ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

Fakulta elektrotechnická
Katedra kybernetiky

Opakování základních pojmů statistiky (a pravděpodobnosti)

Petr Pošík

Části dokumentu jsou převzaty (i doslovně) z
Mirko Navara: Pravděpodobnost a matematická statistika,
`https://cw.felk.cvut.cz/lib/exe/fetch.php/courses/a6m33ssl/pms_print.pdf`
s laskavým svolením autora.

Úvod

Pravděpodobnost a statistika

Teorie pravděpodobnosti: Nástroj pro rozhodování v systémech, jejichž popis známe, ale jejichž *budoucí* stav a chování závisí na okolnostech, které neznáme. Deduktivní uvažování.

Statistika: Nástroj pro hledání a ověřování pravděpodobnostního popisu reálných systémů na základě jejich pozorování. Induktivní uvažování.

Pravděpodobnost a statistika

Teorie pravděpodobnosti: Nástroj pro rozhodování v systémech, jejichž popis známe, ale jejichž *budoucí* stav a chování závisí na okolnostech, které neznáme. Deduktivní uvažování.

Statistika: Nástroj pro hledání a ověřování pravděpodobnostního popisu reálných systémů na základě jejich pozorování. Induktivní uvažování.

Dedukce: Ze znalosti *„obecného“* usuzujeme na vlastnosti *„konkrétního“*. *Specializace* obecných znalostí, zde využití *pravděpodobnosti*.

Indukce: Ze znalosti *„konkrétního“* usuzujeme na vlastnosti *„obecného.“* *Generalizace* poznatků, zde využití *statistického usuzování*.

Základy pravděpodobnosti

Jev

Elementární jevy jsou všechny možné vzájemně se vylučující výsledky nějakého experimentu. Jejich množinu označme Ω.

Jev je podmnožina množiny elementárních jevů, $A \subseteq \Omega$.

- Jakýkoli výrok o výsledku experimentu, u něhož lze vždy rozhodnout, zda platí nebo ne (jev nastal nebo nenastal).
- Ekvivalentně lze místo výroků a výrokových operací používat jim příslušné množiny elementárních jevů a množinové operace.

Některé zvláštní jevy a jejich kombinace:

- **Jev jistý:** Ω, **1**
- **Jev nemožný:** $\emptyset$, **0**
- **Konjunkce jevů („and“):** $A \cap B$
- **Disjunkce jevů („or“):** $A \cup B$
- **Jev opačný** k A: $\overline{A} = \Omega \setminus A$
- $A \Rightarrow B$: $A \subseteq B$
- **Jevy neslučitelné**: $A_1, \ldots, A_n : \bigcap_{i \leq n} A_i = \emptyset$
- **Jevy po dvou neslučitelné** (=**vzájemně se vylučující**): $A_1, \ldots, A_n : \forall i, j \in \{1, \ldots, n\},\ i \neq j : A_i \cap A_j = \emptyset$

Úplný systém jevů

Úplný systém jevů tvoří jevy B_i, $i \in I$, jestliže jsou po dvou neslučitelné a $\bigcup_{i \in I} B_i = \Omega$.

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

Úplný systém jevů

Úplný systém jevů tvoří jevy B_i, $i \in I$, jestliže jsou po dvou neslučitelné a $\bigcup_{i \in I} B_i = \Omega$.

- 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 platí

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

Speciálně pro $\{C, \bar{C}\}$:

$$P(A) = P(A \cap C) + P(A \cap \bar{C}) \, .$$

Definice pravděpodobnosti

Klasická **Laplaceova definice pravděpodobnosti** založená na relativních četnostech výskytu jevu trpí mnoha neduhy:

- Platí jen pro n *stejně možných* elementárních jevů. (Stejně možných?)
- Nedovoluje nekonečné množiny jevů, geometrickou pravděpodobnost, ...
- Nedovoluje iracionální hodnoty pravděpodobnosti.

Definice pravděpodobnosti

Klasická **Laplaceova definice pravděpodobnosti** založená na relativních četnostech výskytu jevu trpí mnoha neduhy:

- Platí jen pro n *stejně možných* elementárních jevů. (Stejně možných?)
- Nedovoluje nekonečné množiny jevů, geometrickou pravděpodobnost, ...
- Nedovoluje iracionální hodnoty pravděpodobnosti.

Kolmogorovova definice pravděpodobnosti:

- Množina elementárních jevů Ω může být nekonečná a el. jevy nemusí být stejně pravděpodobné.
- **Jevy** jsou podmnožiny množiny Ω, ale *ne nutně všechny*; tvoří podmnožinu $\mathcal{A} \subseteq \Omega^2$. ($\Omega^2$ je potenční množina, powerset, množina všech podmnožin množiny Ω.)

Chceme být schopni definovat pravděpodobnost pro jakoukoli konjunkci a disjunkci jevů; **jevové pole $\mathcal{A}$** proto nemůže být jakékoli, musí to být **σ-algebra**, tj. musí splňovat podmínky:

1. $\emptyset \in \mathcal{A}$.
2. $A \in \mathcal{A} \Rightarrow \bar{A} \in \mathcal{A}$.
3. $(\forall n \in \mathbb{N} : A_n \in \mathcal{A}) \Rightarrow \bigcup_{n \in \mathbb{N}} A_n \in \mathcal{A}$. (Uzavřenost na *spočetná* sjednocení.)

Definice pravděpodobnosti

Klasická **Laplaceova definice pravděpodobnosti** založená na relativních četnostech výskytu jevu trpí mnoha neduhy:

- Platí jen pro n *stejně možných* elementárních jevů. (Stejně možných?)
- Nedovoluje nekonečné množiny jevů, geometrickou pravděpodobnost, ...
- Nedovoluje iracionální hodnoty pravděpodobnosti.

Kolmogorovova definice pravděpodobnosti:

- Množina elementárních jevů Ω může být nekonečná a el. jevy nemusí být stejně pravděpodobné.
- **Jevy** jsou podmnožiny množiny Ω, ale *ne nutně všechny*; tvoří podmnožinu $\mathcal{A} \subseteq \Omega^2$. ($\Omega^2$ je potenční množina, powerset, množina všech podmnožin množiny Ω.)

Chceme být schopni definovat pravděpodobnost pro jakoukoli konjunkci a disjunkci jevů; **jevové pole $\mathcal{A}$** proto nemůže být jakékoli, musí to být **σ-algebra**, tj. musí splňovat podmínky:

1. $\emptyset \in \mathcal{A}$.
2. $A \in \mathcal{A} \Rightarrow \bar{A} \in \mathcal{A}$.
3. $(\forall n \in \mathbb{N} : A_n \in \mathcal{A}) \Rightarrow \bigcup_{n \in \mathbb{N}} A_n \in \mathcal{A}$. (Uzavřenost na *spočetná* sjednocení.)

Borelova σ-algebra je nejmenší σ-algebra podmnožin $\mathbb{R}$, která obsahuje všechny intervaly.

Pravděpodobnost

Pravděpodobnost (=pravděpodobnostní míra) je funkce $P : \mathcal{A} \rightarrow \langle 0, 1 \rangle$ splňující podmínky

1. $P(\Omega) = 1$
2. $P\left(\bigcup\limits_{n \in \mathbb{N}} A_n\right) = \sum\limits_{n \in \mathbb{N}} P(A_n)$,

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

Pravděpodobnost

Pravděpodobnost (=pravděpodobnostní míra) je funkce $P : \mathcal{A} \to \langle 0, 1 \rangle$ splňující podmínky

1. $P(\Omega) = 1$
2. $P\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} P(A_n)$,
   pokud jsou množiny (=jevy) $A_n, n \in \mathbb{N}$, po dvou neslučitelné (*spočetná aditivita*).

Pravděpodobnostní prostor je trojice $(\Omega, \mathcal{A}, P)$, kde Ω je neprázdná množina, $\mathcal{A}$ je σ-algebra podmnožin množiny Ω a $P : \mathcal{A} \to \langle 0, 1 \rangle$ je pravděpodobnost.

Pravděpodobnost

Pravděpodobnost (=pravděpodobnostní míra) je funkce $P : \mathcal{A} \rightarrow \langle 0, 1 \rangle$ splňující podmínky

1. $P(\Omega) = 1$
2. $P\left(\bigcup\limits_{n \in \mathbb{N}} A_n\right) = \sum\limits_{n \in \mathbb{N}} P(A_n)$,
   pokud jsou množiny (=jevy) $A_n, n \in \mathbb{N}$, po dvou neslučitelné (*spočetná aditivita*).

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

Vlastnosti pravděpodobnosti:

- $P(A) \in \langle 0, 1 \rangle$

Pravděpodobnost

Pravděpodobnost (=pravděpodobnostní míra) je funkce $P : \mathcal{A} \to \langle 0, 1 \rangle$ splňující podmínky

1. $P(\Omega) = 1$
2. $P\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \sum_{n \in \mathbb{N}} P(A_n)$,
   pokud jsou množiny (=jevy) $A_n, n \in \mathbb{N}$, po dvou neslučitelné (*spočetná aditivita*).

Pravděpodobnostní prostor je trojice $(\Omega, \mathcal{A}, P)$, kde Ω je neprázdná množina, $\mathcal{A}$ je σ-algebra podmnožin množiny Ω a $P : \mathcal{A} \to \langle 0, 1 \rangle$ je pravděpodobnost.

Vlastnosti pravděpodobnosti:

- $P(A) \in \langle 0, 1 \rangle$
- $P(\mathbf{0}) = 0, \qquad P(\mathbf{1}) = 1$

Pravděpodobnost

Pravděpodobnost (=pravděpodobnostní míra) je funkce $P : \mathcal{A} \rightarrow \langle 0, 1 \rangle$ splňující podmínky

1. $P(\Omega) = 1$
2. $P\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \sum_{n \in \mathbb{N}} P(A_n)$,
   pokud jsou množiny (=jevy) $A_n, n \in \mathbb{N}$, po dvou neslučitelné (*spočetná aditivita*).

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

Vlastnosti pravděpodobnosti:

- $P(A) \in \langle 0, 1 \rangle$
- $P(\mathbf{0}) = 0, \qquad P(\mathbf{1}) = 1$
- $P(\overline{A}) = 1 - P(A)$

Pravděpodobnost

Pravděpodobnost (=pravděpodobnostní míra) je funkce $P : \mathcal{A} \rightarrow \langle 0, 1 \rangle$ splňující podmínky

1. $P(\Omega) = 1$
2. $P\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \sum_{n \in \mathbb{N}} P(A_n)$,
   pokud jsou množiny (=jevy) $A_n, n \in \mathbb{N}$, po dvou neslučitelné (*spočetná aditivita*).

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

Vlastnosti pravděpodobnosti:

- $P(A) \in \langle 0, 1 \rangle$
- $P(\mathbf{0}) = 0, \qquad P(\mathbf{1}) = 1$
- $P(\bar{A}) = 1 - P(A)$
- $A \subseteq B \Rightarrow P(A) \leq P(B)$

Pravděpodobnost

Pravděpodobnost (=pravděpodobnostní míra) je funkce $P : \mathcal{A} \rightarrow \langle 0, 1 \rangle$ splňující podmínky

1. $P(\Omega) = 1$
2. $P\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} P(A_n)$,

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

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

Vlastnosti pravděpodobnosti:

- $P(A) \in \langle 0, 1 \rangle$
- $P(\mathbf{0}) = 0, \qquad P(\mathbf{1}) = 1$
- $P(\overline{A}) = 1 - P(A)$
- $A \subseteq B \Rightarrow P(A) \leq P(B)$
- $A \subseteq B \Rightarrow P(B \setminus A) = P(B) - P(A)$

Pravděpodobnost

Pravděpodobnost (=pravděpodobnostní míra) je funkce $P : \mathcal{A} \to \langle 0, 1 \rangle$ splňující podmínky

1. $P(\Omega) = 1$
2. $P\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \sum_{n \in \mathbb{N}} P(A_n)$,
   pokud jsou množiny (=jevy) $A_n, n \in \mathbb{N}$, po dvou neslučitelné (*spočetná aditivita*).

Pravděpodobnostní prostor je trojice $(\Omega, \mathcal{A}, P)$, kde Ω je neprázdná množina, $\mathcal{A}$ je σ-algebra podmnožin množiny Ω a $P : \mathcal{A} \to \langle 0, 1 \rangle$ je pravděpodobnost.

Vlastnosti pravděpodobnosti:

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

Pravděpodobnost

Pravděpodobnost (=pravděpodobnostní míra) je funkce $P : \mathcal{A} \rightarrow \langle 0, 1\rangle$ splňující podmínky

1. $P(\Omega) = 1$
2. $P\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} P(A_n)$,
   pokud jsou množiny (=jevy) $A_n, n \in \mathbb{N}$, po dvou neslučitelné (*spočetná aditivita*).

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

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

Nezávislé jevy

Definice: Jevy A a B jsou **nezávislé**, pokud platí

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

Nezávislé jevy

Definice: Jevy A a B jsou **nezávislé**, pokud platí

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

Jsou-li A, B nezávislé, pak

- $P(A \cup B) = P(A) + P(B) - P(A) \cdot P(B)$,
- jsou nezávislé taky dvojice $A, \bar{B}$ a $\bar{A}, B$ a $\bar{A}, \bar{B}$.

Nezávislé jevy

Definice: Jevy A a B jsou **nezávislé**, pokud platí

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

Jsou-li A, B nezávislé, pak

- $P(A \cup B) = P(A) + P(B) - P(A) \cdot P(B)$,
- jsou nezávislé taky dvojice $A, \bar{B}$ a $\bar{A}, B$ a $\bar{A}, \bar{B}$.

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

Nezávislé jevy

Definice: Jevy A a B jsou **nezávislé**, pokud platí

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

Jsou-li A, B nezávislé, pak

- $P(A \cup B) = P(A) + P(B) - P(A) \cdot P(B)$,
- jsou nezávislé taky dvojice $A, \bar{B}$ a $\bar{A}, B$ a $\bar{A}, \bar{B}$.

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

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

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

pro všechny konečné podmnožiny $\mathcal{K} \subseteq \mathcal{M}$.

