:** Náhodné veličiny $X_j, Y_j$ $(j = 1,\dots,n)$ mají normální rozdělení $\mathrm{N}(\mu_j, \sigma^2)$ se stálým rozptylem $\sigma^2$ a proměnnými středními hodnotami $\mu_j = \mathrm{E}X_j = \mathrm{E}Y_j$.
Můžeme použít náhodné veličiny $U_j := X_j - \mu_j, V_j := Y_j - \mu_j$ $(j = 1,\dots,n)$, které **jsou nezávislé** a mají rozdělení $\mathrm{N}(0,\sigma^2)$.
Náhodné veličiny $\Delta_j := X_j - Y_j = U_j - V_j$ $(j = 1,\dots,n)$ jsou nezávislé a mají rozdělení $\mathrm{N}(0, 2\sigma^2)$.
Výběrový průměr $\bar{\boldsymbol{\Delta}}$ má $\mathrm{N}\left(0, \frac{2\sigma^2}{n}\right)$.

### 6.5.1 Pro známý rozptyl $\sigma^2$

Neznámé parametry sdruženého rozdělení jsou $\mu_1,\dots,\mu_n$, ale nepotřebujeme je.
Dle kapitoly 6.2.1 (pro $c = 0$) testujeme

$$T := \frac{\bar{\boldsymbol{\Delta}}}{\sigma}\sqrt{\frac{n}{2}} = \frac{\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}}{\sigma}\sqrt{\frac{n}{2}}$$

na $\mathrm{N}(0,1)$.

### 6.5.2 Pro neznámý rozptyl

Neznámé parametry sdruženého rozdělení jsou $\boldsymbol{\Theta} = (\sigma^2, \mu_1,\dots,\mu_n)$, potřebujeme z nich pouze $\sigma^2 = \mathrm{D}X$.
Můžeme pracovat přímo s výběrem $(\Delta_1,\dots,\Delta_n)$ z normálního rozdělení.
Dle kapitoly 6.2.2 (pro $c = 0$) testujeme

$$T := \frac{\bar{\boldsymbol{\Delta}}}{S_\Delta}\sqrt{n}$$

na $\mathrm{t}(n-1)$.

**Cvičení:** Maximálně věrohodný odhad parametrů:

$$\ell(\boldsymbol{\Theta}) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\,\pi}\,\sigma} \exp\left(\frac{-(x_j-\mu_j)^2}{2\,\sigma^2}\right) \cdot \prod_{j=1}^{n} \frac{1}{\sqrt{2\,\pi}\,\sigma} \exp\left(\frac{-(y_j-\mu_j)^2}{2\,\sigma^2}\right),$$

$$L(\boldsymbol{\Theta}) = -\sum_{j=1}^{n} \frac{(x_j-\mu_j)^2}{2\,\sigma^2} - \sum_{j=1}^{n} \frac{(y_j-\mu_j)^2}{2\,\sigma^2} - 2\,n\,\ln\sigma - 2\,n\,\ln\sqrt{2\,\pi}\,,$$

$$\begin{aligned} 0 &= \frac{\partial L(\widehat{\boldsymbol{\vartheta}})}{\partial \widehat{\mu_j}} = \frac{\partial}{\partial \widehat{\mu_j}} \left(-\frac{(x_j-\widehat{\mu_j})^2}{2\,\widehat{\sigma^2}} - \frac{(y_j-\widehat{\mu_j})^2}{2\,\widehat{\sigma^2}}\right) \\ &= \frac{1}{\widehat{\sigma^2}} \left((x_j-\widehat{\mu_j}) + (y_j-\widehat{\mu_j})\right) = \frac{1}{\widehat{\sigma^2}} \left(x_j+y_j-2\,\widehat{\mu_j}\right), \\ \widehat{\mu_j} &= \frac{x_j+y_j}{2}\,, \qquad j=1,\ldots,n\,. \end{aligned}$$

