NMAI059 Pravděpodobnost a statistika 1
1. přednáška

Robert Šámal

Přehled

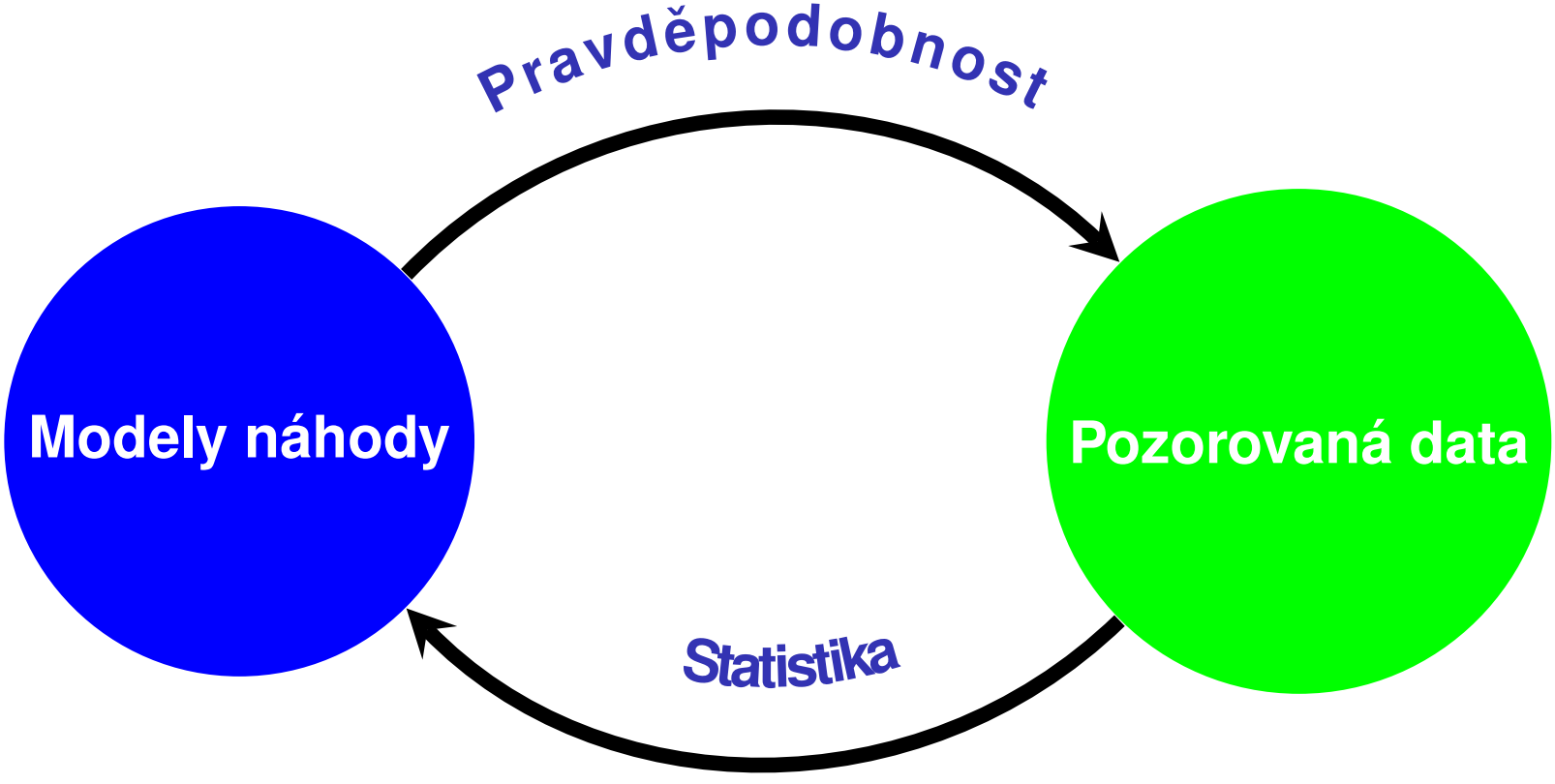
Organizace přednášky

- Přednášky v Zoomu, dokud to bude potřeba (patrně celý semestr). Přednáška má svoji stránku v Moodle (odkaz v SISu). Tam bude všechno.
- Nedojde-li k technickým komplikacím, bude video přednášky dostupné (po přihlášení do SISu).
- Kdyby vám vadilo být nahráni, můžete vypnout svoji kameru, případně dotazy klást v chatu.
- Budu ale rád, pokud si kameru zapnete, abych viděl, jak pomalu/rychle mluvím, co vás překvapilo, atd.
- Používejte též funkce Zoomu – přihlásit se, zpomalit-zrychlit, atd.
- Během přednášky budeme používat krátké ankety.
- Pdf verze „tabule“ bude též k dispozici – už před přednáškou.
- Zkouška bude v ideálním případě prezenční písemka s možností ústního dozkoušení.
- V Moodlu je také prostor pro diskuzi, jak (ne)funguje technologie. Případně se ozvěte emailem.

Organizace cvičení

- Detaily vám sdělí cvičící

Plán přednášky



Přehled

Aplikace na rozehřátí

Příklad

Dány dva polynomy $f(x), g(x)$ stupně d. Chceme zjistit, zda jsou stejné, a to co nejrychleji.

Pravděpodobnost – intuice, definice

Některé jevy neumíme/nechceme popsat kauzálně:

- hod kostkou
- tři hody kostkou, nekonečně mnoho hodů kostkou
- hod šipkou na terč
- počet emailů za den
- dobu běhu programu (v reálném počítači)

Důvody:

- fyzikální vlastnost přírody?
- komplikovaný proces (počasí, medicína, molekuly plynu)
- neznámé vlivy (působení dalších lidí, programů, . . .)
- randomizované algoritmy (test prvočíselnosti, quicksort)
- náhodné grafy (odhady Ramseyových čísel)

Pro popis pomocí teorie pravděpodobnosti napřed vybereme *množinu elementárních jevů (sample space)* Ω.

Prostor jevů

Dále vybereme *prostor jevů (event space)* $\mathcal{F} \subseteq \mathcal{P}(\Omega)$, u kterých budeme měřit jejich pravděpodobnost.
Často $\mathcal{F} = \mathcal{P}(\Omega)$, to je možné vždy, když Ω je spočetná. Ale např. pro $\Omega = \mathbb{R}$ to už nejde.

Definice

$\mathcal{F} \subseteq \mathcal{P}(\Omega)$ je prostor jevů (též σ-algebra), pokud

- $\emptyset \in \mathcal{F}$ *a* $\Omega \in \mathcal{F}$,
- $A \in \mathcal{F} \Rightarrow \Omega \setminus A \in \mathcal{F}$, *a*
- $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Axiomy pravděpodobnosti

Definice

$P : \mathcal{F} \to [0, 1]$ se nazývá pravděpodobnost (probability), pokud

- $P(\emptyset) = 0$, $P(\Omega) = 1$, *a*
- $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$, *pro libovolnou posloupnost po dvou disjunktních jevů* $A_1, A_2, \ldots \in \mathcal{F}$.

Definice

Pravděpodobnostní prostor (probability space) je trojice $(\Omega, \mathcal{F}, P)$ *taková, že*

- $\Omega \neq \emptyset$ *je libovolná množina,*
- $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ *je prostor jevů, a*
- P *je pravděpodobnost.*

Názvosloví

- *Šance (odds)* jevu A je $O(A) = \frac{P(A)}{P(A^c)}$. Např. šance na výhru je 1 ku 2 znamená, že pravděpodobnost výhry je $1/3$; šance, že na kostce padne šestka je 1 ku 5.

- „A je *jistý jev*“ znamená $P(A) = 1$. Také se říká, že A nastává *skoro jistě (almost surely)*, zkráceně s.j. (a.s.).
- „A je *nemožný jev*“ znamená $P(A) = 0$.

Základní vlastnosti

Věta

V pravděpodobnostním prostoru $(\Omega, \mathcal{F}, P)$ *platí pro* $A, B \in \mathcal{F}$

1. $P(A) + P(A^c) = 1 \qquad (A^c = \Omega \setminus A)$
2. $A \subseteq B \Rightarrow P(A) \leq P(B)$
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
4. $P(A_1 \cup A_2 \cup \dots) \leq \sum_i P(A_i)$ *(subaditivita, Booleova nerovnost)*

Příklady pravděpodobnostních prostorů 1

- **Konečný s uniformní pravděpodobností**
  Ω je libovolná konečná množina, $\mathcal{F} = \mathcal{P}(\Omega)$,
  $P(A) = |A|/|\Omega|$.
- **Diskrétní**
  $\Omega = \{\omega_1, \omega_2, \dots\}$ je libovolná spočetná množina. Jsou dány $p_1, p_2, \dots \in [0, 1]$ se součtem 1.
  $P(A) = \sum_{i:\omega_i \in A} p_i$

Příklady pravděpodobnostních prostorů 2

- **Spojitý**

  $\Omega \subseteq R^d$ pro vhodné d (Ω např. uzavřená nebo otevřená)
  $\mathcal{F}$ vhodná (obsahuje např. všechny otevřené množiny)
  $f : \Omega \to [0, 1]$ je funkce taková, že $\int_\Omega f(x)dx = 1$.
  $P(A) = \int_A f(x)dx$

  Spec. případ: $f(x) = 1/V_d(\Omega)$
  $P(A) = V_d(A)/V_d(\Omega)$,
  kde $V_d(A) = \int_A 1$ je d-rozměrný objem A.

- **Bernoulliho krychle – nekonečné opakování**

  $\Omega = S^{\mathbb{N}}$, kde S je diskrétní s pstí Q,
  $\mathcal{F}$ vhodná (obsahuje např. všechny množiny tvaru
  $A = A_1 \times \cdots \times A_k \times S \times S \times \cdots$
  $P(A) = Q(A_1) \cdots Q(A_k)$

  Př.: $\{0, 1\}^{\mathbb{N}}$ nekonečné házení mincí

Nepříklady

- **Náhodné přirozené číslo** můžeme vybrat mnoha způsoby. V přednášce poznáme geometrické a Poissonovo rozdělení. Nemůžeme ale požadovat, aby všechna přirozená čísla měla stejnou pravděpodobnost. (Proč?) „Náhodné přirozené číslo je sudé s pravd. $1/2$.“ ???
- **Náhodné reálné číslo** Opět není žádný preferovaný způsob, jak definovat pravděpodobnost pro $\Omega = \mathbb{R}$. Typicky bude každé reálné číslo mít pravděpodobnost 0! Navíc nejde definovat pravděpodobnost tak, aby nezáležela na posunu, tj. $P([0,1]) = P([1,2]) = \dots$
- **Náhodná tětiva kružnice – Bertrandův paradox** Vybereme náhodnou tětivu zadané kružnice. Jaká je pravděpodobnost, že její délka je větší, než strana vepsaného rovnostranného trojúhelníku?

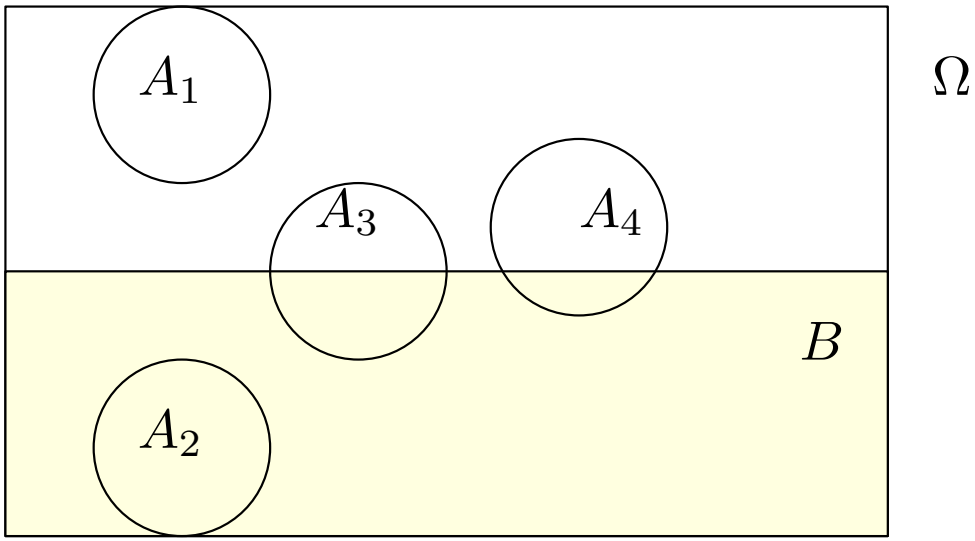
Přehled

Podmíněná pravděpodobnost

Definice

Pokud $A, B \in \mathcal{F}$ a $P(B) > 0$, pak definujeme podmíněnou pravděpodobnost A při B (probability of A given B) jako

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



- $Q(A) := P(A \mid B)$. Pak $(\Omega, \mathcal{F}, Q)$ je pravděpodobnostní prostor.

Zřetězené podmiňování

- $P(A \cap B) = P(B)P(A \mid B)$

Věta

Pokud $A_1, \ldots, A_n \in \mathcal{F}$ *a* $P(A_1 \cap \cdots \cap A_n) > 0$, *tak*

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) =$$

$$P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \ldots P(A_n \mid \bigcap_{i=1}^{n-1} A_i)$$

Rozbor všech možností

Definice

Spočetný systém množin $B_1, B_2, \ldots \in \mathcal{F}$ je rozklad (partition) Ω, pokud

- $B_i \cap B_j = \emptyset$ *pro* $i \neq j$ *a*
- $\bigcup_i B_i = \Omega$.

Věta

Pokud $B_1, B_2, \ldots$ je rozklad Ω a $A \in \mathcal{F}$, tak

$$P(A) = \sum_i P(A \mid B_i) P(B_i)$$

(sčítance s $P(B_i) = 0$ považujeme za 0).

Rozbor všech možností

Bayesova věta

Věta

Pokud $B_1, B_2, \ldots$ *je rozklad* Ω, $A \in \mathcal{F}$ *a* $P(A), P(B_j) > 0$, *tak*

$$P(B_j \mid A) = \frac{P(A \mid B_j)P(B_j)}{\sum_i P(A \mid B_i)P(B_i)}.$$

(sčítance s $P(B_i) = 0$ *považujeme za 0).*

Bayesova věta

Nezávislost jevů

Definice

Jevy $A, B \in \mathcal{F}$ jsou nezávislé (independent) pokud $P(A \cap B) = P(A)P(B)$.