Podmíněná pravděpodobnost

Podmíněná pravděpodobnost jevu A za podmínky B je definována jako

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \qquad P(B) \neq 0.$$

- $P(A)$ známe z pravděpodobnostního popisu systému. Dostaneme-li navíc informaci o tom, že nastal jev B, podmíněná pravděpodobnost $P(A|B)$ je naše aktualizovaná znalost o pravděpodobnosti jevu A.
- $P(\bar{B}|B) = 0$, což odpovídá naší znalosti, že jev $\bar{B}$ nemůže nastat, když nastal jev B.
- Podm. pravděpodobnost je chápána též jako funkce $P(.|B) : \mathcal{A} \rightarrow \langle 0, 1 \rangle$ a je to pravděpodobnost v původním smyslu.

Podmíněná pravděpodobnost

Podmíněná pravděpodobnost jevu A za podmínky B je definována jako

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \qquad P(B) \neq 0.$$

- $P(A)$ známe z pravděpodobnostního popisu systému. Dostaneme-li navíc informaci o tom, že nastal jev B, podmíněná pravděpodobnost $P(A|B)$ je naše aktualizovaná znalost o pravděpodobnosti jevu A.
- $P(\bar{B}|B) = 0$, což odpovídá naší znalosti, že jev $\bar{B}$ nemůže nastat, když nastal jev B.
- Podm. pravděpodobnost je chápána též jako funkce $P(.|B) : \mathcal{A} \to \langle 0, 1 \rangle$ a je to pravděpodobnost v původním smyslu.

Vlastnosti:

- $P(\Omega|B) = 1$, $P(\emptyset|B) = 0$.
- $B \subseteq A \Rightarrow P(A|B) = 1$.
- $P(A \cap B) = 0 \Rightarrow P(A|B) = 0$.
- Pokud se jevy $A_1, \ldots, A_n$ vzájemně vylučují, pak

$$P\left(\bigcup_{n\in\mathbb{N}} A_n | B\right) = \sum_{n\in\mathbb{N}} P(A_n|B).$$

- Je-li $P(A|B)$ definována, jsou **jevy A, B nezávislé** právě tehdy, když $P(A|B) = P(A)$.

Bayesova věta

Věta o úplné pravděpodobnosti: Je-li $B_i, i \in I$, (spočetný) úplný systém jevů a $\forall i \in I : P(B_i) \neq 0$, pak pro každý jev A platí

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

Bayesova věta

Věta o úplné pravděpodobnosti: Je-li $B_i, i \in I$, (spočetný) úplný systém jevů a $\forall i \in I : P(B_i) \neq 0$, pak pro každý jev A platí

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

Platí:

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

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

Bayesova věta

Věta o úplné pravděpodobnosti: Je-li $B_i, i \in I$, (spočetný) úplný systém jevů a $\forall i \in I : P(B_i) \neq 0$, pak pro každý jev A platí

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

Platí:

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

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

Bayesova věta: Je-li $B_i, i \in I$, (spočetný) úplný systém jevů a $\forall i \in I : P(B_i) \neq 0$, pak pro každý jev $A, P(A) \neq 0$, platí

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

Bayesova věta

Věta o úplné pravděpodobnosti: Je-li $B_i, i \in I$, (spočetný) úplný systém jevů a $\forall i \in I : P(B_i) \neq 0$, pak pro každý jev A platí

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

Platí:

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

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

Bayesova věta: Je-li $B_i, i \in I$, (spočetný) úplný systém jevů a $\forall i \in I : P(B_i) \neq 0$, pak pro každý jev $A, P(A) \neq 0$, platí

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

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

Problém: Ke stanovení **aposteriorních pravděpodobností** $P(B_i|A)$ potřebujeme znát **apriorní pravděpodobnosti** $P(B_i)$.

Náhodná veličina

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

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

Náhodná veličina

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

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

Rozdělení náhodné veličiny je popsáno praděpodobnostmi

$$P_X(I) = P[X \in I] = P(\{\omega \in \Omega : X(\omega) \in I\})$$

definovanými pro lib. interval I. Funkce P_X je **pravděpodobnostní míra** na Borelově σ-algebře a splňuje

- $P_X(\mathbb{R}) = 1$, $P_X(\emptyset) = 0$,
- $P_X(\bigcup_{n\in\mathbb{N}} I_n) = \sum_{n\in\mathbb{N}} P_X(I_n)$, pokud jsou množiny $I_n, n \in \mathbb{N}$, navzájem disjunktní,
- $P_X(\mathbb{R} \setminus I) = 1 - P_X(I)$,
- jestliže $I \subseteq J$, pak $P_X(I) \leq P_X(J)$ a $P_X(J \setminus I) = P_X(J) - P_X(I)$.

Náhodná veličina

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

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

Rozdělení náhodné veličiny je popsáno praděpodobnostmi

$$P_X(I) = P[X \in I] = P(\{\omega \in \Omega : X(\omega) \in I\})$$

definovanými pro lib. interval I. Funkce P_X je **pravděpodobnostní míra** na Borelově σ-algebře a splňuje

- $P_X(\mathbb{R}) = 1$, $P_X(\emptyset) = 0$,
- $P_X(\bigcup_{n\in\mathbb{N}} I_n) = \sum_{n\in\mathbb{N}} P_X(I_n)$, pokud jsou množiny $I_n, n \in \mathbb{N}$, navzájem disjunktní,
- $P_X(\mathbb{R} \setminus I) = 1 - P_X(I)$,
- jestliže $I \subseteq J$, pak $P_X(I) \le P_X(J)$ a $P_X(J \setminus I) = P_X(J) - P_X(I)$.

Distribuční funkce náhodné veličiny X je funkce $F_X : \mathbb{R} \to \langle 0, 1\rangle$ definovaná jako

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

Distribuční funkce je

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

Náhodný vektor

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

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

Náhodný vektor lze považovat za vektor náhodných veličin $\boldsymbol{X} = (X_1, \ldots, X_n)$, tj. lze psát $\boldsymbol{X}(\omega) = (X_1(\omega), \ldots, X_n(\omega))$, kde zobrazení $X_k : \Omega \rightarrow \mathbb{R}, k = 1, \ldots, n$, jsou náhodné veličiny.

Náhodný vektor

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

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

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

Rozdělení náhodného vektoru je popsáno pravděpodobnostmi

$$P_{\boldsymbol{X}}(I_1 \times \ldots \times I_n) = P[X_1 \in I_1, \ldots, X_n \in I_n] = P(\{\omega \in \Omega : X_1(\omega) \in I_1, \ldots, X_n(\omega) \in I_n\}),$$

kde $I_1, \ldots, I_n$ jsou intervaly v $\mathbb{R}$, tj.

$$P_{\boldsymbol{X}}(\boldsymbol{I}) = P[\boldsymbol{X} \in \boldsymbol{I}] = P\left(\{\omega \in \Omega : \boldsymbol{X}(\omega) \in \boldsymbol{I}\}\right),$$

definovanými pro libovolnou borelovskou množinu $\boldsymbol{I} \in \mathbb{R}^n$.

Náhodný vektor

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

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

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

Rozdělení náhodného vektoru je popsáno pravděpodobnostmi

$$P_{\boldsymbol{X}}(I_1 \times \ldots \times I_n) = P[X_1 \in I_1, \ldots, X_n \in I_n] = P(\{\omega \in \Omega : X_1(\omega) \in I_1, \ldots, X_n(\omega) \in I_n\}),$$

kde $I_1, \ldots, I_n$ jsou intervaly v $\mathbb{R}$, tj.

$$P_{\boldsymbol{X}}(\boldsymbol{I}) = P[\boldsymbol{X} \in \boldsymbol{I}] = P\left(\{\omega \in \Omega : \boldsymbol{X}(\omega) \in \boldsymbol{I}\}\right),$$

definovanými pro libovolnou borelovskou množinu $\boldsymbol{I} \in \mathbb{R}^n$.

Distribuční funkce náhodného vektoru $\boldsymbol{X}$ je funkce $F_{\boldsymbol{X}} : \mathbb{R}^n \to \langle 0, 1 \rangle$ definovaná jako

$$F_{\boldsymbol{X}}(t_1, \ldots, t_n) = P[X_1 \le t_1, \ldots, X_n \le t_n] = P_{\boldsymbol{X}}\left((-\infty, t_1\rangle \times \ldots \times (-\infty, t_n\rangle\right),$$

která má tyto vlastnosti:

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

Nezávislost náhodných veličin

Náh. veličiny $X_1, \ldots, X_n$ jsou **nezávislé**, pokud pro libovolné intervaly $I_1, \ldots, I_n$ platí

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

Ekvivalentně stačí pro všechna $t_1, \ldots, t_n \in \mathbb{R}$ požadovat

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

takže pro sdruženou distribuční funkci *nezávislých* náhodných veličin musí platit

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

Náhodné veličiny $X_1, \ldots, X_n$ jsou **po dvou nezávislé**, pokud jsou každé dvě z nich nezávislé. To je slabší podmínka než **nezávislost všech veličin** $X_1, \ldots, X_n$.

Druhy náhodných veličin

Diskrétní náhodná veličina má *po částech konstantní distribuční funkci*.

Druhy náhodných veličin

Diskrétní náhodná veličina má *po částech konstantní distribuční funkci*.

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

Druhy náhodných veličin

Diskrétní náhodná veličina má *po částech konstantní distribuční funkci*.

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

Popisuje ji pravděpodobnostní funkce $p_X(t) = P_X(\{t\}) = P[X = t]$, která nabývá nenulových hodnot na nejvýše spočetné množině Ω_X a která splňuje

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

Druhy náhodných veličin

Diskrétní náhodná veličina má *po částech konstantní distribuční funkci*.

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

Popisuje ji pravděpodobnostní funkce $p_X(t) = P_X(\{t\}) = P[X = t]$, která nabývá nenulových hodnot na nejvýše spočetné množině Ω_X a která splňuje

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

Spojitá náhodná veličina má *spojitou distribuční funkci*.

Druhy náhodných veličin

Diskrétní náhodná veličina má *po částech konstantní distribuční funkci*.

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

Popisuje ji pravděpodobnostní funkce $p_X(t) = P_X(\{t\}) = P[X = t]$, která nabývá nenulových hodnot na nejvýše spočetné množině Ω_X a která splňuje

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

Spojitá náhodná veličina má *spojitou distribuční funkci*.

Absolutně spojité náhodné veličiny jsou ty, které mají **hustotu pravděpodobnosti**, což je nezáporná funkce $f_X : \mathbb{R} \to \langle 0, \infty)$ taková, že

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

- Hustota splňuje $\int_{-\infty}^{\infty} f_X(u) du = 1$.
- Není určena jednoznačně, lze volit $f_X(t) = \frac{\mathrm{d} F_X(t)}{\mathrm{d} t}$, pokud existuje.
- $P_X(\{t\}) = 0$ pro všechna t.

Kvantilová funkce

Distribuční funkce $F_X(t)$ říká, jak velká část populace má hodnotu proměnné X menší nebo rovnu určitému limitu t (např. kolik procent studentů FEL má vážený průměr 1.5 nebo lepší).

Kvantilová funkce

Distribuční funkce $F_X(t)$ říká, jak velká část populace má hodnotu proměnné X menší nebo rovnu určitému limitu t (např. kolik procent studentů FEL má vážený průměr 1.5 nebo lepší).

Obráceně se lze ptát, jaká hodnota náhodné veličiny odpovídá určité části populace (např. jaký vážený průměr je třeba, aby se student dostal mezi 5 % nejlepších). Pro určité $\alpha \in \langle 0, 1 \rangle$ tak hledáme takové t, pro které $F_X(t) = \alpha$. Protože ale distribuční funkce může být na nějakém intervalu konstantní, nemusí být hledané t jediné, mohou tvořit omezený interval. Proto:

Kvantilová funkce

Distribuční funkce $F_X(t)$ říká, jak velká část populace má hodnotu proměnné X menší nebo rovnu určitému limitu t (např. kolik procent studentů FEL má vážený průměr 1.5 nebo lepší).

Obráceně se lze ptát, jaká hodnota náhodné veličiny odpovídá určité části populace (např. jaký vážený průměr je třeba, aby se student dostal mezi 5 % nejlepších). Pro určité $\alpha \in \langle 0, 1 \rangle$ tak hledáme takové t, pro které $F_X(t) = \alpha$. Protože ale distribuční funkce může být na nějakém intervalu konstantní, nemusí být hledané t jediné, mohou tvořit omezený interval. Proto:

Kvantilová funkce $q_X : (0, 1) \rightarrow \mathbb{R}$ je definována jako

$$q_X(\alpha) = \frac{1}{2}(\sup\{t \in \mathbb{R} : P[X < t] \le \alpha\} + \inf\{t \in \mathbb{R} : P[X \le t] \ge \alpha\})$$

- Číslo $q_X(\alpha)$ se nazývá **α-kvantil** náhodné veličiny X.
- **Medián** náhodné veličiny je $q_X(0.5)$.
- **Dolní**, $q_X(0.25)$, a **horní kvartil**, $q_X(0.75)$, dále **decily**, **centily** neboli **percentily**, …
- q_X je neklesající.
- F_X a q_X jsou navzájem inverzní tam, kde jsou spojité a rostoucí.

Střední hodnota

Střední hodnota náhodné proměnné X se značí $\mathrm{E}\,X$ nebo μ_X a je definována zvlášť pro

- *diskrétní* náhodnou veličinu X:

$$\mathrm{E}\,X = \sum_{t \in \mathbb{R}} t \cdot p_X(t) = \sum_{t \in \Omega_X} t \cdot p_X(t),$$

- *spojitou* náhodnou veličinu Y:

$$\mathrm{E}\,Y = \int_{-\infty}^{\infty} t \cdot f_Y(t)\,\mathrm{d}\,t.$$

Pro oba případy platí vzorec využívající kvantilovou funkci

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

Střední hodnota

Střední hodnota náhodné proměnné X se značí $\mathrm{E}\,X$ nebo μ_X a je definována zvlášť pro

- *diskrétní* náhodnou veličinu X:

$$\mathrm{E}\,X = \sum_{t\in\mathbb{R}} t \cdot p_X(t) = \sum_{t\in\Omega_X} t \cdot p_X(t),$$

- *spojitou* náhodnou veličinu Y:

$$\mathrm{E}\,Y = \int_{-\infty}^{\infty} t \cdot f_Y(t)\,\mathrm{d}\,t.$$

Pro oba případy platí vzorec využívající kvantilovou funkci

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

Vlastnosti:

- $\mathrm{E}\,r = r$, $\mathrm{E}(\mathrm{E}\,X) = \mathrm{E}\,X$
- $\mathrm{E}(X+Y) = \mathrm{E}\,X + \mathrm{E}\,Y$, $\mathrm{E}(X+r) = \mathrm{E}\,X + r$, $\mathrm{E}(X-Y) = \mathrm{E}\,X - \mathrm{E}\,Y$
- $\mathrm{E}(rX+sY) = r\,\mathrm{E}\,X + s\,\mathrm{E}\,Y$
- Pouze pro *nezávislé* veličiny: $\mathrm{E}(X\cdot Y) = \mathrm{E}\,X \cdot \mathrm{E}\,Y$.