*Odhady* $\widehat{\mu_j}$, $(j=1,\ldots,n)$ **nejsou konzistentní**.
Po jejich dosazení:

$$L(\widehat{\boldsymbol{\vartheta}}) = -\sum_{j=1}^{n} \frac{(x_j-y_j)^2}{4\,\widehat{\sigma^2}} - n\,\ln\widehat{\sigma^2} - 2\,n\,\ln\sqrt{2\,\pi}\,,$$

$$0 = \frac{\partial L(\widehat{\boldsymbol{\vartheta}})}{\partial \left(\widehat{\sigma^2}\right)} = \sum_{j=1}^{n} \frac{(x_j-y_j)^2}{4\,\left(\widehat{\sigma^2}\right)^2} - \frac{2\,n}{\widehat{\sigma^2}}\,,$$

$$\widehat{\sigma^2} = \frac{1}{2\,n} \sum_{j=1}^{n} (x_j-y_j)^2 = \frac{1}{2\,n} \sum_{j=1}^{n} \delta_j^2\,,$$

kde $\delta_j$ je realizace $\Delta_j$. *Odhad* $\widehat{\sigma^2}$ **je konzistentní**.

## 6.6 $\chi^2$-test dobré shody

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 $p_1,\ldots,p_k$.
Testujeme pomocí realizace náhodného výběru rozsahu $n$. Není důležité pořadí výsledků, pouze jejich **četnosti** $n_i$, resp. **relativní četnosti** $\frac{n_i}{n}$ $(i=1,\ldots,k)$. Porovnáváme četnost $n_i$ s **teoretickou četností** $n\,p_i$. Testovací statistikou je

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

Její rozdělení se pro $n\to\infty$ blíží $\chi^2(k-1)$.
Dosažená významnost: $1-F_{\chi^2(k-1)}(t)$. Nulovou hypotézu zamítáme pro $t > q_{\chi^2(k-1)}(1-\alpha)$, tj. $1-F_{\chi^2(k-1)}(t) < \alpha$.

**Cvičení 1.** *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* $\chi^2$ | 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 u voličů, ale poslední má více než* 2× *méně poslanců).*

### 6.6.1 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 (řekněme aspoň 5), aby náš předpoklad byl oprávněný.
**Modifikace:** Vychází-li teoretická četnost některých tříd příliš malá, sloučíme je s jinými třídami (pokud možno „blízkými“).

**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.

### 6.6.2 $\chi^2$-test dobré shody dvou rozdělení

*(dle [Mood a kol.])*
$H_0$ : Dvě diskrétní náhodné veličiny mají stejné diskrétní rozdělení.
Rozsahy výběrů jsou $m, n$ a četnosti výsledků $m_i, n_i$ $(i = 1, \ldots, k)$. Předpokládáme rozdělení

s neznámými teoretickými pravděpodobnostmi $p_i$ $(i = 1, \ldots, k)$.

$$\sum_{i=1}^{k} \frac{(m_i - m\,p_i)^2}{m\,p_i} \text{ se blíží } \chi^2(k-1)\,,$$

$$\sum_{i=1}^{k} \frac{(n_i - n\,p_i)^2}{n\,p_i} \text{ se blíží } \chi^2(k-1)\,,$$

$$T = \sum_{i=1}^{k} \frac{(m_i - m\,p_i)^2}{m\,p_i} + \sum_{i=1}^{k} \frac{(n_i - n\,p_i)^2}{n\,p_i} \text{ se blíží } \chi^2(2(k-1))\,.$$

Neznámé parametry $p_i$ odhadneme pomocí maxima věrohodnosti,

$$p_i = \frac{m_i + n_i}{m + n}\,,$$

z nich je jen $k-1$ nezávislých (neboť $\sum\limits_{i=1}^{k} p_i = 1$), takže výsledný počet stupňů volnosti je $2\,(k-1) - (k-1) = k-1$ a testujeme $T$ na $\chi^2(k-1)$. Nulovou hypotézu zamítáme pro $t > q_{\chi^2(k-1)}(1-\alpha)$, tj. $1 - F_{\chi^2(k-1)}(t) < \alpha$.
Praktičtější (ekvivalentní) vzorec:

$$T = \left(\frac{1}{m} + \frac{1}{n}\right) \sum_{i=1}^{k} \frac{(m_i - m\,p_i)^2}{p_i}\,.$$