- Pak také $P(A \mid B) = P(A)$, pokud $P(B) > 0$.

Nezávislost více jevů

Definice

Jevy $\{A_i : i \in I\}$ jsou (vzájemně) nezávislé, pokud pro každou konečnou množinu $J \subseteq I$

$$P(\bigcap_{i \in J} A_i) = \prod_{i \in J} P(A_i).$$

Pokud podmínka platí jen pro dvouprvkové množiny J, nazýváme jevy $\{A_i\}$ po dvou nezávislé (pairwise independent).

Spojitost pravděpodobnosti

Věta

Nechť pro množiny z prostoru jevů platí

$$A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$$

a $A = \cup_{i=1}^{\infty} A_i$. *Pak platí*

$$P(A) = \lim_{i \to \infty} P(A_i).$$

- $A_n \subset \{P, O\}^{\mathbb{N}}$, $A_n =$ mezi prvními n hody padl aspoň jednou orel.

Přehled

Borel-Cantelliho lemma

Věta

Nechť jevy $A_1, A_2, \ldots$ splňují $P(A_i) = p_i > 0$ pro každé i. Označme Nic jev „nenastal žádný z jevů $\{A_i\}$“ a Inf jev „nastalo nekonečně mnoho z jevů $\{A_i\}$“.

1. *Pokud $\sum_i p_i < \infty$, tak $P(Inf) = 0$.*
2. *Pokud jsou jevy $A_1, A_2, \ldots$ nezávislé a $\sum_i p_i = \infty$, tak $P(Nic) = 0$, $P(Inf) = 1$.*

NMAI059 Pravděpodobnost a statistika 1
2. přednáška

Robert Šámal

Co už víme

- definice pravděpodobnostního prostoru $(\Omega, \mathcal{F}, P)$: dva axiomy
- **naivní** pravděpodobnostní prostor: Ω konečná, $\mathcal{F} = \mathcal{P}(\Omega)$, $P(A) := |A|/|\Omega|$
- **diskrétní** pravděpodobnostní prostor: $\Omega = \{\omega_1, \omega_2, \dots\}$, $\mathcal{F} = \mathcal{P}(\Omega)$, $\sum_i p_i = 1$
  $P(A) := \sum_{i:\omega_i \in A} p_i$
- **geometrický** pravděpodobnostní prostor: $\Omega \subseteq \mathbb{R}^d$ s konečným objemem, $P(A) := V_d(A)/V_d(\Omega)$
- pravděpodobnostní prostor **spojitý s hustotou**: $\Omega \subseteq \mathbb{R}^d$ s funkcí f, kde $\int_\Omega f = 1$, $P(A) := \int_A f$

Co už víme: Základní vlastnosti

V pravděpodobnostním prostoru $(\Omega, \mathcal{F}, P)$ platí pro $A, B \in \mathcal{F}$

- $P(A^c) = 1 - P(A) \qquad (A^c = \Omega \setminus A)$
- $A \subseteq B \Rightarrow P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A_1 \cup A_2 \cup \dots) \leq \sum_i P(A_i)$ (subaditivita, Booleova nerovnost)

- Definujeme podmíněnou pravděpodobnost (pro $P(B) > 0$).

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

- $Q(A) = P(A \mid B)$ splňuje axiomy pro pravděpodobnost

Přehled

Zřetězené podmiňování

- $P(A \cap B) = P(B)P(A \mid B)$

Věta

Pokud $A_1, \ldots, A_n \in \mathcal{F}$ *a* $P(A_1 \cap \cdots \cap A_n) > 0$*, tak*

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) =$$

$$P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \ldots P(A_n \mid \bigcap_{i=1}^{n-1} A_i)$$

- Příklad: vytáhneme 3 karty z balíčku 52 karet. Jaká je P(žádné srdce)?

Věta o úplné pravd. = Rozbor všech možností

Definice

Spočetný systém množin $B_1, B_2, \ldots \in \mathcal{F}$ je rozklad (partition) Ω, pokud

- $B_i \cap B_j = \emptyset$ *pro* $i \neq j$ *a*
- $\bigcup_i B_i = \Omega$.

Věta

Pokud $B_1, B_2, \ldots$ je rozklad Ω a $A \in \mathcal{F}$, tak

$$P(A) = \sum_i P(B_i) P(A \mid B_i)$$

(sčítance s $P(B_i) = 0$ považujeme za 0).

Věta o úplné pravd. = Rozbor všech možností

- 1. aplikace. Máme tři mince: P+O, P+P, O+O. Jaká je pravděpodobnost, že padne orel?

Věta o úplné pravd. = Rozbor všech možností

- 2. aplikace. Gambler's ruin – zbankrotování hazardního hráče.
  Máme a korun, náš protihráč b korun. Hrajeme opakovaně spravedlivou hru o 1 Kč, dokud někdo nepřijde o všechny peníze. Jaká je pravděpodobnost, že vyhrajeme?

Bayesova věta

Věta

Pokud $B_1, B_2, \ldots$ *je rozklad* Ω, $A \in \mathcal{F}$, $P(A) > 0$ *a* $P(B_j) > 0$, *tak*

$$P(B_j \mid A) = \frac{P(B_j)P(A \mid B_j)}{P(A)} = \frac{P(B_j)P(A \mid B_j)}{\sum_i P(B_i)P(A \mid B_i)}.$$

(sčítance s $P(B_i) = 0$ *považujeme za 0).*

Bayesova věta

Nezávislost jevů

Definice

Jevy $A, B \in \mathcal{F}$ jsou nezávislé (independent) pokud $P(A \cap B) = P(A)P(B)$.

- Pak také $P(A \mid B) = P(A)$, pokud $P(B) > 0$.

Příklad: Hodíme dvakrát mincí. Označme

- $A = \{\omega \in \Omega : \omega_1 = P\} =$ „poprvé padla panna“
- $B = \{\omega \in \Omega : \omega_2 = P\} =$ „podruhé padla panna“
- $C = \{\omega \in \Omega : \omega_1 \neq \omega_2\} =$ „padla právě jedna panna“

Nezávislost více jevů

Definice

Jevy $\{A_i : i \in I\}$ jsou (vzájemně) nezávislé, pokud pro každou konečnou množinu $J \subseteq I$

$$P(\bigcap_{i \in J} A_i) = \prod_{i \in J} P(A_i).$$

Pokud podmínka platí jen pro dvouprvkové množiny J, nazýváme jevy $\{A_i\}$ po dvou nezávislé (pairwise independent).

Spojitost pravděpodobnosti

Věta

Nechť pro množiny z prostoru jevů platí

$$A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$$

a $A = \cup_{i=1}^{\infty} A_i$*. Pak platí*

$$P(A) = \lim_{i \to \infty} P(A_i).$$

- $A_n \subset \{P, O\}^{\mathbb{N}}$, $A_n =$ mezi prvními n hody padl aspoň jednou orel.

Přehled

Náhodná veličina/proměnná

Často nás zajímá číslo dané výsledkem náhodného pokusu.

- Hodíme na terč a změříme vzdálenost od středu.
- Házíme kostkou, dokud nepadne šestka, ale pak si všimneme jenom toho, kolik hodů to trvalo.
- U quicksortu (algoritmus na třídění) měříme počet kroků (v závislosti na náhodné volbě pivotu).

Definice

Mějme pravděpodobnostní prostor $(\Omega, \mathcal{F}, P)$*. Funkci* $X : \Omega \to \mathbb{R}$ *nazveme diskrétní náhodná veličina (discrete random variable), pokud* $Im(X)$ *(obor hodnot* X*) je spočetná množina a pokud pro všechna reálná* x *platí*

$$\{\omega \in \Omega : X(\omega) = x\} \in \mathcal{F}.$$

Pravděpodobnostní funkce

Definice

Pravděpodobnostní funkce (probability mass function, pmf) diskrétní náhodné veličiny X je funkce $p_X : \mathbb{R} \to [0,1]$ taková, že

$$p_X(x) = P(X = x) = P(\{\omega \in \Omega : X(\omega) = x\})$$

- $\sum_{x \in Im(X)} p_X(x) = ?$
- $S := Im(X) \qquad Q(A) := \sum_{x \in A} p_X(x)$
  $(S, \mathcal{P}(S), Q)$ je diskrétní pravděpodobnostní prostor.
- Pro $S = \{s_i : i \in I\}$ spočetnou množinu reálných čísel a $c_i \in [0,1]$ splňující $\sum_{i \in I} c_i = 1$ existuje pravděpodobnostní prostor a diskrétní n.v. X na něm taková, že $p_X(s_i) = c_i$ pro $i \in I$.

Přehled

Bernoulliho/alternativní rozdělení

- $X =$ počet orlů při jednom hodu nespravedlivou mincí.
- Značíme $X \sim Bern(p)$. (Někdy se značí $Alt(p)$.)

- Dáno $p \in [0, 1]$.
- $p_X(1) = p$
- $p_X(0) = 1 - p$
- $p_X(k) = 0$ pro $k \neq 0, 1$

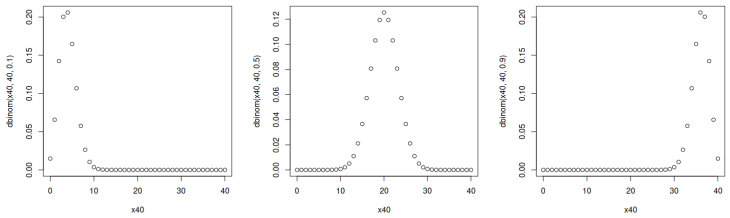
- Pro libovolný jev $A \in \mathcal{F}$ definujeme *indikátorovou n.v.* I_A:
- $I_A(\omega) = 1$ pokud $\omega \in A$, $I_A(\omega) = 0$ jinak.
- $I_A \sim Bern(P(A))$

Binomiální rozdělení

- $X =$ počet orlů při n hodech nespravedlivou mincí.
- Značíme $X \sim Bin(n, p)$.

- Dáno $p \in [0, 1]$.
- $p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$ pro $k \in \{0, 1, \ldots, n\}$

Binomiální rozdělení: pravděpodobnostní funkce



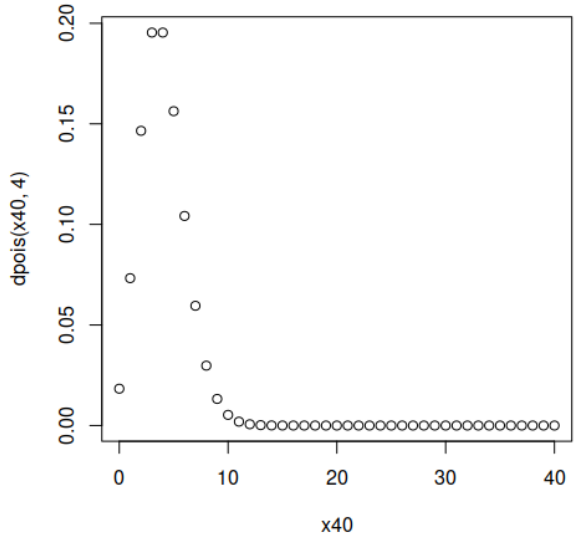
Vygenerováno následujícím kódem v R

```
x40 <- 0:40
plot(x40, dbinom(x40, 40, 0.1))
plot(x40, dbinom(x40, 40, 0.5))
plot(x40, dbinom(x40, 40, 0.9))
```

Poissonovo rozdělení

- Značíme $X \sim Pois(\lambda)$.
- Dáno reálné $\lambda > 0$.
- $p_X(k) = \frac{\lambda^k}{k!} e^{-\lambda}$
- $Pois(\lambda)$ je limitou $Bin(n, \lambda/n)$
- X popisuje např. počet emailů, které dostaneme za jednu hodinu.

Poissonovo rozdělení: pravděpodobnostní funkce



Vygenerováno následujícím kódem v R

```
x40 <- seq(0, 40,by=1)
plot(x40, dpois(x40, 4))
```

Poissonovo paradigma

- $A_1, \ldots, A_n$ jsou (skoro-)nezávislé jevy s $P(A_i) = p_i$, $\lambda = \sum_i p_i$. Nechť n je velké, každé z p_i malé. Pak přibližně platí

$$\sum_{i=1}^{n} I_{A_i} \sim Pois(\lambda).$$

Geometrické rozdělení

- $X =$ kolikátým hodem mincí padl první orel.
- Značíme $X \sim Geom(p)$.

- Dáno $p \in [0, 1]$.
- $p_X(k) = (1 - p)^{k-1} p$

- Někdy se tomuto rozdělení říká posunuté geometrické, a za normální geometrické se považuje rozdělení $X - 1$, tj. počet neúspěšných hodů.

Přehled

Střední hodnota

Definice

Pokud X je diskrétní n.v., tak její střední hodnota (expectation) je označována $\mathbb{E}(X)$ a definována

$$\mathbb{E}(X) = \sum_{x \in Im(X)} x \cdot P(X = x),$$

pokud součet má smysl.

LOTUS

- Pro reálnou funkci g a diskrétní n.v. X je $Y = g(X)$ také diskrétní n.v.

Věta (LOTUS)

Pokud X je diskrétní n.v. a g reálná funkce, tak

$$\mathbb{E}(g(X)) = \sum_{x \in Im(X)} g(x) P(X = x)$$

pokud součet má smysl.

Vlastnosti $\mathbb{E}$

Věta

Nechť X, Y jsou diskrétní n.v. a $a, b \in \mathbb{R}$.

1. *Pokud $P(X \geq 0) = 1$ a $\mathbb{E}(X) = 0$, tak $P(X = 0) = 1$.*
2. *Pokud $\mathbb{E}(X) \geq 0$ tak $P(X \geq 0) > 0$.*
3. $\mathbb{E}(a \cdot X + b) = a \cdot \mathbb{E}(X) + b$.
4. $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.

Podmíněná střední hodnota

Definice

Pokud X je diskrétní n.v. a $P(B) > 0$, tak podmíněná střední hodnota X za předpokladu B (conditional expectation of X given B) je

$$\mathbb{E}(X \mid B) = \sum_{x \in Im(X)} x \cdot P(X = x \mid B),$$

pokud součet má smysl.

Rozbor všech možností

Věta

Pokud $B_1, B_2, \ldots$ je rozklad Ω a $A \in \mathcal{F}$, tak

$$\mathbb{E}(X) = \sum_i \mathbb{E}(X \mid B_i) P(B_i),$$

kdykoliv má součet smysl. (Sčítance s $P(B_i) = 0$ považujeme za 0.)