Rozptyl (disperze)

Rozptyl náhodné proměnné X se značí $\mathrm{D}\,X$, σ_X^2, $\mathrm{var}\,X$, nebo $Var(X)$ a je definován jako

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

$$\mathrm{E}(X^2) = (\mathrm{E}\,X)^2 + \mathrm{D}\,X,$$

nebo také

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

Rozptyl (disperze)

Rozptyl náhodné proměnné X se značí $\mathrm{D}\,X$, σ_X^2, $\mathrm{var}\,X$, nebo $Var(X)$ a je definován jako

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

nebo také

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

Vlastnosti:

- $\mathrm{D}\,X \geq 0$
- $\mathrm{D}\,r = 0$
- $\mathrm{D}(X + r) = \mathrm{D}\,X$
- $\mathrm{D}(rX) = r^2\,\mathrm{D}\,X$
- Pouze pro *nezávislé* veličiny: $\mathrm{D}(X + Y) = \mathrm{D}\,X + \mathrm{D}(Y)$, $\mathrm{D}(X - Y) = \mathrm{D}\,X + \mathrm{D}(Y)$.

Směrodatná odchylka

Směrodatná odchylka náhodné proměnné X se značí σ_X, je definována jako

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

a *má stejný fyzikální rozměr* jako původní náhodná veličina (na rozdíl od rozptylu).

Směrodatná odchylka

Směrodatná odchylka náhodné proměnné X se značí σ_X, je definována jako

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

a *má stejný fyzikální rozměr* jako původní náhodná veličina (na rozdíl od rozptylu).

Vlastnosti:

- $\sigma_X \geq 0$
- $\sigma_r = 0$
- $\sigma_{X+r} = \sigma_X$
- $\sigma_{rX} = |r|\sigma_X$
- Pouze pro *nezávislé* náhodné veličiny: $\sigma_{X+Y} = \sqrt{\mathrm{D}\,X + \mathrm{D}\,Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}$.

Normovaná náhodná veličina

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

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

má-li vzorec smysl. Zpětná transformace je

$$X = \operatorname{E} X + \sigma_X \operatorname{norm} X.$$

Diskrétní rozdělení

- **Diracovo:** jediný možný výsledek $r \in \mathbb{R}$.
- **Alternativní (Bernoulliho):** dva možné výsledky (obvykle označené 0 a 1), jeden z nich (1) má pravděpodobnost q.
- **Rovnoměrné:** m možných, stejně pravděpodobných výsledků.
- **Binomické:** počet úspěchů z m nezávislých pokusů, je-li v každém stejná pravděpodobnost úspěchu $q \in \langle 0, 1 \rangle$. Součet m nezávislých alternativních rozdělení.
- **Poissonovo:** limitní případ binomického rozdělení pro $m \to \infty$ při konstantním $mq = \lambda > 0$ (tedy $q \to 0$).
- **Geometrické:** počet úspěchů do prvního neúspěchu, je-li v každém pokusu stejná pravděpodobnost úspěchu $q \in (0, 1)$.
- **Hypergeometrické:** Počet výskytů v m vzorcích vybraných z M objektů, v nichž je celkem K výskytů ($1 \le m \le K \le M$).

Spojitá rozdělení

- **Rovnoměrné** $R(a, b)$**:** p_X je konstantní na intervalu $\langle a, b\rangle$. Hustota $f_{R(a,b)}$, distribuční funkce $F_{R(a,b)}$.
- **Normální (Gaussovo)**
  - **normované** $\mathrm{N}(0, 1)$**:** hustota ϕ, distribuční funkce Φ.
  - **obecné** $\mathrm{N}(\mu, \sigma^2)$**:** hustota $f_{\mathrm{N}(\mu,\sigma^2)}$, distribuční funkce $F_{\mathrm{N}(\mu,\sigma^2)}$.
- **Logaritmickonormální (lognormální)** $\mathrm{LN}(\mu, \sigma^2)$**:** rozdělení náhodné veličiny $X = e^Y$, kde Y má $\mathrm{N}(\mu, \sigma^2)$.
- **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 δ, nikoli na t.

Náhodný vektor 2

Diskrétní náhodný vektor má všechny složky diskrétní.

- Lze jej popsat **sdruženou pravděpodobnostní funkcí** $p_{\mathbf{X}} : \mathbb{R}^n \rightarrow \langle 0, 1 \rangle$

  $$p_{\mathbf{X}}(t_1, \ldots, t_n) = P[X_1 = t_1, \ldots, X_N = t_N],$$

  která je nenulová jen ve spočetně mnoha bodech.
- *Diskrétní* náh. veličiny $X_1, \ldots, X_n$ jsou **nezávislé** právě tehdy, když pro všechna $t_1, \ldots, t_n \in \mathbb{R}$ platí

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

Náhodný vektor 2

Diskrétní náhodný vektor má všechny složky diskrétní.

- Lze jej popsat **sdruženou pravděpodobnostní funkcí** $p_{\mathbf{X}} : \mathbb{R}^n \to \langle 0, 1 \rangle$

$$p_{\mathbf{X}}(t_1, \ldots, t_n) = P[X_1 = t_1, \ldots, X_N = t_N],$$

  která je nenulová jen ve spočetně mnoha bodech.
- *Diskrétní* náh. veličiny $X_1, \ldots, X_n$ jsou **nezávislé** právě tehdy, když pro všechna $t_1, \ldots, t_n \in \mathbb{R}$ platí

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

Spojitý náhodný vektor má všechny složky spojité.

- Lze jej popsat **sdruženou hustotou pravděpodobnosti**, což je každá nezáporná funkce $f_{\mathbf{X}} : \mathbb{R}^n \to \langle 0, \infty)$ taková, že pro všechny $t_1, \ldots, t_n \in \mathbb{R}$

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

  Pokud to jde, volíme $f_{\mathbf{X}}(u_1, \ldots, u_n) = \frac{\partial}{\partial t_1} \ldots \frac{\partial}{\partial t_n} F_{\mathbf{X}}(t_1, \ldots, t_n)$.
- *Spojité* náh. veličiny $X_1, \ldots, X_n$ jsou **nezávislé** právě tehdy, když pro *skoro* všechna $t_1, \ldots, t_n \in \mathbb{R}$ platí

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

Číselné charakteristiky náhodného vektoru

Pro náhodný vektor $\boldsymbol{X} = (X_1, \ldots, X_n)$ definujeme

- **střední hodnotu** $\mathrm{E}\,\boldsymbol{X} = (\mathrm{E}\,X_1, \ldots, \mathrm{E}\,X_n)$,
- **rozptyl** $\mathrm{D}\,\boldsymbol{X} = (\mathrm{D}\,X_1, \ldots, \mathrm{D}\,X_n)$,
- **kovarianční matici**

$$\boldsymbol{\Sigma}_{\boldsymbol{X}} = \begin{pmatrix} \mathrm{D}\,X_1 & \mathrm{cov}(X_1, X_2) & \cdots & \mathrm{cov}(X_1, X_n) \\ \mathrm{cov}(X_2, X_1) & \mathrm{D}\,X_2 & \cdots & \mathrm{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_n, X_1) & \mathrm{cov}(X_n, X_2) & \cdots & \mathrm{D}\,X_n \end{pmatrix},$$

  která je symetrická, pozitivně semidefinitní a na diagonále má rozptyly $\mathrm{D}\,X_i = \mathrm{cov}(X_i, X_i)$,
- **korelační matici**

$$\boldsymbol{\rho}_{\boldsymbol{X}} = \begin{pmatrix} 1 & \rho(X_1, X_2) & \cdots & \rho(X_1, X_n) \\ \rho(X_2, X_1) & 1 & \cdots & \rho(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(X_n, X_1) & \rho(X_n, X_2) & \cdots & 1 \end{pmatrix},$$

  která je symetrická, pozitivně semidefinitní a na diagonále má jedničky ($\rho(X_i, X_i) = 1$).

Kovariance a korelace

Kovariance dvou náhodných veličin X, Y je míra toho, jak moc se proměnné X, Y společně mění. Je definována (existují-li rozptyly $\mathrm{D}\,X, \mathrm{D}\,Y$) jako

$$\begin{aligned}\operatorname{cov}(X, Y) &= \mathrm{E}((X - \mathrm{E}\,X)(Y - \mathrm{E}\,Y)) = \\ &= \mathrm{E}(XY) - \mathrm{E}\,X \cdot \mathrm{E}\,Y\end{aligned}$$

Kovariance a korelace

Kovariance dvou náhodných veličin X, Y je míra toho, jak moc se proměnné X, Y společně mění. Je definována (existují-li rozptyly $\mathrm{D}\,X, \mathrm{D}\,Y$) jako

$$\begin{aligned}\operatorname{cov}(X,Y) &= \mathrm{E}((X-\mathrm{E}\,X)(Y-\mathrm{E}\,Y)) = \\ &= \mathrm{E}(XY) - \mathrm{E}\,X \cdot \mathrm{E}\,Y\end{aligned}$$

Poznámka: Pro rozptyl součtu 2 náhodných veličin (i závislých) platí

$$\mathrm{D}(X+Y) = \mathrm{D}\,X + \mathrm{D}\,Y + 2\operatorname{cov}(X,Y).$$

Kovariance a korelace

Kovariance dvou náhodných veličin X, Y je míra toho, jak moc se proměnné X, Y společně mění. Je definována (existují-li rozptyly $\mathrm{D}\,X, \mathrm{D}\,Y$) jako

$$\begin{aligned} \operatorname{cov}(X, Y) &= \mathrm{E}((X - \mathrm{E}\,X)(Y - \mathrm{E}\,Y)) = \\ &= \mathrm{E}(XY) - \mathrm{E}\,X \cdot \mathrm{E}\,Y \end{aligned}$$

Poznámka: Pro rozptyl součtu 2 náhodných veličin (i závislých) platí

$$\mathrm{D}(X + Y) = \mathrm{D}\,X + \mathrm{D}\,Y + 2\operatorname{cov}(X, Y).$$

Vlastnosti kovariance:

- $\operatorname{cov}(X, X) = \mathrm{D}\,X$, $\operatorname{cov}(X, -X) = -\mathrm{D}\,X$
- $\operatorname{cov}(X, Y) = \operatorname{cov}(Y, X)$
- $\operatorname{cov}(aX + b, cY + d) = ac\operatorname{cov}(X, Y)$
- Pro *nezávislé* náhodné veličiny X, Y je $\operatorname{cov}(X, Y) = 0$.

Kovariance a korelace

Kovariance dvou náhodných veličin X, Y je míra toho, jak moc se proměnné X, Y společně mění. Je definována (existují-li rozptyly $\mathrm{D}\,X, \mathrm{D}\,Y$) jako

$$\begin{aligned}\operatorname{cov}(X,Y) &= \mathrm{E}((X-\mathrm{E}\,X)(Y-\mathrm{E}\,Y)) = \\ &= \mathrm{E}(XY) - \mathrm{E}\,X \cdot \mathrm{E}\,Y\end{aligned}$$

Poznámka: Pro rozptyl součtu 2 náhodných veličin (i závislých) platí

$$\mathrm{D}(X+Y) = \mathrm{D}\,X + \mathrm{D}\,Y + 2\operatorname{cov}(X,Y).$$

Vlastnosti kovariance:

- $\operatorname{cov}(X,X) = \mathrm{D}\,X$, $\operatorname{cov}(X,-X) = -\mathrm{D}\,X$
- $\operatorname{cov}(X,Y) = \operatorname{cov}(Y,X)$
- $\operatorname{cov}(aX+b, cY+d) = ac\operatorname{cov}(X,Y)$
- Pro *nezávislé* náhodné veličiny X, Y je $\operatorname{cov}(X,Y) = 0$.

Korelace (Pearsonův korelační koeficient) dvou náhodných veličin X, Y popisuje sílu lineární závislosti. Definujeme jej jako kovarianci normovaných veličin X, Y, tj.

$$\begin{aligned}\rho(X,Y) &= \operatorname{cov}(\operatorname{norm} X, \operatorname{norm} Y) = \\ &= \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \\ &= \mathrm{E}(\operatorname{norm} X \cdot \operatorname{norm} Y)\end{aligned}$$

Kovariance a korelace

Kovariance dvou náhodných veličin X, Y je míra toho, jak moc se proměnné X, Y společně mění. Je definována (existují-li rozptyly $\mathrm{D}\,X, \mathrm{D}\,Y$) jako

$$\begin{aligned}\operatorname{cov}(X,Y) &= \mathrm{E}((X-\mathrm{E}\,X)(Y-\mathrm{E}\,Y)) = \\ &= \mathrm{E}(XY) - \mathrm{E}\,X \cdot \mathrm{E}\,Y\end{aligned}$$

Poznámka: Pro rozptyl součtu 2 náhodných veličin (i závislých) platí

$$\mathrm{D}(X+Y) = \mathrm{D}\,X + \mathrm{D}\,Y + 2\operatorname{cov}(X,Y).$$

Vlastnosti kovariance:

- $\operatorname{cov}(X,X) = \mathrm{D}\,X, \operatorname{cov}(X,-X) = -\mathrm{D}\,X$
- $\operatorname{cov}(X,Y) = \operatorname{cov}(Y,X)$
- $\operatorname{cov}(aX+b, cY+d) = ac\operatorname{cov}(X,Y)$
- Pro *nezávislé* náhodné veličiny X, Y je $\operatorname{cov}(X,Y) = 0$.