### 6.6.3 $\chi^2$-test nezávislosti dvou rozdělení

*(dle [Likeš, Machek])*
$H_0$ : Dvě náhodné veličiny (jejichž rozdělení neznáme) jsou nezávislé.
$X$ nabývá $k$ hodnot s pravděpodobnostmi $p_1, \ldots, p_k$,
$Y$ nabývá $m$ hodnot s pravděpodobnostmi $q_1, \ldots, q_m$.
Realizace dvojrozměrného náhodného výběru $((x_1, y_1), \ldots, (x_n, y_n))$ obsahuje dvojice realizací náhodných veličin $X, Y$; z výsledků nás zajímají opět pouze četnosti $n_{ij}$ $(i = 1, \ldots, k;\ j = 1, \ldots, m)$. Ty bývají uspořádány do tzv. **kontingenční tabulky**. Počet tříd je $k\,m$.
Za předpokladu nezávislosti jsou pravděpodobnosti výsledků $p_i\,q_j$ $(i = 1, \ldots, k;\ j = 1, \ldots, m)$,

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

Neznámé parametry $p_i, q_j$ odhadneme pomocí maxima věrohodnosti,

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

z nich je jen $(k-1) + (m-1)$ nezávislých (neboť $\sum\limits_{i=1}^{k} p_i = 1,\ \sum\limits_{j=1}^{m} q_j = 1$), 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 $t > q_{\chi^2((k-1)\,(m-1))}(1-\alpha)$, tj. $1 - F_{\chi^2((k-1)\,(m-1))}(t) < \alpha$.

## 6.7 Korelace, její odhad a testování

*(dle [Likeš, Machek])*
**Korelace** $\varrho(X,Y)$ náhodných veličin $X,Y$ (s nenulovým rozptylem) je střední hodnota součinu odpovídajících normovaných veličin $\frac{X-\mathrm{E}X}{\sigma_X}\cdot\frac{Y-\mathrm{E}Y}{\sigma_Y}$,

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

Je nulová pro nezávislé náhodné veličiny, ale i pro některé jiné, tzv. **nekorelované**.
Extrémní hodnoty $\pm 1$ odpovídají lineární závislosti mezi $X,Y$.

Na základě dvojrozměrného náhodného výběru $((X_1,Y_1),\ldots,(X_n,Y_n))$ můžeme korelaci odhadnout pomocí **výběrového koeficientu korelace**

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

Jeho realizace

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

neboť je to kosinus úhlu vektorů

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

neboli korelace empirického rozdělení, $r_{\boldsymbol{x},\boldsymbol{y}}=\varrho(\mathrm{Emp}(\boldsymbol{x},\boldsymbol{y}))$.
Pro výpočet se používá jednoprůchodový vzorec:

$$R_{\boldsymbol{X},\boldsymbol{Y}}=\frac{n\sum_{j=1}^{n}x_j\,y_j-\left(\sum_{j=1}^{n}x_j\right)\left(\sum_{j=1}^{n}y_j\right)}{\sqrt{\left(n\sum_{j=1}^{n}x_j^2-\left(\sum_{j=1}^{n}x_j\right)^2\right)\left(n\sum_{j=1}^{n}y_j^2-\left(\sum_{j=1}^{n}y_j\right)^2\right)}}\,.$$

### 6.7.1 Test nekorelovanosti dvou normálních rozdělení

**Předpoklad:** Dvojrozměrná náhodná veličina $(X,Y)$ má (dvojrozměrné) normální rozdělení, $n\geq 3$.
$H_0:\varrho(X,Y)=0$ ($X,Y$ jsou nekorelované).
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 6.2.2.

## 6.8 Neparametrické testy

Jsou použitelné bez ohledu na typ rozdělení, jsou však slabší.

### 6.8.1 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 $\mathrm{Bin}\left(n, \frac{1}{2}\right)$. Nulovou hypotézu zamítáme pro

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

(Podobně pro jednostranné testy.) Výpočet kvantilů je pracný, ale kritické hodnoty jsou tabelovány (v závislosti na $n$ a hladině významnosti).
Dosažená významnost se počítá o trochu snáze.
Pro velká $n$ používáme centrální limitní větu a testujeme

$$T_0 := \frac{2T - n}{\sqrt{n}}$$

na $\mathrm{N}(0, 1)$.
Lze použít i k porovnání dvou mediánů u párového pokusu.
**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* $n \ln n$.

### 6.8.2 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

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

porovnáme s tabulkou kritických hodnot pro tento test.

# 7 Co zde nebylo

## 7.1 Více o zobrazení náhodné veličiny funkcí a o součtu náhodných veličin

## 7.2 Diskretizace

## 7.3 Směs pravděpodobností

## 7.4 Charakteristická funkce náhodné veličiny

## 7.5 Důkaz centrální limitní věty

# 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.

[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.

[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.