Rozbor všech možností

Přehled

Bertrand's paradox

Simpsons's paradox

NMAI059 Pravděpodobnost a statistika 1
3. přednáška

Robert Šámal

Přehled

Co už víme

Definice

Mějme pravděpodobnostní prostor $(\Omega, \mathcal{F}, P)$*. Funkci* $X : \Omega \to \mathbb{R}$ *nazveme diskrétní náhodná veličina (discrete random variable), pokud* $Im(X)$ *(obor hodnot* X*) je spočetná množina a pokud pro všechna reálná* x *platí* $\{\omega \in \Omega : X(\omega) = x\} \in \mathcal{F}$.

Definice

Pravděpodobnostní funkce (probability mass function, pmf) diskrétní náhodné veličiny X *je funkce* $p_X : \mathbb{R} \to [0, 1]$ *taková, že*

$$p_X(x) = P(X = x) = P(\{\omega \in \Omega : X(\omega) = x\})$$

- $\sum_{x \in Im(X)} p_X(x) = 1$
- a to je jediné omezení
- X určuje diskrétní pravděpodobnostní prostor na $Im(X)$

Jiný popis – distribuční funkce

Definice

Distribuční funkce (cumulative distribution function, CDF) n.v. X je funkce

$$F_X(x) := P(X \leq x) = P(\{\omega \in \Omega : X(\omega) \leq x).$$

- F_X je neklesající funkce
- $\lim_{x \to -\infty} F_X(x) = 0$
- $\lim_{x \to +\infty} F_X(x) = 1$
- F_X je zprava spojitá

Přehled

Bernoulliho/alternativní rozdělení

- $X =$ počet orlů při jednom hodu nespravedlivou mincí.
- Značíme $X \sim Bern(p)$. (Někdy se značí $Alt(p)$.)

- Dáno $p \in [0, 1]$.
- $p_X(1) = p$
- $p_X(0) = 1 - p$
- $p_X(k) = 0$ pro $k \neq 0, 1$

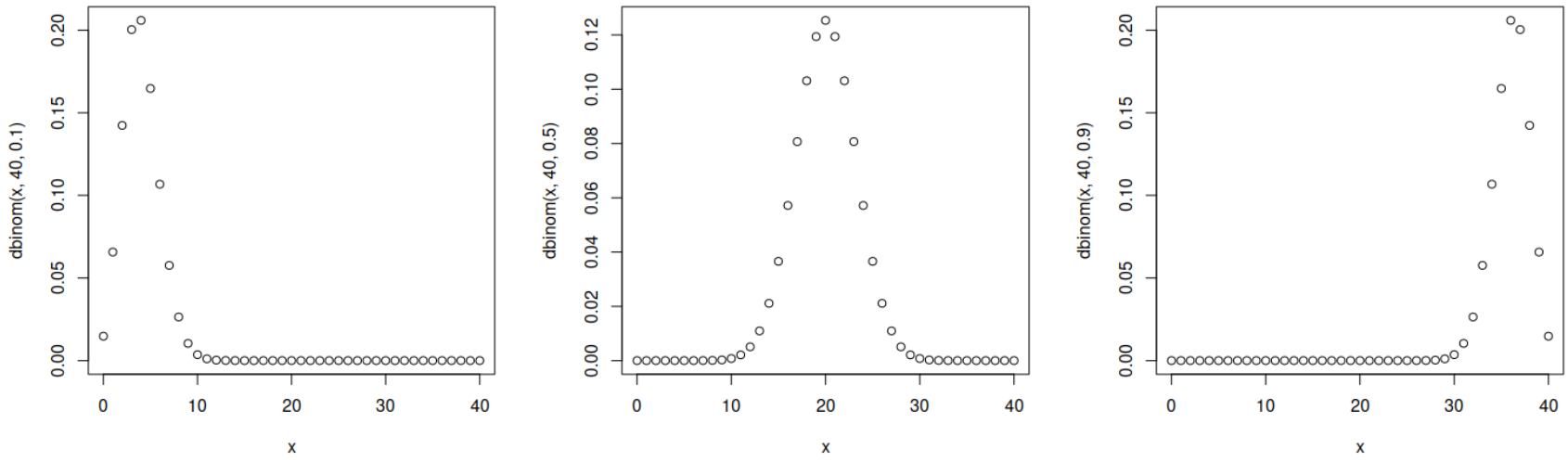
- Pro libovolný jev $A \in \mathcal{F}$ definujeme *indikátorovou n.v.* I_A:
- $I_A(\omega) = 1$ pokud $\omega \in A$, $I_A(\omega) = 0$ jinak.
- $I_A \sim Bern(P(A))$

Binomiální rozdělení

- $X =$ počet orlů při n hodech nespravedlivou mincí.
- Dáno $p \in [0, 1]$ – pravděpodobnost orla při jednom hodu.
- Značíme $X \sim Bin(n, p)$.

- $X = \sum_{i=1}^{n} X_i$ pro nezávislé n.v. $X_1, \ldots, X_n \sim Bern(p)$.
- $p_X(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ pro $k \in \{0, 1, \ldots, n\}$

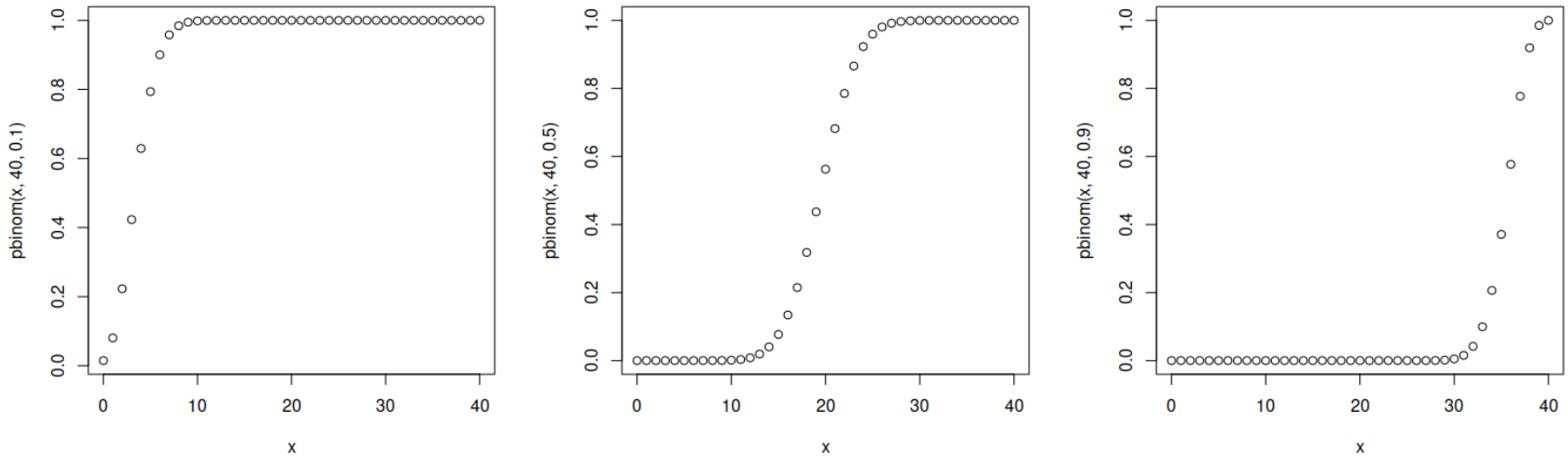
Binomiální rozdělení: pravděpodobnostní funkce



Vygenerováno následujícím kódem v R

```
x <- 0:40
plot(x,dbinom(x,40,0.1))
plot(x,dbinom(x,40,0.5))
plot(x,dbinom(x,40,0.9))
```

Binomiální rozdělení: distribuční funkce



Vygenerováno následujícím kódem v R

```
x <- 0:40
plot(x,pbinom(x,40,0.1))
plot(x,pbinom(x,40,0.5))
plot(x,pbinom(x,40,0.9))
```

Hypergeometrické rozdělení

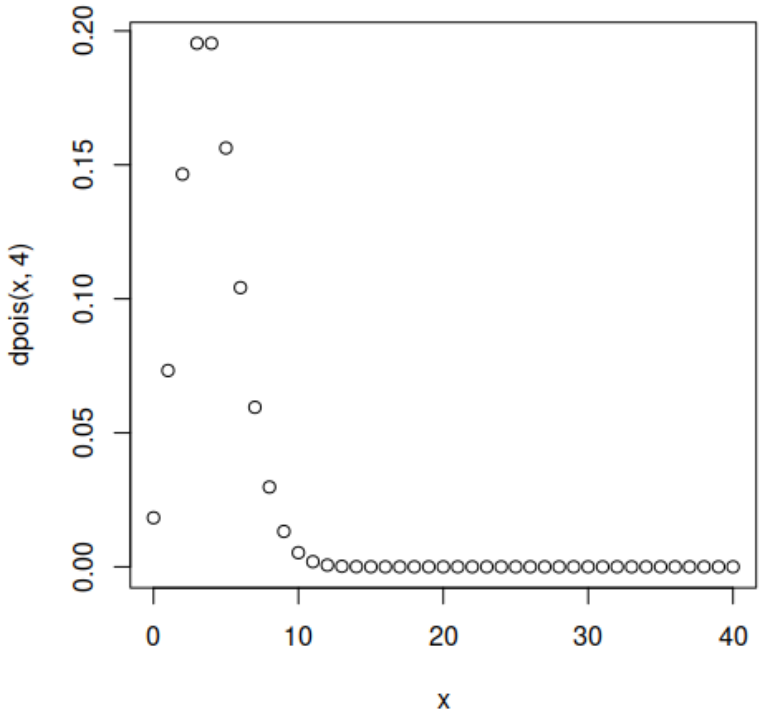
- $X =$ počet vytažených červených míčků při n tazích, v osudí je K červených z N celkových míčků
- Dáno n, N, K.
- Značíme $X \sim Hyper(N, K, n)$.

- $p_X(k) = P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$

Poissonovo rozdělení

- Značíme $X \sim Pois(\lambda)$.
- Dáno reálné $\lambda > 0$.
- $p_X(k) = \frac{\lambda^k}{k!} e^{-\lambda}$
- $Pois(\lambda)$ je limitou $Bin(n, \lambda/n)$
- X popisuje např. počet emailů, které dostaneme za jednu hodinu.

Poissonovo rozdělení: pravděpodobnostní funkce



Vygenerováno následujícím kódem v R

```
x <- seq(0,40,by=1)
plot(x,dpois(x,4))
```

Poissonovo paradigma

- $A_1, \ldots, A_n$ jsou (skoro-)nezávislé jevy s $P(A_i) = p_i$, $\lambda = \sum_i p_i$. Nechť n je velké, každé z p_i malé. Pak přibližně platí

$$\sum_{i=1}^{n} I_{A_i} \sim Pois(\lambda).$$

Geometrické rozdělení

- $X =$ kolikátým hodem mincí padl první orel.
- Značíme $X \sim Geom(p)$.

- Dáno $p \in [0, 1]$.
- $p_X(k) = (1 - p)^{k-1} p$, pro $k = 1, 2, \ldots$

- Někdy se tomuto rozdělení říká posunuté geometrické, a za normální geometrické se považuje rozdělení $X - 1$, tj. počet neúspěšných hodů.

Přehled

Střední hodnota

Definice

Pokud X je diskrétní n.v., tak její střední hodnota (expectation) je označována $\mathbb{E}(X)$ a definována

$$\mathbb{E}(X) = \sum_{x \in Im(X)} x \cdot P(X = x),$$

pokud součet má smysl.

- Nechť X je definována na diskrétním prostoru $(\Omega, \mathcal{F}, P)$. Pak střední hodnotu lze také definovat

$$\mathbb{E}(X) = \sum_{\omega \in \Omega} X(\omega) P(\{\omega\}).$$

LOTUS

- Pro reálnou funkci g a diskrétní n.v. X je $Y = g(X)$ také diskrétní n.v.

Věta (LOTUS)

Pokud X je diskrétní n.v. a g reálná funkce, tak

$$\mathbb{E}(g(X)) = \sum_{x \in Im(X)} g(x) P(X = x)$$

pokud součet má smysl.

Vlastnosti $\mathbb{E}$

Věta

Nechť X, Y jsou diskrétní n.v. a $a, b \in \mathbb{R}$.

1. *Pokud $P(X \geq 0) = 1$ a $\mathbb{E}(X) = 0$, tak $P(X = 0) = 1$.*
2. *Pokud $\mathbb{E}(X) \geq 0$ tak $P(X \geq 0) > 0$.*
3. $\mathbb{E}(a \cdot X + b) = a \cdot \mathbb{E}(X) + b$.
4. $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.

Rozptyl

Definice

Rozptyl (variance) n.v. X nazveme číslo $\mathbb{E}((X - \mathbb{E}X)^2)$.
Značíme jej $var(X)$.

Věta

$$var(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

Podmíněná střední hodnota

Definice

Pokud X je diskrétní n.v. a $P(B) > 0$, tak podmíněná střední hodnota X za předpokladu B (conditional expectation of X given B) je

$$\mathbb{E}(X \mid B) = \sum_{x \in Im(X)} x \cdot P(X = x \mid B),$$

pokud součet má smysl.

Rozbor všech možností

Věta

Pokud $B_1, B_2, \ldots$ je rozklad Ω a $A \in \mathcal{F}$, tak

$$\mathbb{E}(X) = \sum_i \mathbb{E}(X \mid B_i) P(B_i),$$

kdykoliv má součet smysl. (Sčítance s $P(B_i) = 0$ považujeme za 0.)

Rozbor všech možností

Přehled

Parametry rozdělení – Bernoulliho

Pro $X \sim Bern(p)$ je

- $\mathbb{E}(X) = p$
- $var(X) = p - p^2$

Parametry rozdělení – binomické

Pro $X \sim Bin(n, p)$ je

- $\mathbb{E}(X) = np$
- $var(X) = np(1 - p)$

Parametry rozdělení – geometrické

Pro $X \sim Geom(p)$ je

- $\mathbb{E}(X) = 1/p$
- $var(X) = \frac{1-p}{p^2}$

Parametry rozdělení – Poissonovo

Pro $X \sim Pois(\lambda)$ je

- $\mathbb{E}(X) = \lambda$
- $var(X) = \lambda$

Přehled

Základní popis náhodných vektorů

- X, Y – náhodné veličiny na stejném pravděpodobnostním prostoru $(\Omega, \mathcal{F}, P)$.
- Budeme chtít uvažovat (X, Y) jako jeden objekt – náhodný vektor.
- Jak to udělat?
- Příklad: házíme dvakrát čtyřstěnnou kostkou, $X =$ první hod, $Y =$ druhý hod.