Korelace (Pearsonův korelační koeficient) dvou náhodných veličin X, Y popisuje sílu lineární závislosti. Definujeme jej jako kovarianci normovaných veličin X, Y, tj.

$$\begin{aligned}\rho(X,Y) &= \operatorname{cov}(\operatorname{norm} X, \operatorname{norm} Y) = \\ &= \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \\ &= \mathrm{E}(\operatorname{norm} X \cdot \operatorname{norm} Y)\end{aligned}$$

Vlastnosti korelace:

- $\rho(X,Y) \in \langle -1, 1 \rangle$
- $\rho(X,X) = 1, \rho(X,-X) = -1$
- $\rho(X,Y) = \rho(Y,X)$
- $\rho(aX+b, cY+d) = \operatorname{sign} ac\rho(X,Y)$
- Pro *nezávislé* náhodné veličiny X, Y je $\rho(X,Y) = 0$.

Kovariance a korelace

Kovariance dvou náhodných veličin X, Y je míra toho, jak moc se proměnné X, Y společně mění. Je definována (existují-li rozptyly $\mathrm{D}\,X, \mathrm{D}\,Y$) jako

$$\begin{aligned}\operatorname{cov}(X,Y) &= \mathrm{E}((X-\mathrm{E}\,X)(Y-\mathrm{E}\,Y)) = \\ &= \mathrm{E}(XY) - \mathrm{E}\,X \cdot \mathrm{E}\,Y\end{aligned}$$

Poznámka: Pro rozptyl součtu 2 náhodných veličin (i závislých) platí

$$\mathrm{D}(X+Y) = \mathrm{D}\,X + \mathrm{D}\,Y + 2\operatorname{cov}(X,Y).$$

Vlastnosti kovariance:

- $\operatorname{cov}(X,X) = \mathrm{D}\,X$, $\operatorname{cov}(X,-X) = -\mathrm{D}\,X$
- $\operatorname{cov}(X,Y) = \operatorname{cov}(Y,X)$
- $\operatorname{cov}(aX+b, cY+d) = ac\operatorname{cov}(X,Y)$
- Pro *nezávislé* náhodné veličiny X, Y je $\operatorname{cov}(X,Y) = 0$.

Korelace (Pearsonův korelační koeficient) dvou náhodných veličin X, Y popisuje sílu lineární závislosti. Definujeme jej jako kovarianci normovaných veličin X, Y, tj.

$$\begin{aligned}\rho(X,Y) &= \operatorname{cov}(\operatorname{norm} X, \operatorname{norm} Y) = \\ &= \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \\ &= \mathrm{E}(\operatorname{norm} X \cdot \operatorname{norm} Y)\end{aligned}$$

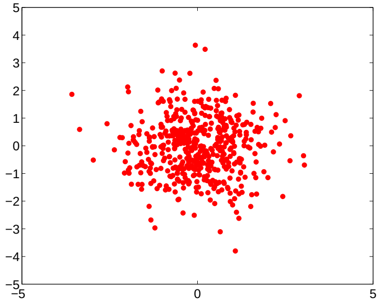
Vlastnosti korelace:

- $\rho(X,Y) \in \langle -1, 1 \rangle$
- $\rho(X,X) = 1$, $\rho(X,-X) = -1$
- $\rho(X,Y) = \rho(Y,X)$
- $\rho(aX+b, cY+d) = \operatorname{sign} ac\,\rho(X,Y)$
- Pro *nezávislé* náhodné veličiny X, Y je $\rho(X,Y) = 0$.

Pokud $\rho(X,Y) = 0$ a $\operatorname{cov}(X,Y) = 0$, *neznamená to*, že veličiny X, Y jsou nezávislé! Takové veličiny nazýváme **nekorelované**.

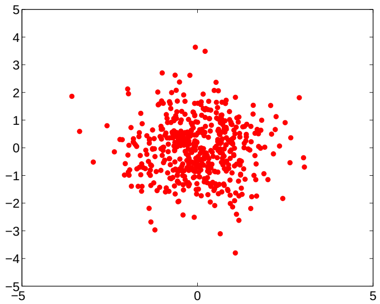
Korelace: příklady

Korelační koeficienty pro náhodné výběry (viz později) ze sdruženého rozdělení náhodných veličin X a Y:



Korelace: příklady

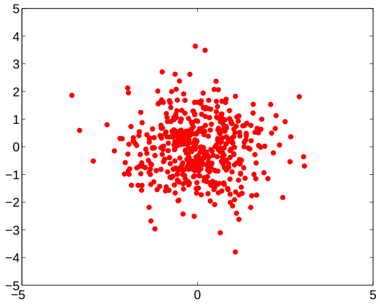
Korelační koeficienty pro náhodné výběry (viz později) ze sdruženého rozdělení náhodných veličin X a Y:

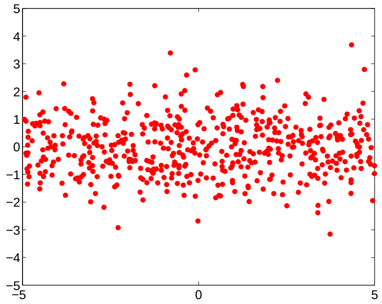


$r = 0$

Korelace: příklady

Korelační koeficienty pro náhodné výběry (viz později) ze sdruženého rozdělení náhodných veličin X a Y:

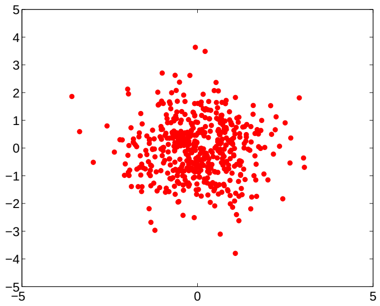




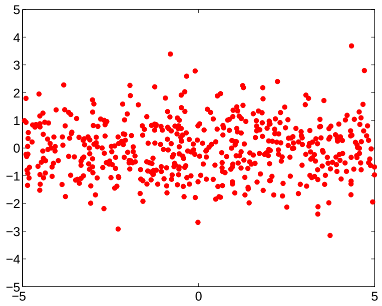
$$r = 0$$

Korelace: příklady

Korelační koeficienty pro náhodné výběry (viz později) ze sdruženého rozdělení náhodných veličin X a Y:



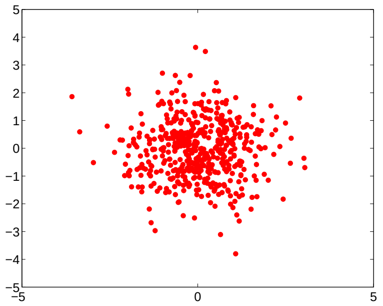
$r = 0$



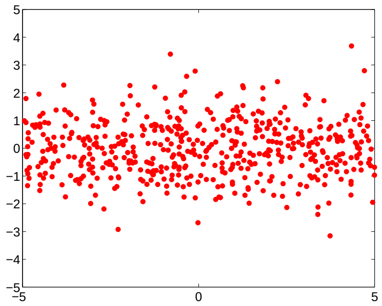
$r = 0$

Korelace: příklady

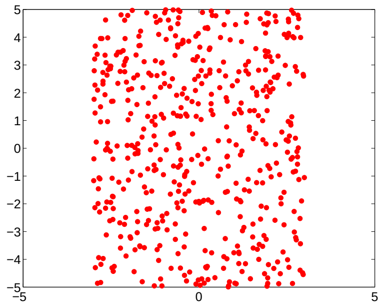
Korelační koeficienty pro náhodné výběry (viz později) ze sdruženého rozdělení náhodných veličin X a Y:



$r = 0$

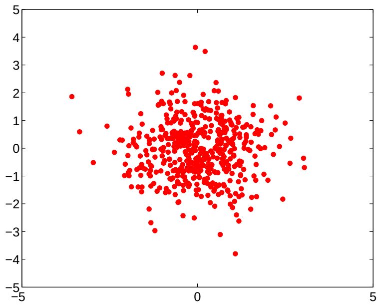


$r = 0$

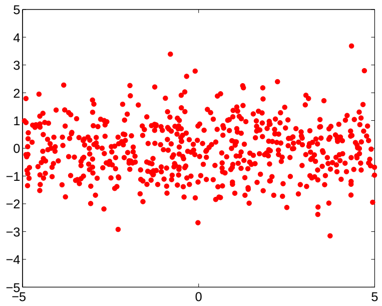


Korelace: příklady

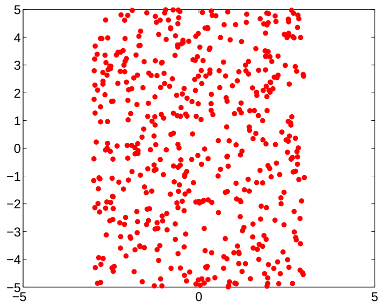
Korelační koeficienty pro náhodné výběry (viz později) ze sdruženého rozdělení náhodných veličin X a Y:



$r = 0$



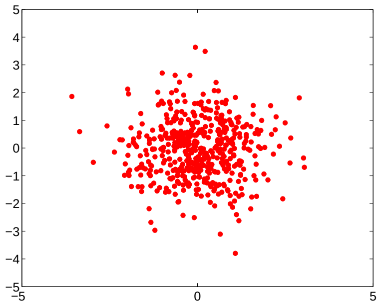
$r = 0$



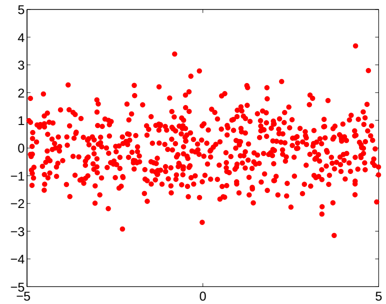
$r = 0$

Korelace: příklady

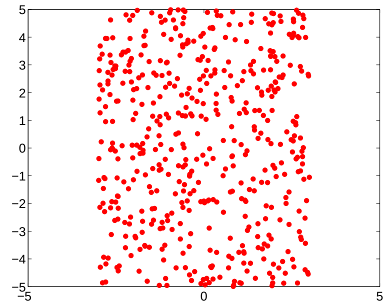
Korelační koeficienty pro náhodné výběry (viz později) ze sdruženého rozdělení náhodných veličin X a Y:



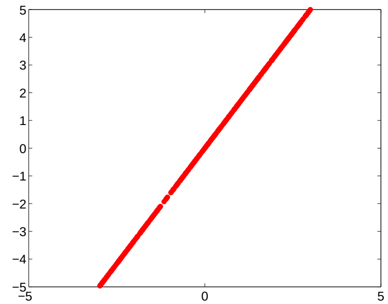
$r = 0$



$r = 0$

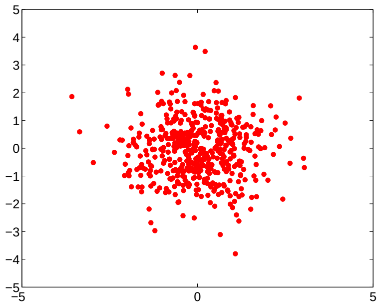


$r = 0$

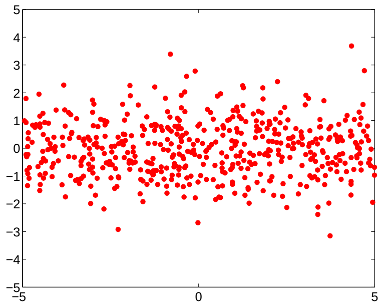


Korelace: příklady

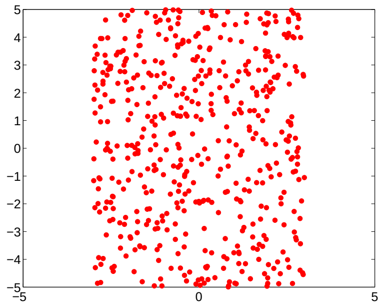
Korelační koeficienty pro náhodné výběry (viz později) ze sdruženého rozdělení náhodných veličin X a Y:



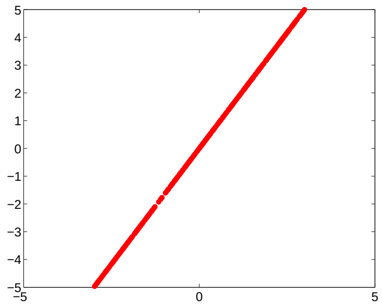
$r = 0$



$r = 0$



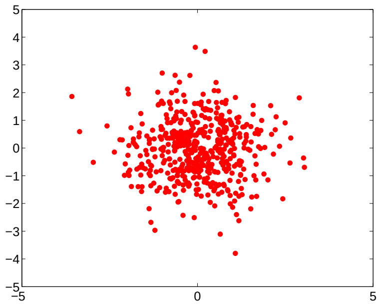
$r = 0$



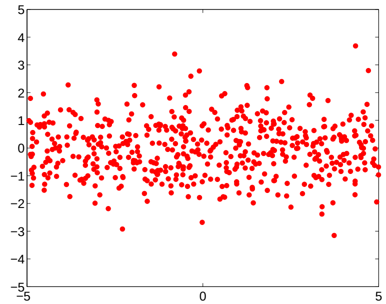
$r = 1$

Korelace: příklady

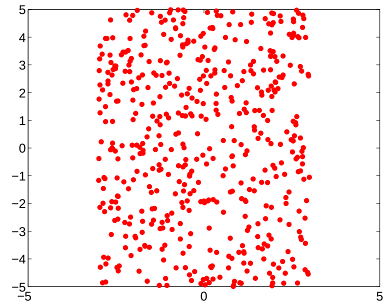
Korelační koeficienty pro náhodné výběry (viz později) ze sdruženého rozdělení náhodných veličin X a Y:



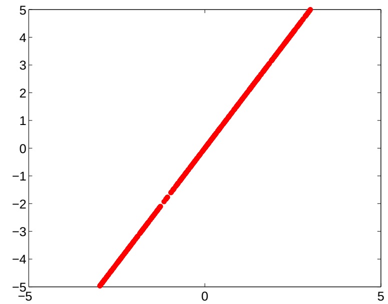
$r = 0$



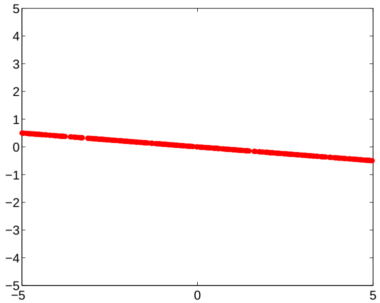
$r = 0$



$r = 0$

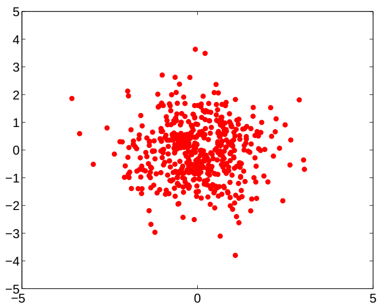


$r = 1$

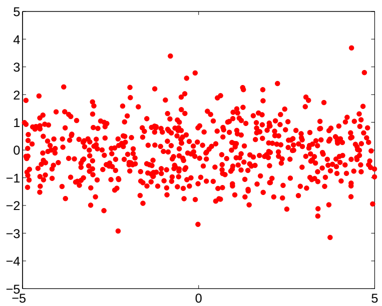


Korelace: příklady

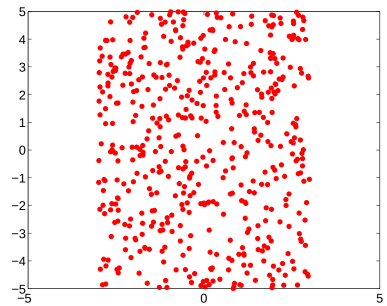
Korelační koeficienty pro náhodné výběry (viz později) ze sdruženého rozdělení náhodných veličin X a Y:



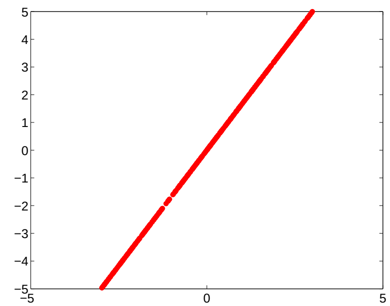
$r = 0$



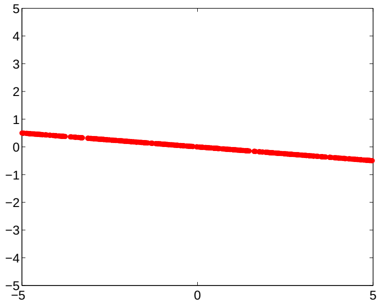
$r = 0$



$r = 0$



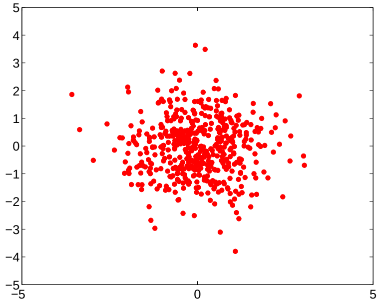
$r = 1$



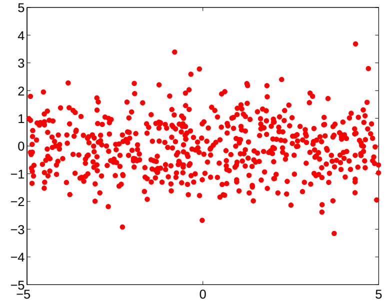
$r = -1$

Korelace: příklady

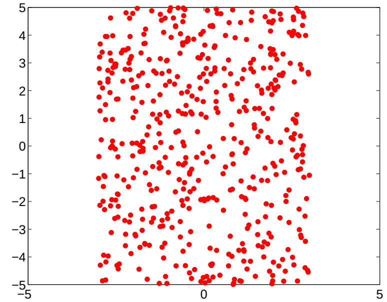
Korelační koeficienty pro náhodné výběry (viz později) ze sdruženého rozdělení náhodných veličin X a Y:



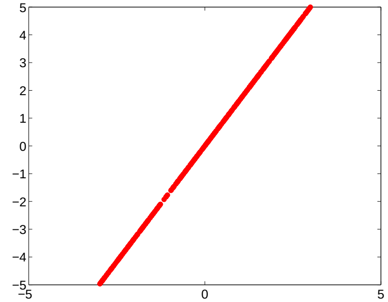
$r = 0$



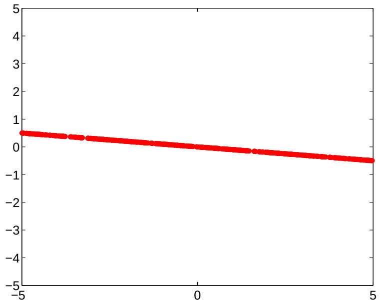
$r = 0$



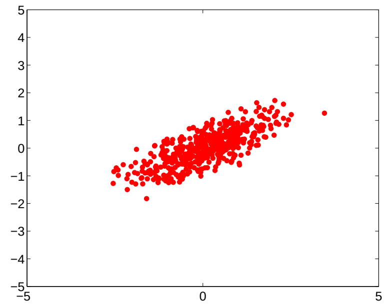
$r = 0$



$r = 1$

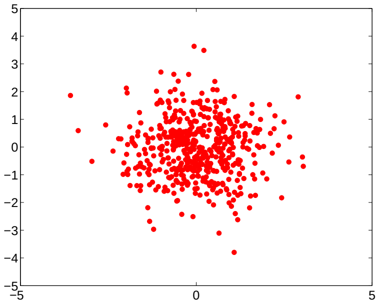


$r = -1$

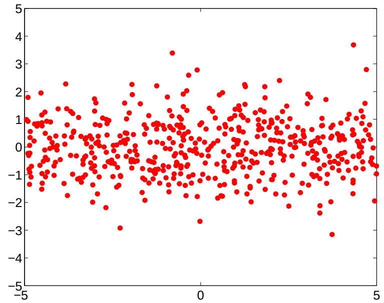


Korelace: příklady

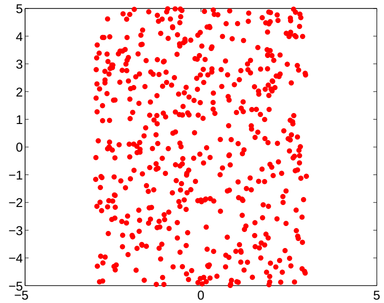
Korelační koeficienty pro náhodné výběry (viz později) ze sdruženého rozdělení náhodných veličin X a Y:



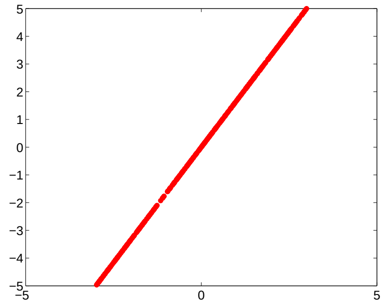
$r = 0$



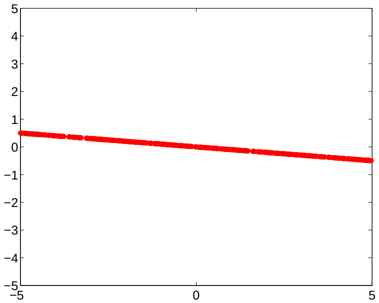
$r = 0$



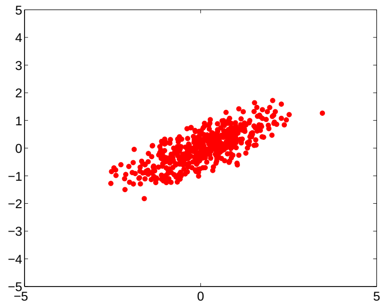
$r = 0$



$r = 1$



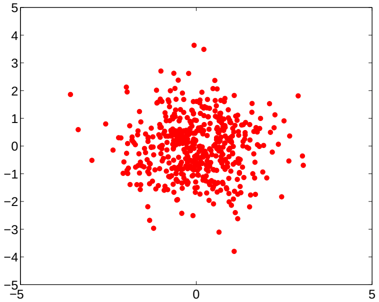
$r = -1$



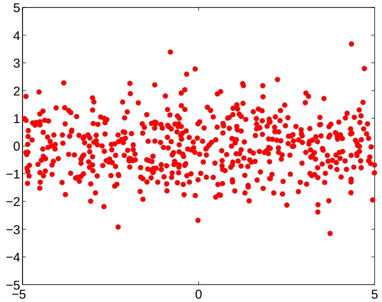
$r = 0.76$

Korelace: příklady

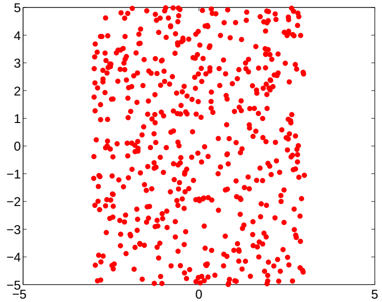
Korelační koeficienty pro náhodné výběry (viz později) ze sdruženého rozdělení náhodných veličin X a Y:



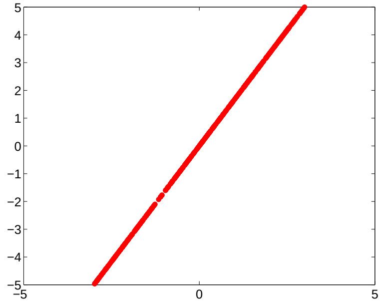
$r = 0$



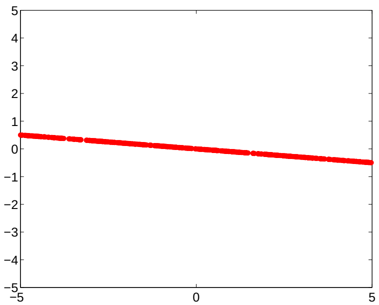
$r = 0$



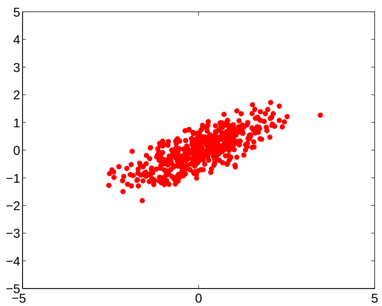
$r = 0$



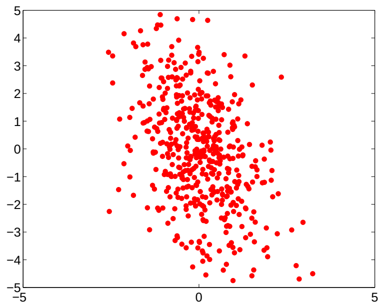
$r = 1$



$r = -1$

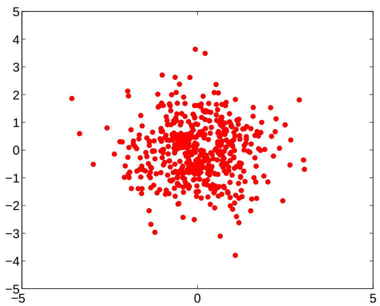


$r = 0.76$

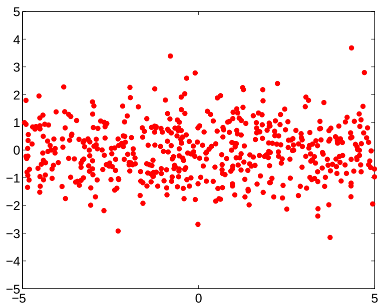


Korelace: příklady

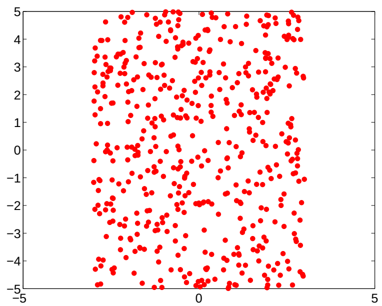
Korelační koeficienty pro náhodné výběry (viz později) ze sdruženého rozdělení náhodných veličin X a Y:



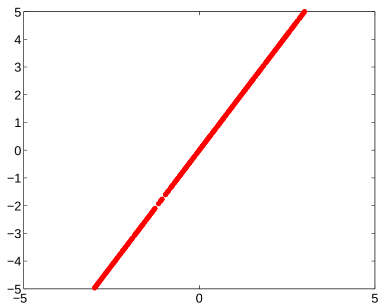
$r = 0$



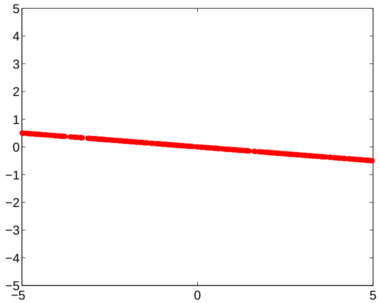
$r = 0$



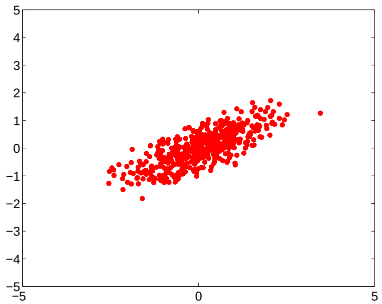
$r = 0$



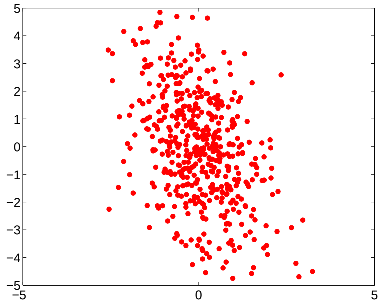
$r = 1$



$r = -1$



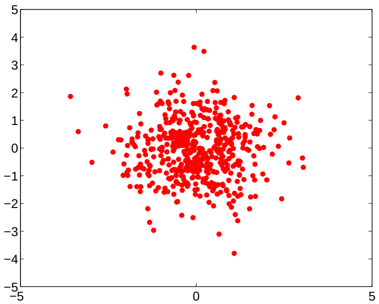
$r = 0.76$



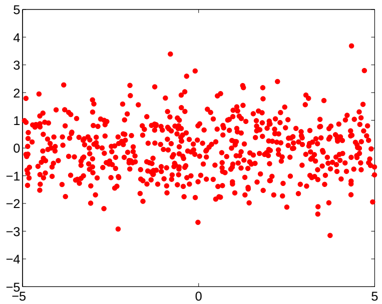
$r = -0.44$

Korelace: příklady

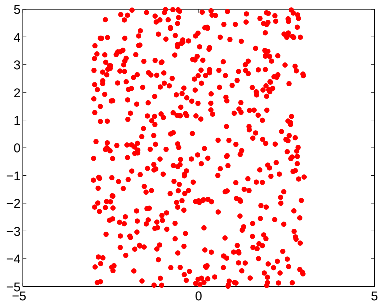
Korelační koeficienty pro náhodné výběry (viz později) ze sdruženého rozdělení náhodných veličin X a Y:



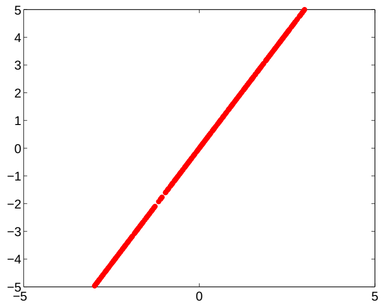
$r = 0$



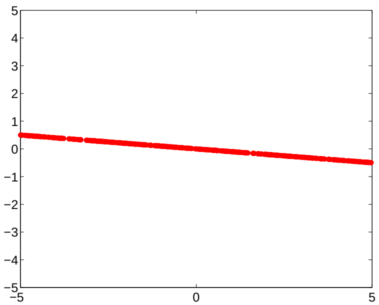
$r = 0$



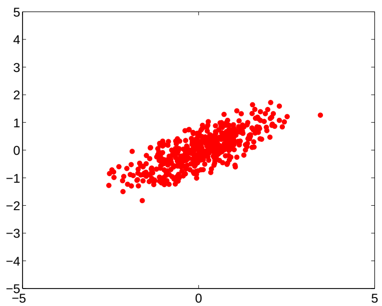
$r = 0$



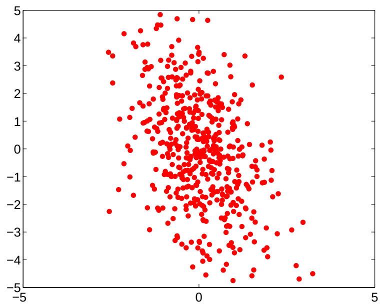
$r = 1$



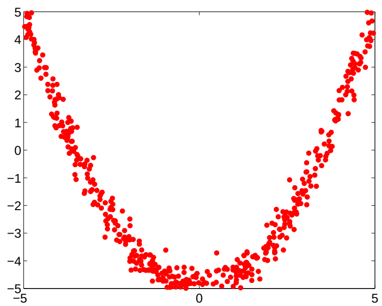
$r = -1$



$r = 0.76$

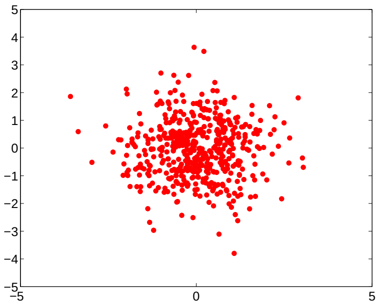


$r = -0.44$

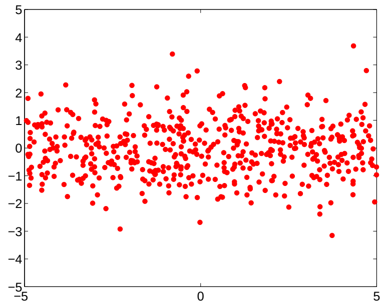


Korelace: příklady

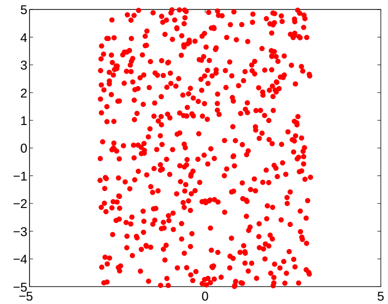
Korelační koeficienty pro náhodné výběry (viz později) ze sdruženého rozdělení náhodných veličin X a Y:



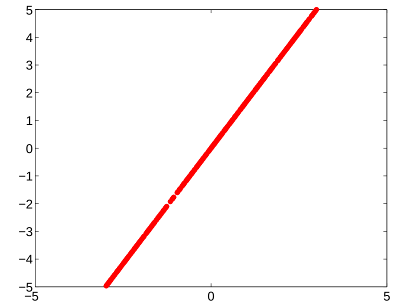
$r = 0$



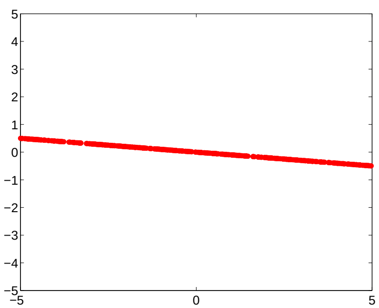
$r = 0$



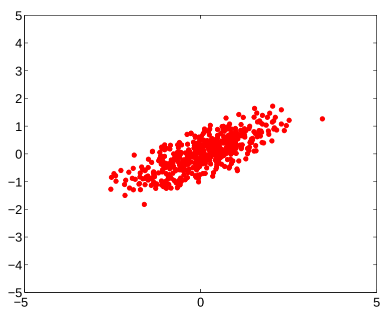
$r = 0$



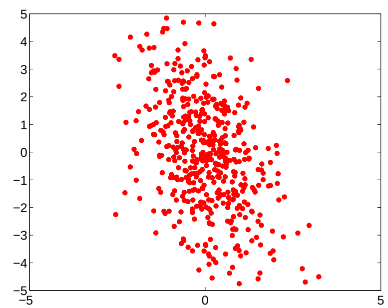
$r = 1$



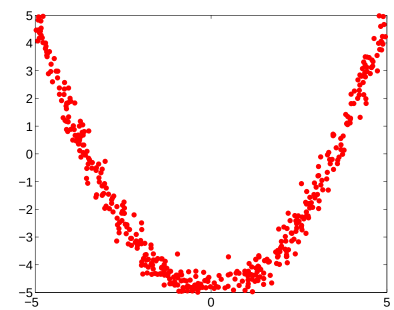
$r = -1$



$r = 0.76$



$r = -0.44$



$r = 0$

Statistika

Povaha statistiky

Nemusím sníst celého vola, abych poznal, že je tuhý.

Samuel Johnson

Statistika: účel a členění

Statistika jako matematická disciplína:

- Zkoumá *společné vlastnosti* velkého počtu obdobných jevů.
- Využívá jen *vybraný vzorek* jevů, nikoli všechny.
- Zabývá se sběrem, prezentací, analýzou a interpretací dat popisujících jevy či vlastnosti pozorovaných objektů.
- Typické úlohy:
  - Odhad parametrů pravděpodobnostního modelu
  - Testování hypotéz

Matematická (teoretická) statistika: výzkum a popis nových metod.

Aplikovaná statistika: použití stat. metod v konkrétních problémech různých oborů, např. například v přírodních či společenských vědách, v politice nebo v lékařství.

Deskriptivní statistika se zabývá numerickým nebo grafickým popisem získaných dat

Inferenční (induktivní) statistika se zabývá *vyhledáváním zákonitostí* v datech naměřených na vzorku jedinců nebo objektů a *zobecňováním* těchto zákonitostí na skupinu, z níž byl vzorek vybrán. Inferenční statistika vychází z počtu pravděpodobnosti.

Základní pojmy

Statistické jednotky: Objekty, jejichž vlastnosti zkoumáme. (Lidé, buňky, hřídele, . . .)

Statistický soubor: Specificky vymezená množina statistických jednotek; o libovolném prvku musíme být schopni rozhodnout, zda do statistického souboru patří či nikoliv. (Např. učitelé FEL, kteří měli na FEL v roce 2013 nadpoloviční úvazek.)

Základní soubor (populace): *Úplný* statistický soubor (soubor všech jednotek). Může být i nekonečný.

Výběrový soubor (výběr, vzorek) rozsahu n: Konečný soubor obsahující jen těch n prvků, které skutečně pozorujeme nebo měříme. Jejich výběr ze základního souboru musí být proveden *náhodně* (s rovnoměrným rozdělením), nebo podle znaku, který se studovanými znaky nesouvisí.

Proč výběr?

1. Omezené zdroje
   Nejsou prostředky nebo čas na zkoumání celé populace.
2. Destruktivní zkoušky
   Můžeme použít *všechny* červené krvinky pacienta, abychom zjistili jejich skutečnou průměrnou velikost?
3. Vzorek bývá přesnější
   Sběr dat pro menší vzorek lze provést menší skupinou lépe proškolených lidí.

Druhy veličin

Znak	**Škála**	**Možné operace**	**Příklady**
Kval.	Nominální	Popsat příslušnost	
	Ordinální	Seřadit	
Kvant.	Intervalová	Porovnat vzdálenosti	
	Poměrová	Porovnat velikosti	

- Spojité vs. diskrétní
- Nezávislé (vstupy) vs. závislé (výstupy)

Druhy veličin

Znak	Škála	Možné operace	Příklady
Kval.	Nominální	Popsat příslušnost	Barva očí, národnost, pohlaví, místo narození
	Ordinální	Seřadit	
Kvant.	Intervalová	Porovnat vzdálenosti	
	Poměrová	Porovnat velikosti	

- Spojité vs. diskrétní
- Nezávislé (vstupy) vs. závislé (výstupy)

Druhy veličin

Znak	Škála	Možné operace	Příklady
Kval.	Nominální	Popsat příslušnost	Barva očí, národnost, pohlaví, místo narození
	Ordinální	Seřadit	Popis velikosti (S,M,L,XL,XXL), vzdělání (ZŠ, SŠ, VŠ)
Kvant.	Intervalová	Porovnat vzdálenosti	
	Poměrová	Porovnat velikosti	

- Spojité vs. diskrétní
- Nezávislé (vstupy) vs. závislé (výstupy)

Druhy veličin

Znak	Škála	Možné operace	Příklady
Kval.	Nominální	Popsat příslušnost	Barva očí, národnost, pohlaví, místo narození
	Ordinální	Seřadit	Popis velikosti (S,M,L,XL,XXL), vzdělání (ZŠ, SŠ, VŠ)
Kvant.	Intervalová	Porovnat vzdálenosti	Kalendářní datum, teplota, úhel
	Poměrová	Porovnat velikosti	

- Spojité vs. diskrétní
- Nezávislé (vstupy) vs. závislé (výstupy)

Druhy veličin

Znak	Škála	Možné operace	Příklady
Kval.	Nominální	Popsat příslušnost	Barva očí, národnost, pohlaví, místo narození
	Ordinální	Seřadit	Popis velikosti (S,M,L,XL,XXL), vzdělání (ZŠ, SŠ, VŠ)
Kvant.	Intervalová	Porovnat vzdálenosti	Kalendářní datum, teplota, úhel
	Poměrová	Porovnat velikosti	Objem prodeje, průměr hřídele, hmotnost, teplota v Kelvinech, úhel vzhledem k . . .

- Spojité vs. diskrétní
- Nezávislé (vstupy) vs. závislé (výstupy)

Šetření vs. experiment

Šetření Průzkumy veřejného mínění, potvrzovací studie prováděné mezi obyvatelstvem...

- Pouze sledujeme, nijak nezasahujeme.
- Nemůžeme ovlivnit rozdělení subjektů do skupin.
- Rozdíl mezi skupinami bývá často ovlivněn tzv. matoucí (confounding) veličinou.
- Částečně lze vyřešit zahrnutím matoucích veličin do modelu; problém je v tom, že nevíme, co může být matoucí veličinou.

Experiment Klinické studie, laboratorní testy

- Aktivně zasahujeme a ovlivňujeme podmínky.
- Snažíme se vyloučit vlivy, které nejsou předmětem výzkumu. Náhodným rozdělením subjektů se tyto vlivy vyruší.
- Zkoumané veličiny aktivně nastavujeme tak, aby vzorek nebyl vychýlený.

Randomizovaný experiment

Vzorek zkoumaných subjektů

- se *náhodně rozdělí* do skupin,
- ke kterým se snažíme *chovat* naprosto *shodným způsobem*.
- Pokud zkoumané subjekty neví, ve které skupině jsou zařazeny (což je obvyklé), mluvíme o **slepém experimentu**.
- Pokud navíc ani lidé, kteří subjekty hodnotí, příp. se o ně starají, neví, do které skupiny jsou subjekty zařazeny, mluvíme o **dvojitě slepém experimentu**.

Randomizované experimenty dávají přesnější představu o sledovaném jevu než výběrová šetření, ale:

Randomizace nemožná, např. souvislost pohlaví s výší platů: pohlaví nelze přiřadit náhodně.

Randomizace možná, ale nepraktická, např. studie kriminality ve městech a na vesnici: lidé se nepřestěhují jen kvůli průzkumu.

Randomizace možná, praktická, přesto se nedělá, např. studie výhod předškolních vzdělávacích programů poskytovaných zdarma, kdy poptávka převyšuje nabídku: náhodné přidělení míst je férový postup, ale lidé odmítají připustit, že generátor náhodných čísel udělá jejich práci lépe než oni.

Etické problémy Experimenty na lidech a na zvířatech ano či ne? Správné otázky by měly znít: *Experimentujeme s rozmyslem nebo hazardujeme? Provádíme experimenty, abychom se z nich dozvěděli maximum, nebo jsou naše experimenty chabé, poskytují špatné informace a poškozují lidi?*

Etika: ilustrace

Jednoho dne naši školu navštívil jeden významný chirurg z Bostonu a udělal nám skvělou přednášku o velké skupině pacientů, na kterých vyzkoušel svou novou metodu vaskulární rekonstrukce. Na konci přednášky se zeptal jeden ze studentů:

- „Měl jste nějakou kontrolní skupinu?"

Etika: ilustrace

Jednoho dne naši školu navštívil jeden významný chirurg z Bostonu a udělal nám skvělou přednášku o velké skupině pacientů, na kterých vyzkoušel svou novou metodu vaskulární rekonstrukce. Na konci přednášky se zeptal jeden ze studentů:

- „Měl jste nějakou kontrolní skupinu?"

Chirurg se postavil, výhružně se opřel o stůl a řekl:

- „Myslíte tím, zda jsem operoval jen polovinu pacientů?"