Sdružené rozdělení

Definice

Pro diskrétní n.v. X, Y na pravděpodobnostním prostoru $(\Omega, \mathcal{F}, P)$ definujeme jejich sdruženou pravděpodobnostní funkci (joint pmf) $p_{X,Y} : \mathbb{R}^2 \to [0, 1]$ předpisem

$$p_{X,Y}(x, y) = P(\{\omega \in \Omega : X(\omega) = x \& Y(\omega) = y).$$

- Mohli bychom definovat i pro více než dvě n.v. $p_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$.

Marginální rozdělení

- Máme-li dáno $p_{X,Y}$, jak zjistit rozdělení jednotlivých složek, tj. p_X a p_Y?

Funkce náhodného vektoru

Věta

Nechť X, Y jsou n.v. na $(\Omega, \mathcal{F}, P)$, nechť $g : \mathbb{R}^2 \to \mathbb{R}$ je funkce.

- *Pak $Z = g(X, Y)$ je n.v. na $(\Omega, \mathcal{F}, P)$*
- *a platí pro ni*

$$\mathbb{E}(g(X, Y)) = \sum_{x \in Im X} \sum_{y \in Im Y} g(x, y) P(X = x, Y = y).$$

Věta

Pro X, Y n.v. a $a, b \in \mathbb{R}$ platí

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y).$$

Důkaz věty o rozptylu

Nezávislost náhodných veličin

Definice

Diskrétní n.v. X, Y jsou nezávislé (independent) pokud pro každé $x, y \in \mathbb{R}$ jsou jevy $\{X = x\}$ a $\{Y = y\}$ nezávislé. To nastane, právě když

$$P(X = x, Y = y) = P(X = x)P(Y = y).$$

Součin nezávislých n.v.

Věta

Pro nezávislé diskrétní n.v. X, Y platí

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

Součet nezávislých n.v.

- Máme-li dáno $p_{X,Y}$, jak zjistit rozdělení součtu, $Z = X + Y$?

Součet nezávislých n.v. – konvoluce

Věta

Pokud X, Y jsou diskrétní nezávislé náhodné veličiny (zkráceně n.n.v.), tak jejich součet $Z = X + Y$ má pravděpodobnostní funkci

$$P(Z = z) = \sum_{x \in Im X} P(X = x) P(Y = z - x).$$

NMAI059 Pravděpodobnost a statistika 1
4. přednáška

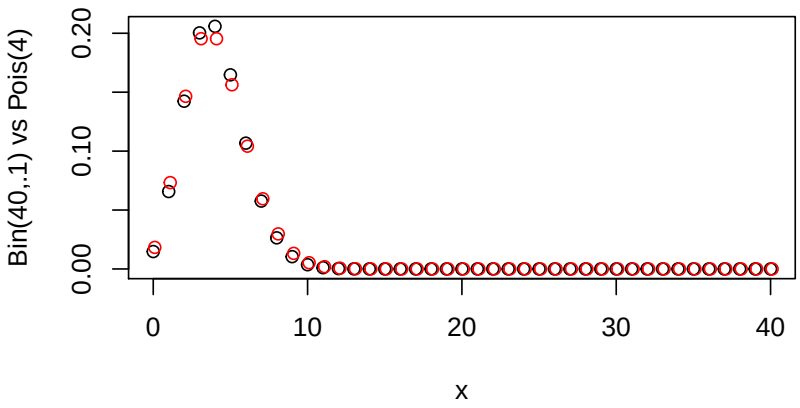
Robert Šámal

Přehled

Co už víme

- Co je diskrétní n.v.
- Jak je popisovat pomocí pravděpodobnostní a/nebo distribuční funkce.
- Příklady rozdělení: Bernoulliho, binomické, hypergeometrické, Poissonovo, geometrické.
- Co je střední hodnota: dvě možné definice:
- $\mathbb{E}(X) = \sum_{x \in Im(X)} x \cdot P(X = x)$
- $\mathbb{E}(X) = \sum_{\omega \in \Omega} X(\omega) P(\{\omega\})$
- $\mathbb{E}(g(X)) = \sum_{x \in Im(X)} g(x) P(X = x)$ (LOTUS)
- „Kolik čekáme, že průměrně dostaneme, když budeme opakovat nezávislé pokusy s výsledkem popsaným X“ ... bude tzv. zákon velkých čísel

Srovnání binomického a Poissonova rozdělení: pravděpodobnostní funkce



Vygenerováno následujícím kódem v R

```
x = 0:40
bin = dbinom(x,40,0.1)
pois = dpois(x,4)
plot(x,bin, ylab="Bin(40,.1) vs Pois(4)")
points(x+.1,pois,col="red")
```

Vlastnosti $\mathbb{E}$

Věta

Nechť X, Y jsou diskrétní n.v. a $a, b \in \mathbb{R}$.

1. *Pokud $P(X \geq 0) = 1$ a $\mathbb{E}(X) = 0$, tak $P(X = 0) = 1$.*
2. *Pokud $\mathbb{E}(X) \geq 0$ tak $P(X \geq 0) > 0$.*
3. $\mathbb{E}(a \cdot X + b) = a \cdot \mathbb{E}(X) + b$.
4. $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.

Alternativní formulka pro střední hodnotu

Věta

Nechť X je diskrétní n.v. nabývající jen hodnot z $\mathbb{N}_0 = \{0, 1, 2, \dots\}$. Pak platí

$$\mathbb{E}(X) = \sum_{n=0}^{\infty} P(X > n).$$

Rozptyl

Definice

Rozptyl (variance) n.v. X nazveme číslo $\mathbb{E}((X - \mathbb{E}X)^2)$.
Značíme jej $var(X)$.

- *Směrodatná odchylka (standard deviation)* $\sigma_X = \sqrt{var(X)}$ – „stejné jednotky jako X“.
- Měří, jak je daleko „typicky“ je X od $\mathbb{E}(X)$. Mohli bychom to měřit i jinak (např. $\mathbb{E}(|X - \mathbb{E}(X)|)$, ale rozptyl je výhodnější).

Věta

$$var(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

Podmíněná střední hodnota

Definice

Pokud X je diskrétní n.v. a $P(B) > 0$, tak podmíněná střední hodnota X za předpokladu B (conditional expectation of X given B) je

$$\mathbb{E}(X \mid B) = \sum_{x \in Im(X)} x \cdot P(X = x \mid B),$$

pokud součet má smysl.

Rozbor všech možností

Věta

Pokud $B_1, B_2, \ldots$ je rozklad Ω a $A \in \mathcal{F}$, tak

$$\mathbb{E}(X) = \sum_i \mathbb{E}(X \mid B_i) P(B_i),$$

kdykoliv má součet smysl. (Sčítance s $P(B_i) = 0$ považujeme za 0.)

Rozbor všech možností

Přehled

Parametry rozdělení – Bernoulliho

Pro $X \sim Bern(p)$ je

- $\mathbb{E}(X) = p$
- $var(X) = p(1-p)$

Parametry rozdělení – binomické

Pro $X \sim Bin(n, p)$ je

- $\mathbb{E}(X) = np$
- $var(X) = np(1-p)$

- První postup: $X = \sum_{i=1}^{n} X_i$, kde $X_i =$
- $\mathbb{E}(X_i) = P(X_i = 1) =$

- Druhý postup:
  $\mathbb{E}(X) = \sum_{k=0}^{n} k \cdot P(X = k) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}$

Parametry rozdělení – hypergeometrické

Pro $X \sim Hyper(N, K, n)$

- $\mathbb{E}(X) = n\frac{K}{N}$
- $var(X) = n\frac{K}{N}(1 - \frac{K}{N})\frac{N-n}{N-1}$

- První postup: $X = \sum_{i=1}^{n} X_i$, kde $X_i =$
- $\mathbb{E}(X_i) = P(X_i = 1) =$

- Druhý postup: $X = \sum_{j=1}^{K} Y_j$, kde $Y_j =$
- $\mathbb{E}(Y_j) = P(Y_j = 1) =$

Parametry rozdělení – geometrické

Pro $X \sim Geom(p)$ je

- $\mathbb{E}(X) = 1/p$
- $var(X) = \frac{1-p}{p^2}$

Parametry rozdělení – Poissonovo

Pro $X \sim Pois(\lambda)$ je

- $\mathbb{E}(X) = \lambda$
- $var(X) = \lambda$

Přehled

Základní popis náhodných vektorů

- X, Y – náhodné veličiny na stejném pravděpodobnostním prostoru $(\Omega, \mathcal{F}, P)$.
- Budeme chtít uvažovat (X, Y) jako jeden objekt – náhodný vektor.
- Jak to udělat?
- Příklad: házíme dvakrát čtyřstěnnou kostkou, $X =$ první hod, $Y =$ druhý hod.

Sdružené rozdělení

Definice

Pro diskrétní n.v. X, Y *na pravděpodobnostním prostoru* $(\Omega, \mathcal{F}, P)$ *definujeme jejich sdruženou pravděpodobnostní funkci (joint pmf)* $p_{X,Y} : \mathbb{R}^2 \to [0, 1]$ *předpisem*

$$p_{X,Y}(x, y) = P(\{\omega \in \Omega : X(\omega) = x \& Y(\omega) = y).$$

- Mohli bychom definovat i pro více než dvě n.v. $p_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$.

Marginální rozdělení

- Máme-li dáno $p_{X,Y}$, jak zjistit rozdělení jednotlivých složek, tj. p_X a p_Y?

Marginální rozdělení

Věta

Nechť X, Y jsou diskrétní n.v. Pak

$$p_X(x) = P(X = x) = \sum_{Y \in Im(Y)} P(X = x \& Y = y) = \sum_{Y \in Im(Y)} p_{X,Y}(x, y)$$

$$p_Y(y) = P(Y = y) = \sum_{X \in Im(X)} P(X = x \& Y = y) = \sum_{X \in Im(X)} p_{X,Y}(x, y)$$

Funkce náhodného vektoru

Věta

Nechť X, Y jsou n.v. na $(\Omega, \mathcal{F}, P)$, nechť $g : \mathbb{R}^2 \to \mathbb{R}$ je funkce.

- *Pak $Z = g(X, Y)$ je n.v. na $(\Omega, \mathcal{F}, P)$*
- *a platí pro ni*

$$\mathbb{E}(g(X,Y)) = \sum_{x \in ImX} \sum_{y \in ImY} g(x,y) P(X = x, Y = y).$$

Věta

Pro X, Y n.v. a $a, b \in \mathbb{R}$ platí

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y).$$

Důkaz věty o rozptylu

Nezávislost náhodných veličin

Definice

Diskrétní n.v. X, Y jsou nezávislé (independent) pokud pro každé $x, y \in \mathbb{R}$ jsou jevy $\{X = x\}$ a $\{Y = y\}$ nezávislé. To nastane, právě když

$$P(X = x, Y = y) = P(X = x)P(Y = y).$$

Součin nezávislých n.v.

Věta

Pro nezávislé diskrétní n.v. X, Y platí

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

Součet nezávislých n.v.

- Máme-li dáno $p_{X,Y}$, jak zjistit rozdělení součtu, $Z = X + Y$?

Součet nezávislých n.v. – konvoluce

Věta

Pokud X, Y jsou diskrétní nezávislé náhodné veličiny (zkráceně n.n.v.), tak jejich součet $Z = X + Y$ má pravděpodobnostní funkci

$$P(Z = z) = \sum_{x \in ImX} P(X = x)P(Y = z - x).$$

Přehled

Podmíněné rozdělení

X, Y – diskrétní náhodné veličiny na $(\Omega, \mathcal{F}, P)$, $A \in \mathcal{F}$

- $p_{X|A}(x) := P(X = x \mid A)$
  příklad: X je výsledek hodu kostkou, $A =$ padlo sudé číslo
- $p_{X|Y}(x|y) = P(X = x \mid Y = y)$ příklad: X, Z jsou výsledky dvou nezávislých hodů kostkou, $Y = X + Z$.

$$p_{X|Y}(6|10) =$$

- $p_{X|Y}$ z $p_{X,Y}$:

Sdružené vs. podmíněné rozdělení

$p_{X,Y}$	...	10	11	12
1				
2				
3				
4				
5				
6				

$p_{X\|Y}$	...	10	11	12
1				
2				
3				
4				
5				
6				

NMAI059 Pravděpodobnost a statistika 1
5. přednáška

Robert Šámal

Přehled

Co už víme

- *Náhodný vektor* = vektor, kde každá souřadnice je náhodná veličina, **stále jen diskrétní.**
- *Sdružená pravděpodobnostní funkce (joint pmf)* $p_{X,Y} : \mathbb{R}^2 \to [0, 1]$ je definována předpisem

$$p_{X,Y}(x, y) = P(X = x \,\&\, Y = y)$$

- *Marginální rozdělení* = rozdělení jednotlivých souřadnic
- Sdružené rozdělení má víc informace, obecně nejde z jednotlivých složek rekonstruovat.
- Jde to pro *nezávislé n.v.* – tam platí (podle definice)

$$p_{X,Y}(x, y) = p_X(x)p_Y(y)$$

a taky (cvičení)

$$P(X \leq x \,\&\, Y \leq y) = P(X \leq x)P(Y \leq y)$$

Marginální rozdělení ze sdruženého

Věta

Nechť X, Y jsou diskrétní n.v. Pak

$$p_X(x) = \sum_{Y \in Im(Y)} P(X = x \,\&\, Y = y) = \sum_{Y \in Im(Y)} p_{X,Y}(x, y)$$