Etika: ilustrace

Jednoho dne naši školu navštívil jeden významný chirurg z Bostonu a udělal nám skvělou přednášku o velké skupině pacientů, na kterých vyzkoušel svou novou metodu vaskulární rekonstrukce. Na konci přednášky se zeptal jeden ze studentů:

- „Měl jste nějakou kontrolní skupinu?“

Chirurg se postavil, výhružně se opřel o stůl a řekl:

- „Myslíte tím, zda jsem operoval jen polovinu pacientů?“

V posluchárně se rozhostilo ticho. Studentův hlas odpověděl:

- „Ano, to je přesně to, co mám na mysli.“

Lékař praštil pěstí do stolu a zahřměl:

- „Samozřejmě, že ne! Tím bych odsoudil polovinu z nich k smrti!“

Etika: ilustrace

Jednoho dne naši školu navštívil jeden významný chirurg z Bostonu a udělal nám skvělou přednášku o velké skupině pacientů, na kterých vyzkoušel svou novou metodu vaskulární rekonstrukce. Na konci přednášky se zeptal jeden ze studentů:

- „Měl jste nějakou kontrolní skupinu?“

Chirurg se postavil, výhružně se opřel o stůl a řekl:

- „Myslíte tím, zda jsem operoval jen polovinu pacientů?“

V posluchárně se rozhostilo ticho. Studentův hlas odpověděl:

- „Ano, to je přesně to, co mám na mysli.“

Lékař praštil pěstí do stolu a zahřměl:

- „Samozřejmě, že ne! Tím bych odsoudil polovinu z nich k smrti!“

Ovšem pak do ticha opět promluvil studentův hlas:

- „A kterou polovinu?“

Kontrolní skupina

Statistikovi a jeho ženě se narodila dvojčata. Ihned po návratu z porodnice volá muž do kostela a oznamuje tu skvělou zprávu. Kněz má samozřejmě radost:

- „To je skvělé! Tak je co nejdříve přivezte a pokřtíme je!“
- „Ne,“ řekl statistik, „pokřtíme jen jedno. To druhé si necháme jako kontrolní skupinu.“

Statistika

Statistika je

- matematická disciplína ... — to už víme. Ale také
- každá *měřitelná*[1] funkce G definovaná na náhodném výběru libovolného (dostatečného) rozsahu, tj. počítá se z náhodných veličin výběru, a tudíž *sama je náhodnou veličinou*.
- Obvykle se používá jako *odhad parametrů rozdělení* (které nám zůstávají skryty).

Značení:

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

[1] V praxi se setkáte jen s měřitelnými funkcemi. **Měřitelná funkce** G je taková funkce, že pro každé $t \in \mathbb{R}$ je definována pravděpodobnost

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

Náhodný výběr

Náhodný výběr $\boldsymbol{X} = (X_1, \ldots, X_n)$ je vektor náhodných veličin, které jsou *nezávislé* a mají *stejné rozdělení* (independent and identically distributed, i.i.d., IID).

Realizace $\boldsymbol{x} = (x_1, \ldots, x_n)$ náhodného výběru $\boldsymbol{X}$ je výsledkem konkrétního pokusu.

- Popisuje ji *empirické rozdělení*: Vybereme $j \in \{1, \ldots, n\}$ s rovnoměrným rozdělením, výsledkem je x_j.
- Je to diskrétní rozdělení, směs Diracových: $\mathrm{Mix}_{(1/n,\ldots,1/n)}(x_1, \ldots, x_n)$.

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

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

Histogram a empirické rozdělení

V realizaci náhodného výběru $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ *nezáleží na pořadí* hodnot, ale *záleží na jejich četnostech*. Náhodný výběr tak lze popsat

1. množinou (nejvýše n) hodnot $H = \{x_1, \ldots, x_n\}$ a
2. jejich četnostmi n_t, $t \in H$,

které se obvykle prezentují ve formě **tabulky četností** nebo **histogramu**.

Histogram a empirické rozdělení

V realizaci náhodného výběru $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ *nezáleží na pořadí* hodnot, ale *záleží na* jejich *četnostech*. Náhodný výběr tak lze popsat

1. množinou (nejvýše n) hodnot $H = \{x_1, \ldots, x_n\}$ a
2. jejich četnostmi n_t, $t \in H$,

které se obvykle prezentují ve formě **tabulky četností** nebo **histogramu**.

Empirické rozdělení $\mathrm{Emp}(\boldsymbol{x})$, přesněji jeho pstní funkce $p_{\mathrm{Emp}(\boldsymbol{x})}$, vznikne normováním četností: $r_t := \dfrac{n_t}{n} = p_{\mathrm{Emp}(\boldsymbol{x})}(t)$. Jak uvidíme později:

- Obecné momenty empirického rozdělení jsou rovny výběrovým momentům původního rozdělení.

$$\mathrm{E}\left(\mathrm{Emp}(\boldsymbol{x})\right)^k = \sum_{t \in H} t^k \cdot r_t = \frac{1}{n} \sum_{t \in H} t^k \cdot n_t = \frac{1}{n} \sum_{j=1}^{n} x_j^k = m_{\boldsymbol{x}}^k \text{, z čehož plyne } \mathrm{E}\,\mathrm{Emp}(\boldsymbol{x}) = \bar{\boldsymbol{x}}.$$

- Rozptyl empirického rozdělení odpovídá odhadu $\widehat{\sigma_{\boldsymbol{X}}}^2 = \frac{n-1}{n} S_{\boldsymbol{X}}^2$ rozptylu původního rozdělení, ale odlišnému od $S_{\boldsymbol{X}}^2$.

$$\mathrm{D}\,\mathrm{Emp}(\boldsymbol{x}) = \sum_{t \in H} (t - \bar{\boldsymbol{x}})^2 \cdot r_t = \frac{1}{n} \sum_{t \in H} (t - \bar{\boldsymbol{x}})^2 \cdot n_t = \frac{1}{n} \sum_{j=1}^{n} \left(x_j - \bar{\boldsymbol{x}}\right)^2 = \widehat{\sigma_{\boldsymbol{x}}}^2 = \frac{n-1}{n} s_{\boldsymbol{x}}^2.$$

Statistické odhady

Jednou jsem potkal hezkou a milou statističku. Tak jsem ji hned požádal o telefonní číslo.

Ale ona mi dala jenom odhad.

Anonym

Odhady

Žádoucí vlastnosti:

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

Odhady

Žádoucí vlastnosti:

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

Rozlišujeme odhady

- **bodové** (výsledkem je hodnota aproximující skutečnou hodnotu optimálně ve smyslu jistého kritéria) a
- **intervalové** (výsledkem je interval, v němž se skutečná hodnota nachází s danou pravděpodobností).

Výběrový průměr

Výběrový průměr $\bar{X}$ je statistika (náhodná veličina) definovaná jako aritmetický průměr náhodného výběru:

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

Realizace výběrového průměru je rovna aritmetickému průměru realizace náhodného výběru a také střední hodnotě empirického rozdělení:

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

Výběrový průměr

Výběrový průměr $\bar{X}$ je statistika (náhodná veličina) definovaná jako aritmetický průměr náhodného výběru:

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

Realizace výběrového průměru je rovna aritmetickému průměru realizace náhodného výběru a také střední hodnotě empirického rozdělení:

$$\bar{\boldsymbol{x}} = \frac{1}{n} \sum_{j=1}^{n} x_j = \operatorname{E} \operatorname{Emp}(x)$$

Platí:

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

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

$$\sigma_{\bar{\boldsymbol{X}}_n} = \sqrt{\frac{1}{n} \operatorname{D} X} = \frac{1}{\sqrt{n}} \sigma_X$$, pokud existují. (Zde $\operatorname{E} X = \operatorname{E} X_j$ atd.)

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

Centrální limitní věta

Věta: Výběrový prům. z *normálního* rozdělení $\mathrm{N}(\mu, \sigma^2)$ má normální rozdělení $\mathrm{N}(\mu, \frac{1}{n}\sigma^2)$.

Centrální limitní věta

Věta: Výběrový prům. z *normálního* rozdělení $\mathrm{N}(\mu, \sigma^2)$ má normální rozdělení $\mathrm{N}(\mu, \frac{1}{n}\sigma^2)$.

Podobná věta platí i pro jiná rozdělení alespoň asymptoticky.

Centrální limitní věta

Věta: Výběrový prům. z *normálního* rozdělení $\mathrm{N}(\mu, \sigma^2)$ má normální rozdělení $\mathrm{N}(\mu, \frac{1}{n}\sigma^2)$.

Podobná věta platí i pro jiná rozdělení alespoň asymptoticky.

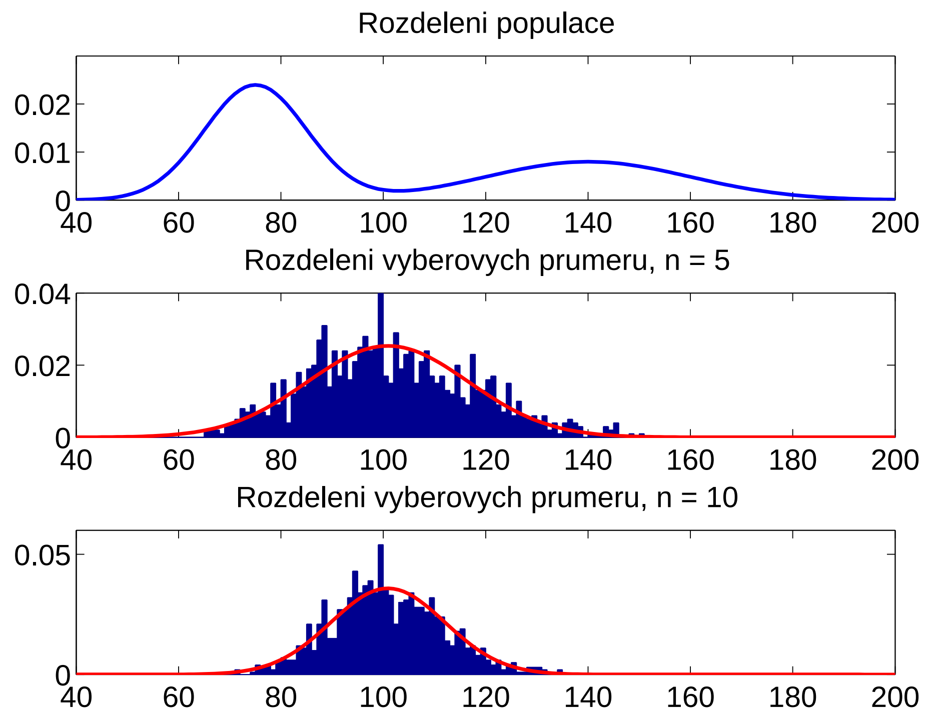
Centrální limitní věta: Nechť $X_j,\ j \in \mathbb{N}$, jsou nezávislé stejně rozdělené náhodné veličiny se střední hodnotou $\mathrm{E}\,X$ a směrodatnou odchylkou $\sigma_X \neq 0$. Pak normované náhodné veličiny

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

konvergují k normovanému normálnímu rozdělení v následujícím smyslu:

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

Ilustrace:



Výběrový rozptyl

Výběrový rozptyl S_X^2 je statistika (náhodná veličina) definovaná jako

$$S_X^2 = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \bar{X}_n)^2$$

a jeho **realizace**:

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

Věta: Výběrový rozptyl je *nestranný* ($\mathrm{E}\, S_X^2 = \mathrm{D}\, X$) *konzistentní* odhad rozptylu původního rozdělení (má-li původní rozdělení rozptyl a 4. centrální moment).

Výběrový rozptyl

Výběrový rozptyl S_X^2 je statistika (náhodná veličina) definovaná jako

$$S_X^2 = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \bar{X}_n)^2$$

a jeho **realizace**:

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

Věta: Výběrový rozptyl je *nestranný* ($\mathrm{E}\, S_X^2 = \mathrm{D}\, X$) *konzistentní* odhad rozptylu původního rozdělení (má-li původní rozdělení rozptyl a 4. centrální moment).

POZOR: Odhad rozptylu pomocí rozptylu empirického rozdělení

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

který je realizací odhadu

$$\widehat{\sigma_X^2} = \frac{1}{n} \sum_{j=1}^{n} (X_j - \bar{\boldsymbol{X}})^2,$$

je *vychýlený* (pouze *asymptoticky nestranný*) odhad rozptylu!

Rozdělení výběrového rozptylu

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

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

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

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

Rozdělení výběrového rozptylu

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

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

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

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

Rozdělení χ^2 s 1 stupněm volnosti, $\chi^2(1)$, je rozdělení náhodné veličiny $V = U^2$, kde U má normované normální rozdělení $\mathrm{N}(0,1)$.

Vlastnosti:

$$\mathrm{E}\, V = \mathrm{E}\, U^2 = \mathrm{D}\, U + (\mathrm{E}\, U)^2 = 1 \qquad (\text{protože } \mathrm{E}\, U = 0, \mathrm{D}\, U = 1),$$

$$\mathrm{D}\, V = 2.$$

Rozdělení χ^2

Rozdělení χ^2 s η stupni volnosti, $\chi^2(\eta)$:

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

Vlastnosti:

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

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

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

Výběrový rozptyl z normálního rozdělení

Pro výběrový rozptyl z *normálního* rozdělení $\mathrm{N}(\mathrm{E}\,X, \mathrm{D}\,X)$ platí:

$$\frac{(n-1)S_X^2}{\mathrm{D}\,X} = \frac{n\widehat{\sigma_X^2}}{\mathrm{D}\,X} \text{ má rozdělení } \chi^2(n-1).$$

Z vlastností rozdělení χ^2 plyne

- pro střední hodnotu výběrového rozptylu

$$\mathrm{E}\,\frac{(n-1)S_X^2}{\mathrm{D}\,X} = n-1 \text{ takže}$$
$$\mathrm{E}\,S_X^2 = \mathrm{D}\,X, \qquad \text{což potvrzuje nestrannost odhadu.}$$

Výběrový rozptyl z normálního rozdělení

Pro výběrový rozptyl z *normálního* rozdělení $\mathrm{N}(\mathrm{E}\,X, \mathrm{D}\,X)$ platí:

$$\frac{(n-1)S_X^2}{\mathrm{D}\,X} = \frac{n\widehat{\sigma_X^2}}{\mathrm{D}\,X} \text{ má rozdělení } \chi^2(n-1).$$

Z vlastností rozdělení χ^2 plyne

- pro střední hodnotu výběrového rozptylu

$$\mathrm{E}\,\frac{(n-1)S_X^2}{\mathrm{D}\,X} = n-1 \text{ takže}$$
$$\mathrm{E}\,S_X^2 = \mathrm{D}\,X, \qquad \text{což potvrzuje nestrannost odhadu.}$$

- pro rozptyl výběrového rozptylu

$$\mathrm{D}\,\frac{(n-1)S_X^2}{\mathrm{D}\,X} = 2(n-1),$$
$$\frac{(n-1)^2\,\mathrm{D}\,S_X^2}{(\mathrm{D}\,X)^2} = 2(n-1), \qquad \text{takže}$$
$$\mathrm{D}\,S_X^2 = \frac{2}{n-1}(\mathrm{D}\,X)^2.$$

Výběrový rozptyl z normálního rozdělení

Pro výběrový rozptyl z *normálního* rozdělení $\mathrm{N}(\mathrm{E}\,X, \mathrm{D}\,X)$ platí:

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

Z vlastností rozdělení χ^2 plyne

- pro střední hodnotu výběrového rozptylu

$$\mathrm{E}\,\frac{(n-1)S_{\boldsymbol{X}}^2}{\mathrm{D}\,X} = n-1 \text{ takže}$$
$$\mathrm{E}\,S_{\boldsymbol{X}}^2 = \mathrm{D}\,X\,, \qquad \text{což potvrzuje nestrannost odhadu.}$$

- pro rozptyl výběrového rozptylu

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

Věta: Pro náhodný výběr $\boldsymbol{X}_n$ z *normálního* rozdělení je $\bar{\boldsymbol{X}}$ nejlepší nestranný odhad střední hodnoty, $S_{\boldsymbol{X}}^2$ je nejlepší nestranný odhad rozptylu a statistiky $\bar{\boldsymbol{X}}$ a $S_{\boldsymbol{X}}^2$ jsou konzistentní a *nezávislé*.

Výběrová směrodatná odchylka

Výběrová směrodatná odchylka S_X je statistika (náhodná veličina) definovaná jako

$$S_X = \sqrt{S_X^2} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2}$$

a její **realizace**:

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

Věta: Výběrová směrodatná ochylka je *vychýleným* ($\mathrm{E}\, S_X \leq \sigma_X$) *konzistentním* odhadem směrodatné odchylky původního rozdělení (má-li původní rozdělení rozptyl a 4. centrální moment).

Důkaz: $\mathrm{D}\, X = \mathrm{E}\, S_X^2 = (\mathrm{E}\, S_X)^2 + \mathrm{D}\, S_X$, a protože $\mathrm{D}\, S_X \geq 0$, tak $\mathrm{D}\, X \geq (\mathrm{E}\, S_X)^2$, takže $\sigma_X \geq \mathrm{E}\, S_X$.

Výběrová směrodatná odchylka

Výběrová směrodatná odchylka S_X je statistika (náhodná veličina) definovaná jako

$$S_X = \sqrt{S_X^2} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2}$$

a její **realizace**:

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

Věta: Výběrová směrodatná ochylka je *vychýleným* ($\mathrm{E}\, S_X \le \sigma_X$) *konzistentním* odhadem směrodatné odchylky původního rozdělení (má-li původní rozdělení rozptyl a 4. centrální moment).

Důkaz: $\mathrm{D}\, X = \mathrm{E}\, S_X^2 = (\mathrm{E}\, S_X)^2 + \mathrm{D}\, S_X$, a protože $\mathrm{D}\, S_X \ge 0$, tak $\mathrm{D}\, X \ge (\mathrm{E}\, S_X)^2$, takže $\sigma_X \ge \mathrm{E}\, S_X$.

POZOR: Odhad směrodatné odchylky pomocí sm. odch. empirického rozdělení

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

je taktéž *vychýlený* odhad směrodatné odchylky!

Výběrový medián

Výběrový medián je statistika (náhodná veličina), která se používá jako odhad mediánu původního rozdělení. Je to 50% kvantil empirického rozdělení, $q_{\mathrm{Emp}(x)}\left(\frac{1}{2}\right)$.

- Je *robustnější* než výběrový průměr (odolnější vůči vlivu odlehlých hodnot).
- Víme, jak se změní po transformaci monotónní funkcí.
- Má vyšší výpočetní náročnost než výběrový průměr: seřazení hodnot má náročnost $\mathcal{O}(n \log n)$, výpočet průměru jen n.
- Má vyšší paměťovou náročnost než výběrový průměr: musíme si pamatovat všech n čísel, u průměru stačí 2 registry.
- Špatně se decentralizuje/paralelizuje.

Míry polohy

1. *Výběrový modus* je nejčastější hodnota ve výběru. Lze jej stanovit i pro celou populaci.

Míry polohy

1. *Výběrový modus* je nejčastější hodnota ve výběru. Lze jej stanovit i pro celou populaci.
2. *Výběrový medián* je hodnota, pod níž (i nad níž) leží 50% hodnot. 50%-ní kvantil, 50. percentil, 5. decil.

Míry polohy

1. *Výběrový modus* je nejčastější hodnota ve výběru. Lze jej stanovit i pro celou populaci.
2. *Výběrový medián* je hodnota, pod níž (i nad níž) leží 50% hodnot. 50%-ní kvantil, 50. percentil, 5. decil.
3. *Výběrový průměr* je takový díl, že naskládáme-li jich za sebe stejný počet, jako je původních hodnot, dostaneme stejný součet, jako dávají původní hodnoty.

Míry polohy

1. *Výběrový modus* je nejčastější hodnota ve výběru. Lze jej stanovit i pro celou populaci.
2. *Výběrový medián* je hodnota, pod níž (i nad níž) leží 50% hodnot. 50%-ní kvantil, 50. percentil, 5. decil.
3. *Výběrový průměr* je takový díl, že naskládáme-li jich za sebe stejný počet, jako je původních hodnot, dostaneme stejný součet, jako dávají původní hodnoty.

Která míra polohy je vhodná? Modus, medián nebo průměr?
Příklad: Rozdělení platů v republice. Je poměrně velká část populace, která nedostává žádný plat (děti, důchodci, nezaměstnaní, lidé, co pracovat nechtějí).

- Modus je

Míry polohy

1. *Výběrový modus* je nejčastější hodnota ve výběru. Lze jej stanovit i pro celou populaci.
2. *Výběrový medián* je hodnota, pod níž (i nad níž) leží 50% hodnot. 50%-ní kvantil, 50. percentil, 5. decil.
3. *Výběrový průměr* je takový díl, že naskládáme-li jich za sebe stejný počet, jako je původních hodnot, dostaneme stejný součet, jako dávají původní hodnoty.

Která míra polohy je vhodná? Modus, medián nebo průměr?
Příklad: Rozdělení platů v republice. Je poměrně velká část populace, která nedostává žádný plat (děti, důchodci, nezaměstnaní, lidé, co pracovat nechtějí).

- Modus je 0. Největší část populace nedostává plat. Nestane-li se s platy něco opravdu převratného, zůstane modus nulový. (Což nemusí platit pro jiné veličiny.)
- Medián je

Míry polohy

1. *Výběrový modus* je nejčastější hodnota ve výběru. Lze jej stanovit i pro celou populaci.
2. *Výběrový medián* je hodnota, pod níž (i nad níž) leží 50% hodnot. 50%-ní kvantil, 50. percentil, 5. decil.
3. *Výběrový průměr* je takový díl, že naskládáme-li jich za sebe stejný počet, jako je původních hodnot, dostaneme stejný součet, jako dávají původní hodnoty.

Která míra polohy je vhodná? Modus, medián nebo průměr?
Příklad: Rozdělení platů v republice. Je poměrně velká část populace, která nedostává žádný plat (děti, důchodci, nezaměstnaní, lidé, co pracovat nechtějí).

- Modus je 0. Největší část populace nedostává plat. Nestane-li se s platy něco opravdu převratného, zůstane modus nulový. (Což nemusí platit pro jiné veličiny.)
- Medián je asi nejlepší ukazatel toho, co si vydělá typický občan. I když se vysoké platy 10x zvýší, zůstane stejný.
- Průměr je

Míry polohy

1. *Výběrový modus* je nejčastější hodnota ve výběru. Lze jej stanovit i pro celou populaci.
2. *Výběrový medián* je hodnota, pod níž (i nad níž) leží 50% hodnot. 50%-ní kvantil, 50. percentil, 5. decil.
3. *Výběrový průměr* je takový díl, že naskládáme-li jich za sebe stejný počet, jako je původních hodnot, dostaneme stejný součet, jako dávají původní hodnoty.

Která míra polohy je vhodná? Modus, medián nebo průměr?
Příklad: Rozdělení platů v republice. Je poměrně velká část populace, která nedostává žádný plat (děti, důchodci, nezaměstnaní, lidé, co pracovat nechtějí).

- Modus je 0. Největší část populace nedostává plat. Nestane-li se s platy něco opravdu převratného, zůstane modus nulový. (Což nemusí platit pro jiné veličiny.)
- Medián je asi nejlepší ukazatel toho, co si vydělá typický občan. I když se vysoké platy 10x zvýší, zůstane stejný.
- Průměr je dobrá míra pro ekonomiku a daňové úřady, protože z něj mohou spočítat celkový plat. Není to ale dobrá míra typického platu, je snadno ovlivnitelný (především vysokými platy).

Míry polohy

1. *Výběrový modus* je nejčastější hodnota ve výběru. Lze jej stanovit i pro celou populaci.
2. *Výběrový medián* je hodnota, pod níž (i nad níž) leží 50% hodnot. 50%-ní kvantil, 50. percentil, 5. decil.
3. *Výběrový průměr* je takový díl, že naskládáme-li jich za sebe stejný počet, jako je původních hodnot, dostaneme stejný součet, jako dávají původní hodnoty.

Která míra polohy je vhodná? Modus, medián nebo průměr?
Příklad: Rozdělení platů v republice. Je poměrně velká část populace, která nedostává žádný plat (děti, důchodci, nezaměstnaní, lidé, co pracovat nechtějí).

- Modus je 0. Největší část populace nedostává plat. Nestane-li se s platy něco opravdu převratného, zůstane modus nulový. (Což nemusí platit pro jiné veličiny.)
- Medián je asi nejlepší ukazatel toho, co si vydělá typický občan. I když se vysoké platy 10x zvýší, zůstane stejný.
- Průměr je dobrá míra pro ekonomiku a daňové úřady, protože z něj mohou spočítat celkový plat. Není to ale dobrá míra typického platu, je snadno ovlivnitelný (především vysokými platy).

Teď už přesně víme, co je průměr a co je medián. Takže nás vůbec nepřekvapí, že:

- Naprostá většina lidí má nadprůměrný počet nohou. *Je to tak?*

Míry polohy

1. *Výběrový modus* je nejčastější hodnota ve výběru. Lze jej stanovit i pro celou populaci.
2. *Výběrový medián* je hodnota, pod níž (i nad níž) leží 50% hodnot. 50%-ní kvantil, 50. percentil, 5. decil.
3. *Výběrový průměr* je takový díl, že naskládáme-li jich za sebe stejný počet, jako je původních hodnot, dostaneme stejný součet, jako dávají původní hodnoty.

Která míra polohy je vhodná? Modus, medián nebo průměr?
Příklad: Rozdělení platů v republice. Je poměrně velká část populace, která nedostává žádný plat (děti, důchodci, nezaměstnaní, lidé, co pracovat nechtějí).

- Modus je 0. Největší část populace nedostává plat. Nestane-li se s platy něco opravdu převratného, zůstane modus nulový. (Což nemusí platit pro jiné veličiny.)
- Medián je asi nejlepší ukazatel toho, co si vydělá typický občan. I když se vysoké platy 10x zvýší, zůstane stejný.
- Průměr je dobrá míra pro ekonomiku a daňové úřady, protože z něj mohou spočítat celkový plat. Není to ale dobrá míra typického platu, je snadno ovlivnitelný (především vysokými platy).

Teď už přesně víme, co je průměr a co je medián. Takže nás vůbec nepřekvapí, že:

- Naprostá většina lidí má nadprůměrný počet nohou. *Je to tak?*
- Polovina populace má inteligenci nižší, než medián. Většina populace má ovšem nadprůměrnou inteligenci. To je jen důsledek toho, že lidská inteligence je shora omezená. Pro lidskou hloupost ovšem žádné limity neexistují. *Jak by muselo vypadat rozdělení inteligence v populaci, aby to pravda byla?*

Míry polohy

1. *Výběrový modus* je nejčastější hodnota ve výběru. Lze jej stanovit i pro celou populaci.
2. *Výběrový medián* je hodnota, pod níž (i nad níž) leží 50% hodnot. 50%-ní kvantil, 50. percentil, 5. decil.
3. *Výběrový průměr* je takový díl, že naskládáme-li jich za sebe stejný počet, jako je původních hodnot, dostaneme stejný součet, jako dávají původní hodnoty.

Která míra polohy je vhodná? Modus, medián nebo průměr?
Příklad: Rozdělení platů v republice. Je poměrně velká část populace, která nedostává žádný plat (děti, důchodci, nezaměstnaní, lidé, co pracovat nechtějí).

- Modus je 0. Největší část populace nedostává plat. Nestane-li se s platy něco opravdu převratného, zůstane modus nulový. (Což nemusí platit pro jiné veličiny.)
- Medián je asi nejlepší ukazatel toho, co si vydělá typický občan. I když se vysoké platy 10x zvýší, zůstane stejný.
- Průměr je dobrá míra pro ekonomiku a daňové úřady, protože z něj mohou spočítat celkový plat. Není to ale dobrá míra typického platu, je snadno ovlivnitelný (především vysokými platy).

Teď už přesně víme, co je průměr a co je medián. Takže nás vůbec nepřekvapí, že:

- Naprostá většina lidí má nadprůměrný počet nohou. *Je to tak?*
- Polovina populace má inteligenci nižší, než medián. Většina populace má ovšem nadprůměrnou inteligenci. To je jen důsledek toho, že lidská inteligence je shora omezená. Pro lidskou hloupost ovšem žádné limity neexistují. *Jak by muselo vypadat rozdělení inteligence v populaci, aby to pravda byla?*

Už jste slyšeli o tom politikovi, který ve své předvolební kampani sliboval, že se po svém zvolení zasadí o to, aby měl každý občan nadprůměrný příjem?

Místo závěru

Existují tři druhy lží: lži, naprosté lži a statistiky.

Benjamin Disraeli