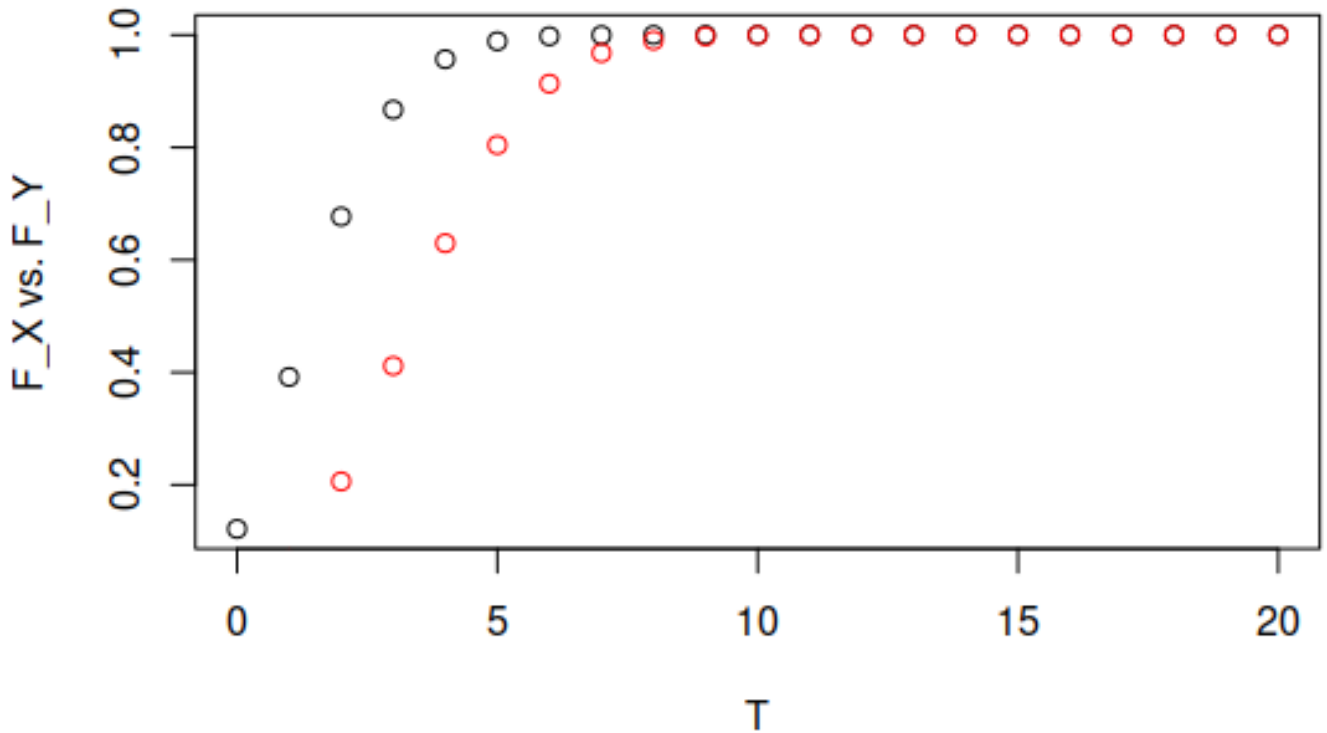
$$p_Y(y) = \sum_{X \in Im(X)} P(X = x \,\&\, Y = y) = \sum_{X \in Im(X)} p_{X,Y}(x, y)$$

Příklad: Multinomické rozdělení

- Na kostce padne číslo i s pravděpodobností p_i pro $i = 1, \ldots, 6$. Hodíme n-krát a označíme X_i počet hodů, kdy padlo i.

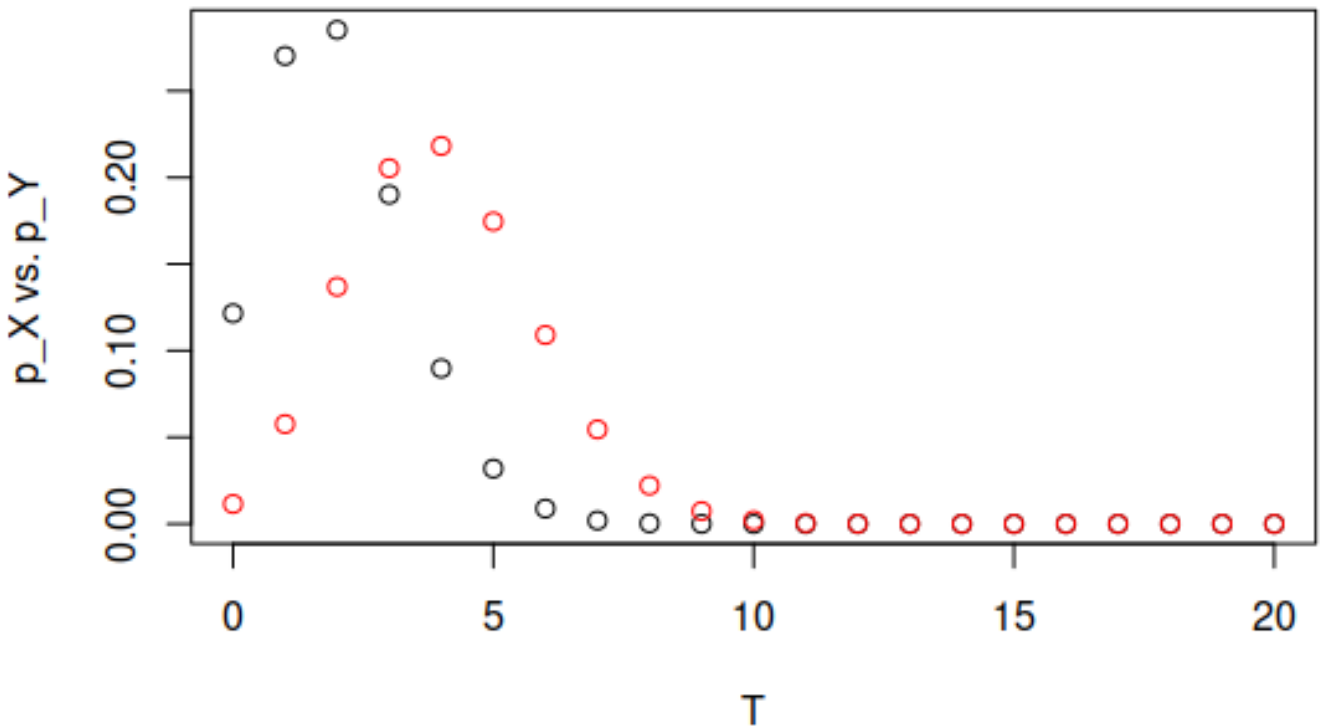
Sdružování/Coupling – netriviální využití sdružených rozdělení

- $X \sim Bin(n,p)$ a $Y \sim Bin(n,q)$ a pro $p < q$
- Co můžeme říct o F_X a F_Y?
- $\sum_{i=0}^{k} \binom{n}{k} p^k (1-p)^{n-k}$ je rostoucí funkce p – ale proč?



Sdružování/Coupling – netriviální využití sdružených rozdělení

- $X \sim Bin(n,p)$ a $Y \sim Bin(n,q)$ a pro $p < q$
- Co můžeme říct o F_X a F_Y?
- $\sum_{i=0}^{k} \binom{n}{k} p^k (1-p)^{n-k}$ je rostoucí funkce p – ale proč?



Coupling

- $X = \sum_{i=1}^{n} X_i$, kde $X_1, \ldots, X_n$ jsou n.n.v
- $Y = \sum_{i=1}^{n} Y_i$, kde $Y_1, \ldots, Y_n$ jsou n.n.v
- Vztah X a Y je není určen – můžou být jakékoliv.
- Zařídíme, že nebudou nezávislé, dokonce bude vždy $X \leq Y$.
- Stačí definovat $Y_i =$

Funkce náhodného vektoru

Věta

Nechť X, Y jsou n.v. na $(\Omega, \mathcal{F}, P)$, nechť $g : \mathbb{R}^2 \to \mathbb{R}$ je funkce.

- *Pak $Z = g(X, Y)$ je n.v. na $(\Omega, \mathcal{F}, P)$*
- *a platí pro ni*

$$\mathbb{E}(g(X,Y)) = \sum_{x \in ImX} \sum_{y \in ImY} g(x,y) P(X = x, Y = y).$$

Věta

Pro X, Y n.v. a $a, b \in \mathbb{R}$ platí

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y).$$

Součin nezávislých n.v.

Věta

Pro nezávislé diskrétní n.v. X, Y platí

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

Důkaz byl minule, využili jsme tehdy nejasný krok

$$\mathbb{E}(XY) = \sum_{x \in Im(X), y \in Im(Y)} xy \cdot P(X = x \,\&\, Y = y)$$

Součet nezávislých n.v.

- Máme-li dáno $p_{X,Y}$, jak zjistit rozdělení součtu, $Z = X + Y$?

Součet nezávislých n.v. – konvoluce

Věta

Pokud X, Y jsou diskrétní náhodné veličiny, tak pro $Z = X + Y$ platí

$$P(Z = z) = \sum_{x \in Im(X)} P(X = x, Y = z - x).$$

Pokud X, Y jsou navíc nezávislé, tak

$$P(Z = z) = \sum_{x \in ImX} P(X = x)P(Y = z - x).$$

Ukázka konvoluce

Přehled

Podmíněné rozdělení

X, Y – diskrétní náhodné veličiny na $(\Omega, \mathcal{F}, P)$, $A \in \mathcal{F}$

- $p_{X|A}(x) := P(X = x \mid A)$
  příklad: X je výsledek hodu kostkou, $A =$ padlo sudé číslo
- $p_{X|Y}(x|y) = P(X = x \mid Y = y)$ příklad: X, Z jsou výsledky dvou nezávislých hodů kostkou, $Y = X + Z$.

$$p_{X|Y}(6|10) =$$

- $p_{X|Y}$ z $p_{X,Y}$:

Sdružené vs. podmíněné rozdělení

$p_{X,Y}$	...	10	11	12
1				
2				
3				
4				
5				
6				

$p_{X\|Y}$	...	10	11	12
1				
2				
3				
4				
5				
6				

Přehled

Obecná náhodná veličina

Definice

Náhodná veličina (random variable) na $(\Omega, \mathcal{F}, P)$ *je zobrazení* $X : \Omega \to \mathbb{R}$*, které pro každé* $x \in \mathbb{R}$ *splňuje*

$$\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}.$$

- diskrétní n.v. je n.v.

Distribuční funkce

Definice

Distribuční funkce (cumulative distribution function, CDF) n.v. X je funkce

$$F_X(x) := P(X \leq x) = P(\{\omega \in \Omega : X(\omega) \leq x).$$

- F_X je neklesající funkce
- $\lim_{x \to -\infty} F_X(x) = 0$
- $\lim_{x \to +\infty} F_X(x) = 1$
- F_X je zprava spojitá

Distribuční funkce – další ukázky

Kvantilová funkce

Pro nádhnou veličinu X definujeme *kvantilovou funkci* $Q_X : [0,1] \to \mathbb{R}$ pomocí

$$Q_X(p) := \min\{x \in \mathbb{R} : p \leq F_X(x)\}$$

- Pokud F_X je spojitá, tak $Q_X = F_X^{-1}$.
- $Q_X(1/2) =$ medián (pozor, když F_X není rostoucí)
- $Q_X(10/100) =$ desátý percentil, atd.

Spojitá náhodná veličina

Definice

N.v. X se nazývá spojitá (continuous), pokud existuje nezáporná reálná funkce f_X tak, že

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt.$$

(Někdy se též používá pojem absolutně spojitá veličina.)
Funkce f_X se nazývá hustota (probability density function, pdf) náhodné veličiny X.

Práce s hustotou

Věta

Nechť spojitá n.v. X má hustotu f_X. Pak

1. $P(X = x) = 0$ *pro každé* $x \in \mathbb{R}$.
2. $P(a \leq X \leq b) = \int_a^b f_X(t)dt$ *pro každé* $a, b \in \mathbb{R}$.

Uniformní rozdělení

- N.v. X má uniformní rozdělení na intervalu $[a, b]$, píšeme $X \sim U(a, b)$, pokud $f_X(x) = 1/(b - a)$ pro $x \in [a, b]$ a $f_X(x) = 0$ jinak.

Universalita unif.

Věta

Nechť F je funkce „typu distribuční funkce“: neklesající zprava spojitá funkce s $\lim_{x\to-\infty} F(x) = 0$ a $\lim_{x\to+\infty} F(x) = 1$. Nechť Q je odpovídající kvantilová funkce.

1. *Nechť $U \sim U(0,1)$ a $X = Q(U)$. Pak X má distribuční funkci F.*
2. *Nechť X je n.v. s distribuční funkcí $F_X = F$, nechť F je rostoucí. Pak $F(X) \sim U(0,1)$.*

Střední hodnota spojité n.v.

Definice

Nechť spojitá n.v. X má hustotu f_X. Pak její střední hodnota (expectation, expected value, mean) je označována $\mathbb{E}(X)$ a definována

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x \; f_X(x) dx,$$

pokud integrál má smysl, tj. pokud se „nejedná o typ $\infty - \infty$“.

- Analogie s výpočtem těžiště tyče ze znalosti hustoty.

Spojitý LOTUS

Věta (LOTUS)

Pokud X je spojitá n.v. s hustotou f_X a g reálná funkce, tak

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx,$$

pokud integrál má smysl.

(Důkaz vynecháme.)

Přehled

Shrnutí, co už víme

- $F_X(x) := P(X \leq x)$ distribuční funkce, existuje vždy, je neklesající, zprava spojitá, $F_X(-\infty) = 0$, $F_X(+\infty) = 1$
- $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$ pro tzv. spojité n.v. Pro ty dále platí:
- $P(a \leq X \leq b) = \int_a^b f_X(t)dt$, spec. tedy $P(X = x) = 0$ pro každé $x \in \mathbb{R}$
- obecněji: $P(X \in A) = \int_A f_X(t)dt$, kdykoli umíme přes množinu A integrovat
- spec. tedy $\int_{-\infty}^{\infty} f_X(t)dt = 1$
- $\mathbb{E}(X) = \int_{-\infty}^{\infty} t\, f_X(t)dt$
- $\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(t) f_X(t)dt$
- Pokud je hustota spojitá, tak navíc platí: $f_X = F_X'$ (základní věta kalkulu).

Uniformní rozdělení

- N.v. X má uniformní rozdělení na intervalu $[a, b]$, píšeme $X \sim U(a, b)$, pokud $f_X(x) = 1/(b - a)$ pro $x \in [a, b]$ a $f_X(x) = 0$ jinak.

Rozptyl spojité n.v.

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x \; f_X(x) dx$$

$$\mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 \; f_X(x) dx$$

Označíme-li $\mu = \mathbb{E}(X)$, tak

$$var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \; f_X(x) dx = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$$

NMAI059 Pravděpodobnost a statistika 1
6. přednáška

Robert Šámal

Přehled

Obecná náhodná veličina – co už víme

- N.v. je zobrazení $X : \Omega \to \mathbb{R}$, které pro každé $x \in \mathbb{R}$ splňuje $\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$.
- Diskrétní n.v. je n.v.
- Distribuční funkce n.v. X je funkce $F_X(x) := P(X \leq x)$.
- Distr. fce je F_X je neklesající, zprava spojitá, s definovanými limitami v $\pm\infty$.

Kvantilová funkce

Pro náhodnou veličinu X definujeme *kvantilovou funkci* $Q_X : [0,1] \to \mathbb{R}$ pomocí

$$Q_X(p) := \min\{x \in \mathbb{R} : p \leq F_X(x)\}$$

- Pokud F_X je spojitá, tak $Q_X = F_X^{-1}$.
- Obecně platí: $Q_X(p) \leq x \Leftrightarrow p \leq F_X(x)$.
- $Q_X(1/2) =$ medián (pozor, když F_X není rostoucí)
- Pokud F_X je spojitá, tak $Q_X = F_X^{-1}$.

Spojitá náhodná veličina

Definice

N.v. X se nazývá spojitá (continuous), pokud existuje nezáporná reálná funkce f_X tak, že

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt.$$

(Někdy se též používá pojem absolutně spojitá veličina.)
Funkce f_X se nazývá hustota (probability density function, pdf) náhodné veličiny X.

- Alternativně: máme zadanou funkci $f \geq 0$ s $\int_{-\infty}^{\infty} f = 1$.
- Vybereme náhodný bod pod grafem f.
- Označíme jeho souřadnice (X, Y).
- Pak X je n.v. s hustotou f.

Práce s hustotou

Věta

Nechť spojitá n.v. X má hustotu f_X. Pak

1. $P(X = x) = 0$ *pro každé* $x \in \mathbb{R}$.
2. $P(a \leq X \leq b) = \int_a^b f_X(t)dt$ *pro každé* $a, b \in \mathbb{R}$.

- V důsledku taky platí (pro "rozumnou množinu A")

$$P(X \in A) = \int_A f_X(t)dt$$

Střední hodnota spojité n.v.

Definice

Nechť spojitá n.v. X má hustotu f_X. Pak její střední hodnota (expectation, expected value, mean) je označována $\mathbb{E}(X)$ a definována

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x \; f_X(x) dx,$$

pokud integrál má smysl, tj. pokud se „nejedná o typ $\infty - \infty$“.

- Analogie s výpočtem těžiště tyče ze znalosti hustoty.
- Diskretizace.

Vlastnosti střední hodnoty

Věta (LOTUS)

Pokud X je spojitá n.v. s hustotou f_X a g reálná funkce, tak

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx,$$

pokud integrál má smysl.

(Důkaz vynecháme, dělal by se pomocí substituce v integrálu.)

Věta (Linearita střední hodnoty)

Pro $X_1, \ldots, X_n$ diskrétní nebo spojité n.v. platí

$$\mathbb{E}(X_1 + \cdots + X_n) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n).$$

(Důkaz bude později.)

Rozptyl spojité n.v.

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x \; f_X(x) dx$$

$$\mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 \; f_X(x) dx$$

Označíme-li $\mu = \mathbb{E}(X)$, tak

$$var(X) := \mathbb{E}((X-\mu)^2) = \int_{-\infty}^{\infty} (x-\mu)^2 \; f_X(x) dx.$$

Věta

I pro spojité n.v. platí $var(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$.

(Důkaz jako pro diskrétní n.v.)

Rozptyl součtu

Věta (Rozptyl součtu)

Pro $X_1, \ldots, X_n$ nezávislé diskrétní nebo spojité n.v. platí

$$var(X_1 + \cdots + X_n) = var(X_1) + \cdots + var(X_n).$$

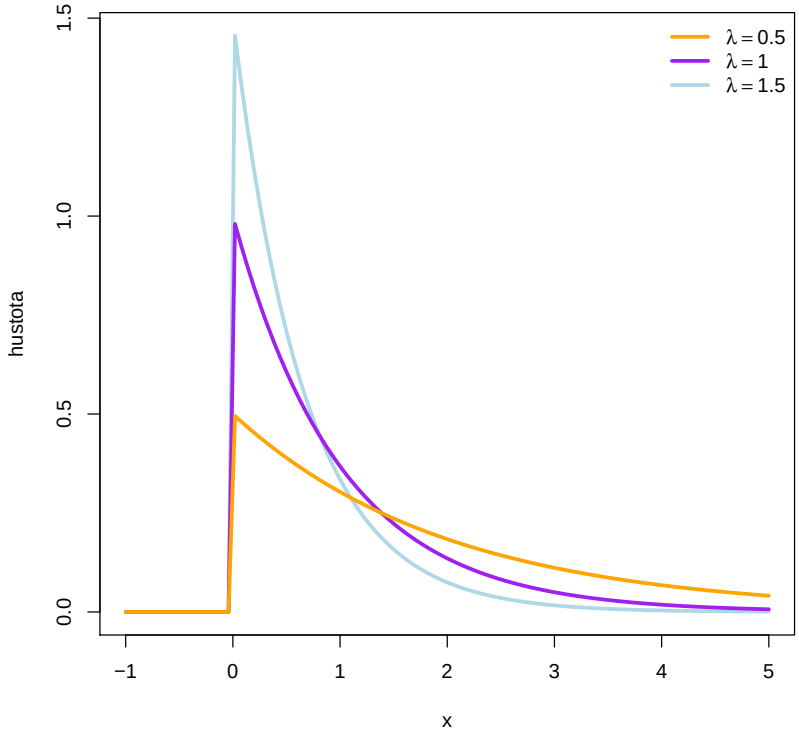
Přehled

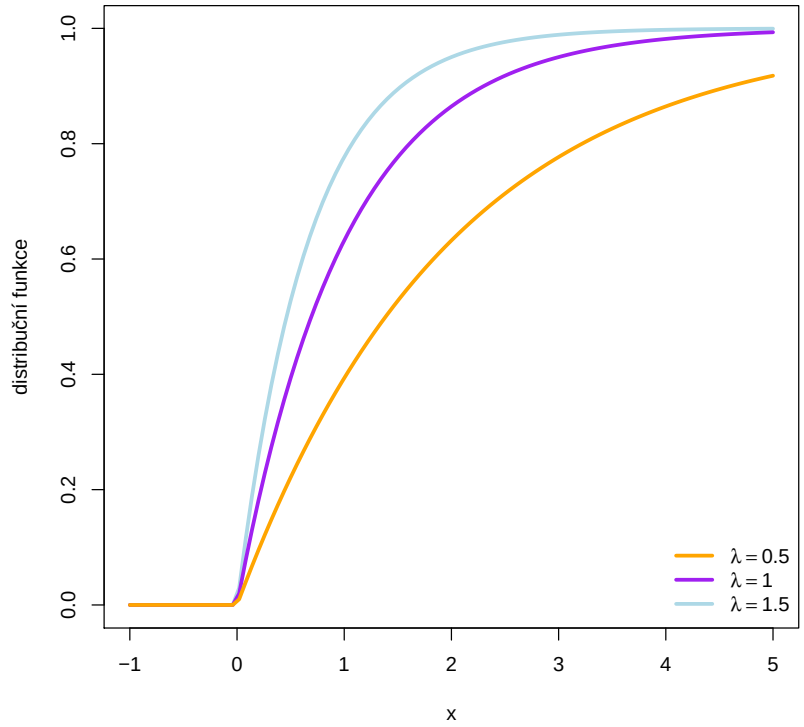
Uniformní rozdělení

- N.v. X má uniformní rozdělení na intervalu $[a, b]$, píšeme $X \sim U(a, b)$, pokud $f_X(x) = 1/(b - a)$ pro $x \in [a, b]$ a $f_X(x) = 0$ jinak.

Exponenciální rozdělení

$$F_X(x) = \begin{cases} 0 & \text{pro } x \leq 0 \\ 1 - e^{-\lambda x} & \text{pro } x \geq 0 \end{cases}$$

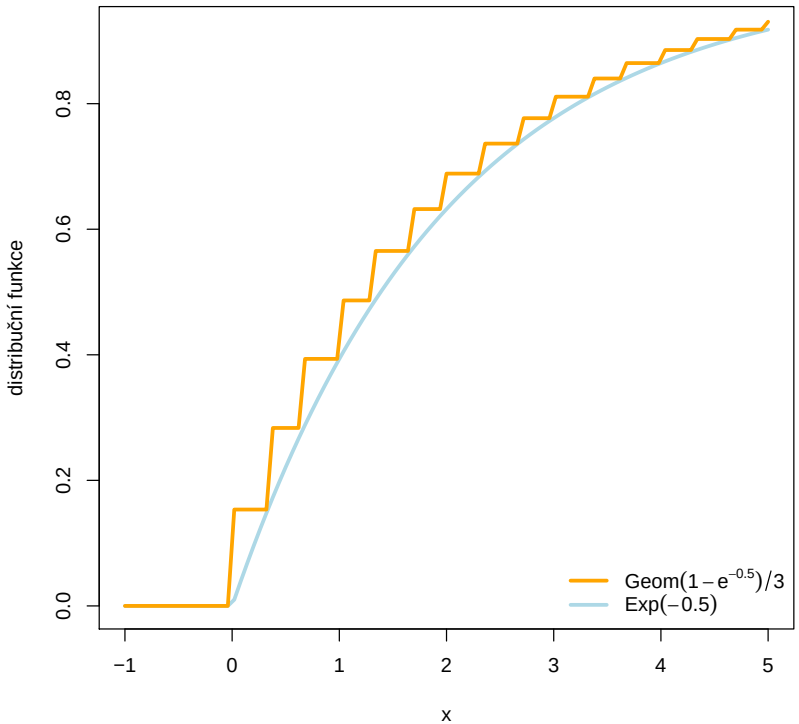




- X modeluje např. čas před příchodem dalšího telefonního hovoru do call-centra/dotazu na web-server/čas do dalšího blesku v bouřce/...

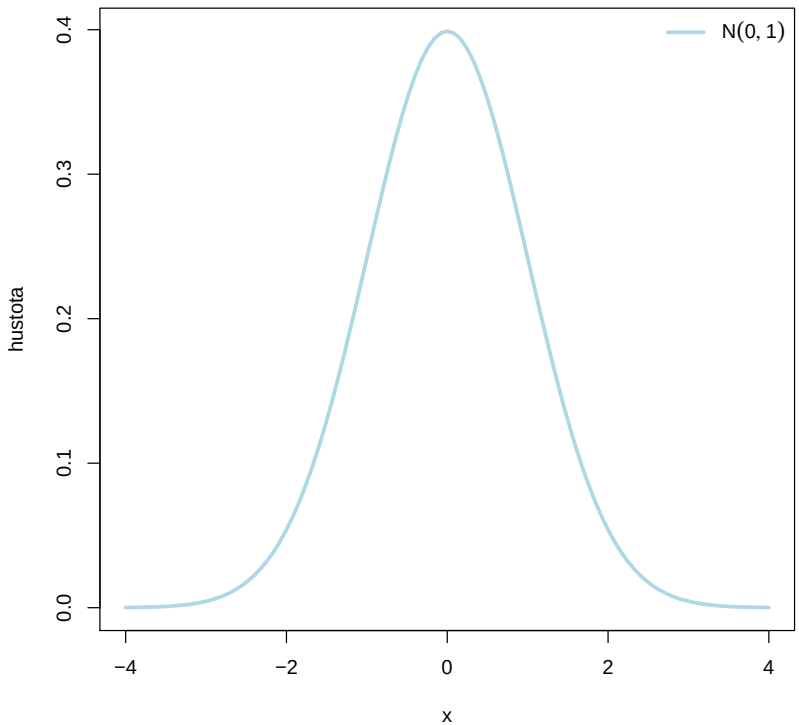
Souvislost $X \sim Exp(\lambda)$ a $Y \sim Geom(p)$

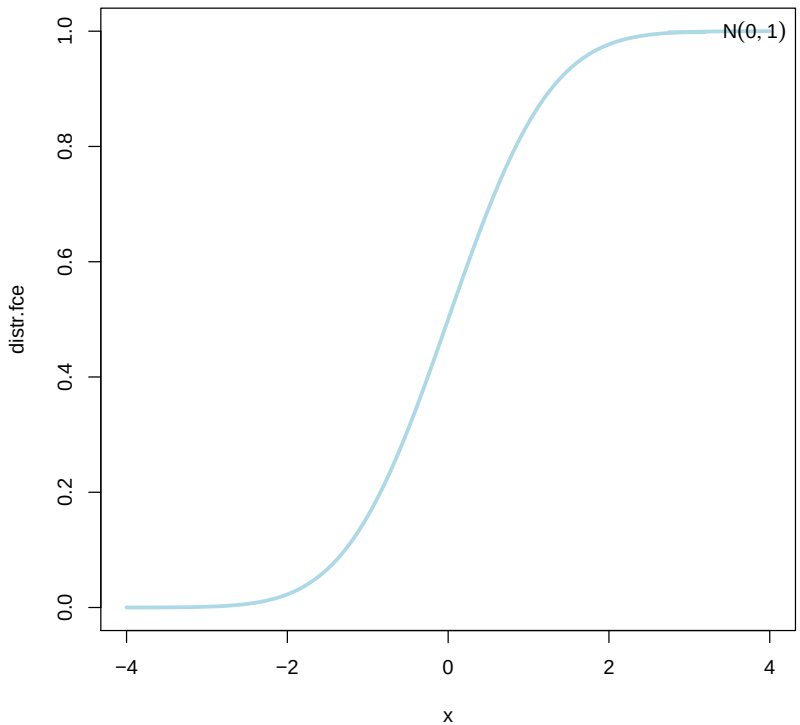
- $P(X > x) = e^{-\lambda x}$ pro $x > 0$
- $P(Y > n) = (1-p)^n$ pro $n \in \mathbb{N}$



Standardní normální rozdělení

- $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$
- $\Phi(x)$ – primitivní funkce k φ
- *Standardní normální rozdělení* $N(0,1)$ má hustotu φ a distribuční funkci Φ.
- Pokud $Z \sim N(0,1)$, tak $\mathbb{E}(Z) = 0$, $var(Z) = 1$





Obecné normální rozdělení

- Pro $\mu, \sigma \in \mathbb{R}$, $\sigma > 0$ položme $X = \mu + \sigma \cdot Z$, kde $Z \sim N(0,1)$.
- Píšeme $X \sim N(\mu, \sigma^2)$ – obecné normální rozdělení.
- Normální rozdělení $N(\mu, \sigma^2)$ má hustotu $\frac{1}{\sigma}\varphi(\frac{x-\mu}{\sigma})$.

Odolnost vůči součtu

- Pokud $X_1, \ldots, X_k$ jsou n.n.v., kde $X_i \sim N(\mu_i, \sigma_i^2)$, pak

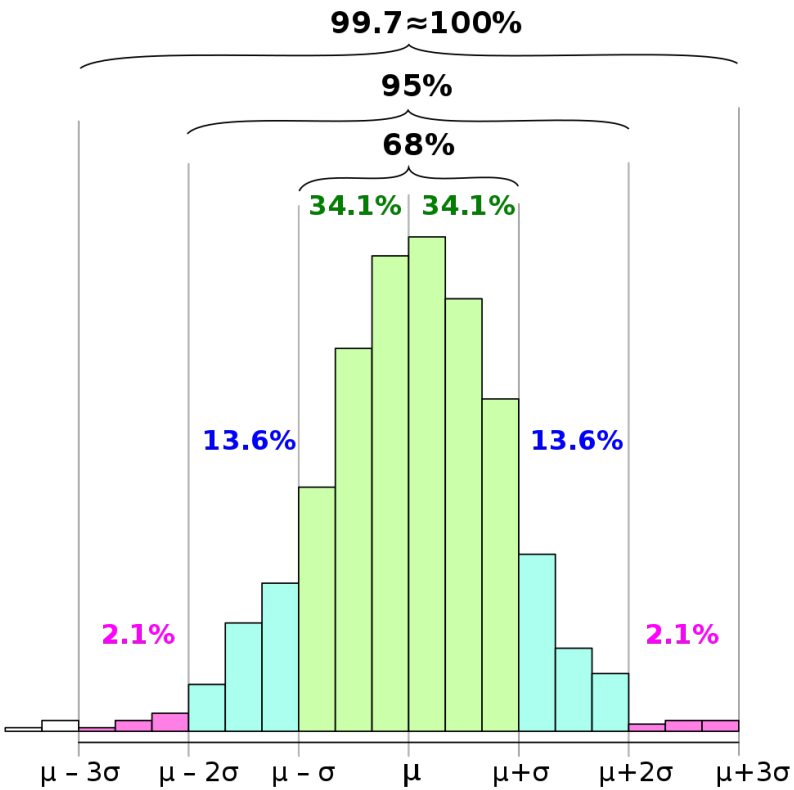
$$X_1 + \cdots + X_k \sim N(\mu, \sigma^2),$$

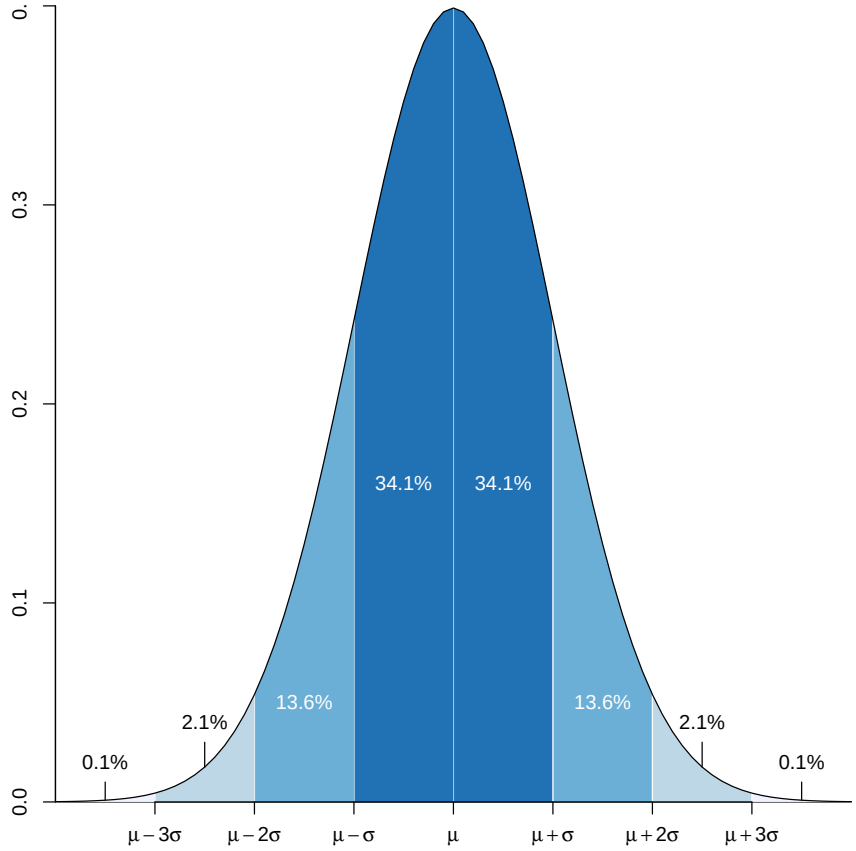
kde $\mu =$

$\sigma =$

Normální rozdělení – klíčové vlastnosti

- Pravidlo 3σ (68–95–99.7 rule)
- Centrální limitní věta

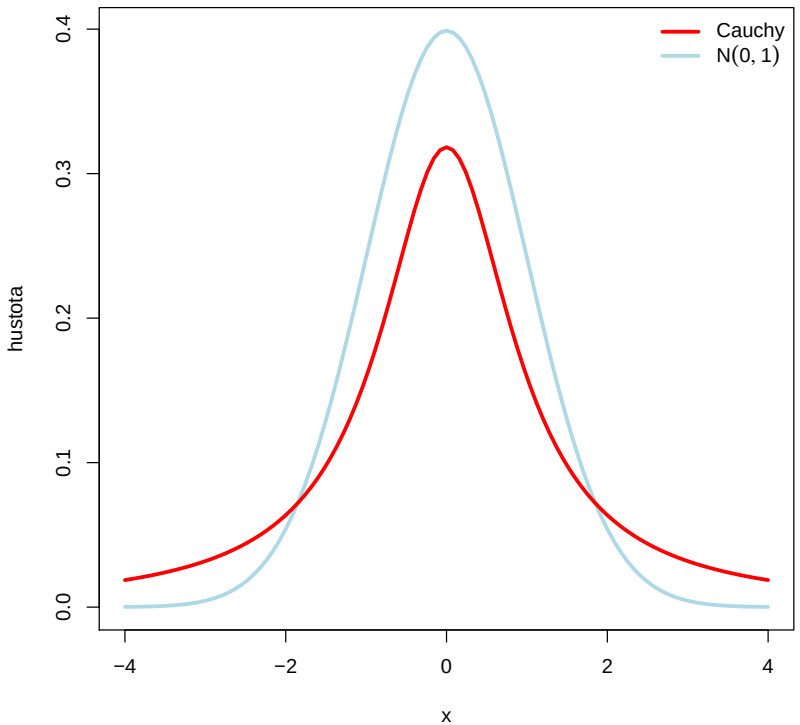




(Obrázek vlevo z Wikipedie, autor Melikamp.)

Cauchyho rozdělení

- *Cauchyho rozdělení:* hustota $f(x) = \frac{1}{\pi(1+x^2)}$
- nemá střední hodnotu!!!



Gamma rozdělení

- $Gamma(w, \lambda)$, *gamma rozdělení s parametry* $w > 0$ *a* $\lambda > 0$ má hustotu

$$f(x) = \begin{cases} 0 & \text{pro } x \leq 0 \\ \frac{1}{\Gamma(w)} \lambda^w x^{w-1} e^{-\lambda x} & \text{pro } x \geq 0 \end{cases}$$

  kde $\Gamma(w) = (w-1)! = \int_0^\infty x^{w-1} e^{-x} dx$.
- Pro $w = 1$ dostáváme znovu exponenciální rozdělení.
- Pokud $X_1, \ldots, X_n$ jsou n.n.v. s rozdělením $Exp(\lambda)$, tak $X_1 + \cdots + X_n \sim Gamma(n, \lambda)$.
- Modeluje mj. životnost součástky, souhrn dešťových srážek za rok, latenci webového serveru.

A mnoho dalších

- $Beta(s,t)$ – beta rozdělení
- χ^2 rozdělení s k stupni volnosti = chí-kvadrát (χ^2_k) je jiné jméno pro $Gamma(\frac{1}{2}k, \frac{1}{2})$. Je to rozdělení $Z_1^2 + \cdots + Z_k^2$, kde $Z_i \sim N(0,1)$ jsou n.n.v.
- Studentova t-distribuce
- atd. atd.

Uniformní rozdělení

- N.v. X má uniformní rozdělení na intervalu $[a, b]$, píšeme $X \sim U(a, b)$, pokud $f_X(x) = 1/(b - a)$ pro $x \in [a, b]$ a $f_X(x) = 0$ jinak.

Universalita unif.

Věta

Nechť X je n.v. s distribuční funkcí $F_X = F$, nechť F je spojitá a rostoucí. Pak $F(X) \sim U(0,1)$.

Věta

Nechť F je funkce „typu distribuční funkce“: neklesající zprava spojitá funkce s $\lim_{x\to-\infty} F(x) = 0$ a $\lim_{x\to+\infty} F(x) = 1$. Nechť Q je odpovídající kvantilová funkce.
Nechť $U \sim U(0,1)$ a $X = Q(U)$. Pak X má distribuční funkci F.

Přehled

Sdružená distribuční funkce (Joint cdf)

Definice

Pro n.v. X, Y na pravděpodobnostním prostoru $(\Omega, \mathcal{F}, P)$ definujeme jejich sdruženou distribuční funkci (joint cdf) $F_{X,Y} : \mathbb{R}^2 \to [0,1]$ předpisem

$$F_{X,Y}(x,y) = P(\{\omega \in \Omega : X(\omega) \leq x \,\&\, Y(\omega) \leq y).$$

- Mohli bychom definovat i pro více než dvě n.v. $F_{X_1,\ldots,X_n}(x_1,\ldots,x_n)$.
- Můžeme odsud odvodit pravděpodobnost obdélníku: $P(X \in (a,b] \,\&\, Y \in (c,d]) =$

Sdružená hustota (Joint pdf)

- Často můžeme sdruženou distribuční funkci psát jako integrál pomocí nezáporné funkce $f_{X,Y}$

$$F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(x,y)dxdy.$$

- Pak nazýváme n.v. X, Y sdruženě spojité. Funkce $f_{X,Y}$ je jejich *sdružená hustota.*
- Jako u jednorozměrného případu může být $f_{X,Y} > 1$.
- Stejně jako u jednorozměrného případu můžeme pak pomocí hustoty vyjádřit i další pravděpodobnosti:

$$P((X,Y) \in S) = \int_{S} f_{X,Y}(x,y)dxdy.$$

- $f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}$

LOTUS

- Stejně jako v diskrétním případu platí pro střední hodnotu funkce dvou n.v.

$$\mathbb{E}(g(X,Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)dxdy.$$

- A stejně jako v diskrétním případu odsud odvodíme $\mathbb{E}(aX + bY + c) = a \cdot \mathbb{E}(X) + b \cdot \mathbb{E}(Y) + c$.

Vícerozměrné normální rozdělení

NMAI059 Pravděpodobnost a statistika 1
7. přednáška

Robert Šámal

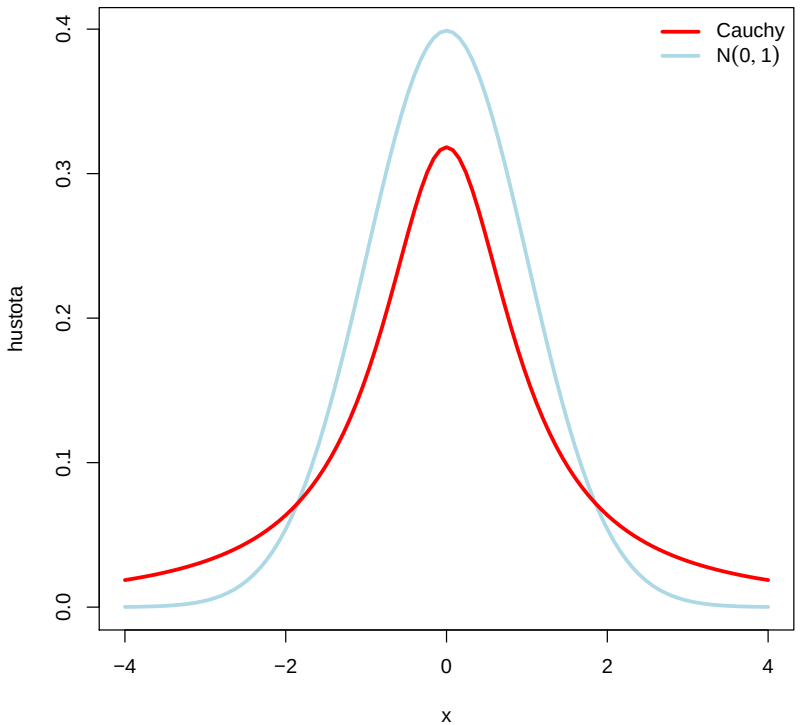
Přehled

Jaká rozdělení jsme už potkali

- $U(a, b)$ – uniformní – rovnoměrné na intervalu $[a, b]$
- $Exp(\lambda)$ – exponenciální – za jak dlouho se utrhne ucho u džbánu
- $N(\mu, \sigma^2)$ – normální – kolik váží chleba

Cauchyho rozdělení

- *Cauchyho rozdělení:* hustota $f(x) = \frac{1}{\pi(1+x^2)}$
- nemá střední hodnotu!!!

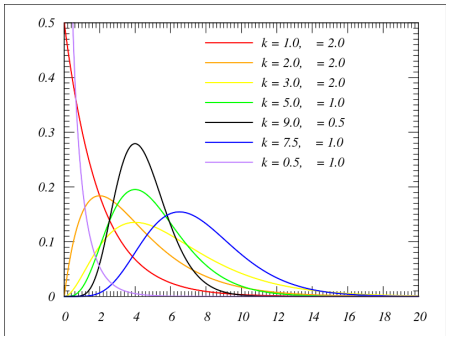


Gamma rozdělení

- $Gamma(w, \lambda)$, *gamma rozdělení s parametry* $w > 0$ *a* $\lambda > 0$ má hustotu

$$f(x) = \begin{cases} 0 & \text{pro } x \leq 0 \\ \frac{1}{\Gamma(w)} \lambda^w x^{w-1} e^{-\lambda x} & \text{pro } x \geq 0 \end{cases}$$

  kde $\Gamma(w) = (w-1)! = \int_0^\infty x^{w-1} e^{-x} dx$.
- Pro $w = 1$ dostáváme znovu exponenciální rozdělení.
- Pokud $X_1, \ldots, X_n$ jsou n.n.v. s rozdělením $Exp(\lambda)$, tak $X_1 + \cdots + X_n \sim Gamma(n, \lambda)$.
- Modeluje mj. životnost součástky, souhrn dešťových srážek za rok, latenci webového serveru.



A mnoho dalších

- $Beta(s,t)$ – beta rozdělení
- χ^2 rozdělení s k stupni volnosti = chí-kvadrát (χ^2_k) je jiné jméno pro $Gamma(\frac{1}{2}k, \frac{1}{2})$. Je to rozdělení $Z_1^2 + \cdots + Z_k^2$, kde $Z_i \sim N(0,1)$ jsou n.n.v.
- Studentova t-distribuce
- atd. atd.

Uniformní rozdělení

- N.v. X má uniformní rozdělení na intervalu $[a, b]$, píšeme $X \sim U(a, b)$, pokud $f_X(x) = 1/(b - a)$ pro $x \in [a, b]$ a $f_X(x) = 0$ jinak.

Universalita unif.

Věta

Nechť X je n.v. s distribuční funkcí $F_X = F$, nechť F je spojitá a rostoucí. Pak $F(X) \sim U(0,1)$.

Věta

Nechť F je funkce „typu distribuční funkce“: neklesající zprava spojitá funkce s $\lim_{x\to-\infty} F(x) = 0$ a $\lim_{x\to+\infty} F(x) = 1$. Nechť Q je odpovídající kvantilová funkce.
Nechť $U \sim U(0,1)$ a $X = Q(U)$. Pak X má distribuční funkci F.

Přehled

Sdružená distribuční funkce (Joint cdf)

Definice

Pro n.v. X, Y na pravděpodobnostním prostoru $(\Omega, \mathcal{F}, P)$ definujeme jejich sdruženou distribuční funkci (joint cdf) $F_{X,Y} : \mathbb{R}^2 \to [0,1]$ předpisem

$$F_{X,Y}(x,y) = P(\{\omega \in \Omega : X(\omega) \leq x \,\&\, Y(\omega) \leq y\}).$$

- Formální podmínka: potřebujeme $\{X \leq x \,\&\, Y \leq y\} \in \mathcal{F}$, jinak (X, Y) není náhodný vektor.
- Mohli bychom definovat i pro více než dvě n.v. $F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) =$
- Můžeme odsud odvodit pravděpodobnost obdélníku: $P(X \in (a,b] \,\&\, Y \in (c,d]) =$

Sdružená hustota (Joint pdf)

- Často můžeme sdruženou distribuční funkci psát jako integrál pomocí nezáporné funkce $f_{X,Y}$

$$F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(s,t)dsdt.$$

- Pak nazýváme n.v. X, Y *sdruženě spojité*. Funkce $f_{X,Y}$ je jejich *sdružená hustota.*
- Jako u jednorozměrného případu může být $f_{X,Y} > 1$.
- Stejně jako u jednorozměrného případu můžeme pak pomocí hustoty vyjádřit i další pravděpodobnosti, pro „rozumnou množinu A“.

$$P((X,Y) \in A) = \int_{A} f_{X,Y}(x,y)dxdy.$$

- $f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}$

- $f_{X,Y}(x,y) \doteq \frac{P(x \leq X \leq x+\Delta_x \;\&\; y \leq Y \leq y+\Delta_y)}{\Delta_x \Delta_y}$

LOTUS

- Analogicky jako v diskrétním případu platí pro střední hodnotu funkce dvou n.v.

$$\mathbb{E}\big(g(X,Y)\big) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,dx\,dy.$$

- A tak jako v diskrétním případu odsud odvodíme

$$\mathbb{E}\big(aX + bY + c\big) = a \cdot \mathbb{E}\big(X\big) + b \cdot \mathbb{E}\big(Y\big) + c.$$

Nezávislost spojitých náhodných veličin

Definice

Libovolné náhodné veličiny nazveme nezávislé (independent), pokud jevy $\{X \leq x\}$ *a* $\{Y \leq y\}$ *jsou nezávislé pro libovolná* $x, y \in \mathbb{R}$. *Ekvivalentně,*

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y),$$
$$F_{X,Y}(x, y) = F_X(x)F_Y(y)$$

Věta

Nechť X, Y *mají sdruženou hustotu* $f_{X,Y}$ *(a hustoty* f_X, f_Y*). Následující tvrzení jsou ekvivalentní:*

- X, Y *jsou nezávislé*
- $f_{X,Y}(x, y) = f_X(x)f_Y(y)$

Vícerozměrné normální rozdělení

- $\varphi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$
- $f(t_1, \ldots, t_n) = \varphi(t_1)\varphi(t_2)\cdots\varphi(t_n) = \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{t_1^2+\cdots+t_n^2}{2}}$
- $f(t_1, \ldots, t_n) = (2\pi)^{-n/2} e^{-r^2/2}$, kde $r^2 = t_1^2 + \cdots + t_n^2$
  radiálně symetrická funkce

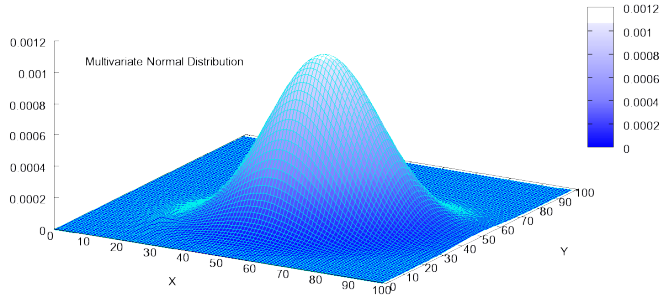


Image by Wikipedia editor Piotrg.

- Nechť $Z = (Z_1, \ldots, Z_n)$ má hustotu f.
- $Z_1, \ldots, Z_n$ jsou n.n.v., $Z_i \sim N(0,1)$

- $Z/\|Z\|$ je uniformně náhodný bod na n-rozměrné sféře.
- tudíž skal. součin Z s libovolným jednotkovým vektorem je $N(0,1)$
- $\langle u, Z\rangle = \sum_{i=1}^n u_i Z_i$ má také rozdělení $N(0,1)$

Vícerozměrné normální rozdělení obecné

- Obecněji můžeme vzít náhodný vektor s hustotou $c \cdot e^{Q(t)}$, kde $c > 0$ je vhodná konstanta a $Q(t)$ je obecná kvadratická funkce.
- Používá se ve strojovém učení.
- Souřadnice nejsou nezávislé!

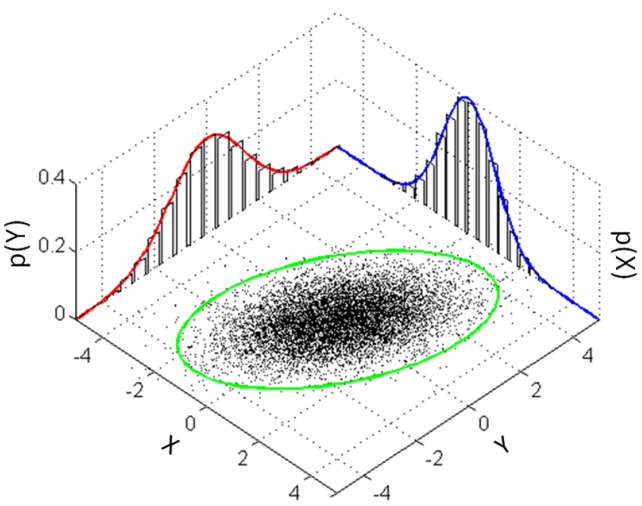


Image by Wikipedia editor Bscan.

Součet spojitých n.v.

Věta

Nechť spojité X, Y jsou n.n.v. Pak $Z = X + Y$ je také spojitá n.v. a její hustotu dostaneme jako konvoluci funkcí f_X, f_Y, neboli

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) dx.$$

Podmiňování

Definice

X je n.v. na $(\Omega, \mathcal{F}, P)$, $B \in \mathcal{F}$.

$$F_{X|B}(x) := P(X \leq x \mid B)$$

K tomu přísluší hustotní funkce $f_{X|B}$.

Věta

Nechť $B_1, B_2, \ldots$ je rozklad Ω. Pak

$$F_X(x) = \sum_i F_{X|B_i} P(B_i) \quad \textit{a}$$

$$f_X(x) = \sum_i f_{X|B_i} P(B_i).$$

Důkaz: věta o úplné pravděpodobnosti. (Spec. případ byl na cvičení – dva algoritmy.)

Přehled

Kovariance

Definice

Pro n.v. X, Y definujeme jejich kovarianci (covariance) předpisem

$$cov(X,Y) = \mathbb{E}\big((X - \mathbb{E}X)(Y - \mathbb{E}Y)\big).$$

Věta

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

- $var(X) = cov(X,X)$
- $cov(X, aY + bZ + c) = a\,cov(X,Y) + b\,cov(X,Z)$
- $cov(X,Y) = 0$ pokud X,Y jsou nezávislé
- ale nejen tehdy

Korelace

Definice

Korelace náhodných veličin X, Y je definována předpisem

$$\varrho(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X)\,var(Y)}}.$$

- je to přenormovaná kovariance
- $-1 \leq \varrho(X,Y) \leq 1$ (cvič.)
- Korelace neznamená příčinnou souvislost! (Např., korelace je symetrická, kauzalita nikoli!)
- Naopak, nekorelace neznamená nezávislost. (Př: X libovolná, $Y = +X$ nebo $Y = -X$, obojí se stejnou pravděpodobností.)

Rozptyl součtu

Věta

Nechť $X = \sum_{i=1}^{n} X_i$. *Pak*

$$var(X) = \sum_{i=1}^{n}\sum_{j=1}^{n} cov(X_i, X_j) = \sum_{i=1}^{n} var(X_i) + \sum_{i\neq j} cov(X_i, X_j).$$

Spec. jsou-li $X_1, \ldots, X_n$ *nezávislé, pak*

$$var(X) = \sum_{i=1}^{n} var(X_i).$$

Přehled

Podmíněná hustota

Definice

Pro spojité n.v. X, Y definujeme podmíněnou hustotu (conditional pdf) předpisem

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

pokud je $f_Y(y) > 0$, jinak ji nedefinujeme.

- připomeňme, že $f_Y(y) = \int_{x \in \mathbb{R}} f_{X,Y}(x,y)dx$
- pro fixované y je $f_{X|Y}(x|y)$ hustota

Podmíněná, sdružená a marginální hustota

Věta

$$f_{X,Y}(x,y) = f_Y(y) f_{X|Y}(x|y)$$

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x|y) f_Y(y) dy$$

Podmíněná hustota a střední hodnota

- $F_{X|B}(x) := P(X \leq x \mid B)$ ... distr. funkce n.v. X zúžené na $B \subseteq \Omega$, potřebujeme $P(B) > 0$.
- $f_{X|B}$ odpovídající hustota
- pokud $B = \{X \in S\}$, tak

$$f_{X|B}(x) = \begin{cases} \frac{f_X(x)}{P(X \in S)} & \text{pokud } x \in S \\ 0 & \text{jinak} \end{cases}$$

- $\mathbb{E}\big(X \mid B\big) = \int_{-\infty}^{\infty} x \cdot f_{X|B}(x)dx$
- $\mathbb{E}\big(g(X)|B\big) = \int_{-\infty}^{\infty} g(x) f_{X|B}(x)dx$
- Pokud $B_1, B_2, \ldots$ je rozklad, tak

$$\mathbb{E}\big(X\big) = \sum_i \mathbb{E}\big(X \mid B_i\big) P(B_i).$$

Podmíněná hustota a střední hodnota

- $f_{X|Y}(x|y) := \frac{f_{X,Y}(x,y)}{f_Y(y)}$ je hustota n.v. X, pokud $Y = y$
- $\mathbb{E}\big(X \mid Y = y\big) := \int_{-\infty}^{\infty} x \cdot f_{X|Y}(x,y)dx$ je střední hodnota této veličiny
- $\mathbb{E}\big(g(X)|Y = y\big) = \int_{-\infty}^{\infty} g(x) \cdot f_{X|Y}(x,y)dx$
- $$\mathbb{E}\big(X\big) = \int_{-\infty}^{\infty} \mathbb{E}\big(X \mid Y = y\big) f_Y(y)dy$$
- $\mathbb{E}\big(X\big) = \mathbb{E}(\mathbb{E}\big(X \mid Y\big))$

Přehled

Cauchyho nerovnost

Věta

Nechť X, Y mají konečnou střední hodnotu a rozptyl. Pak

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

- Důsledek pro korelaci: $-1 \leq \varrho(X, Y) \leq 1$

Jensenova nerovnost

Věta

Nechť X má konečnou střední hodnotu a nechť g je konvexní reálná funkce. Pak

$$\mathbb{E}\big(g(X)\big) \geq g(\mathbb{E}\big(X\big)).$$

(Pro konkávní platí opačná nerovnost.)

Markovova nerovnost

Věta

Nechť náhodná veličina X spolňuje $X \geq 0$. Pak

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

Čebyševova (Chebyshev) nerovnost

Věta

Nechť X má konečnou střední hodnotu μ a rozptyl σ^2. Pak

$$P(|X - \mu| \geq a \cdot \sigma) \leq \frac{1}{a^2}.$$

Chernoffova (Černovova) nerovnost

Věta

Nechť $X = \sum_{i=1}^{n} X_i$, *kde* X_i *jsou n.n.v. nabývající hodnot* ± 1 *s pravděpodobností* $1/2$. *Pak pro* $t > 0$ *platí*

$$P(X \leq -t) = P(X \geq t) \leq e^{-t^2/2\sigma^2},$$

kde $\sigma = \sigma_X = \sqrt{n}$.

Bez dk.

Přehled

Silný zákon velkých čísel (strong law of large numbers)

Věta

Nechť $X_1, \ldots, X_n$ jsou n.n.v. se stř. hodnotou μ a rozptylem σ^2. Označme $S_n = (X_1 + \cdots + X_n)/n$ tzv. výběrový průměr (sample mean). Pak platí

$$\lim_{n\to\infty} S_n = \mu \quad \textit{skoro jistě (tj. s pravděpodobností 1).}$$

Říkáme, že posloupnost S_n konverguje k μ skoro jistě (almost surely).

Monte Carlo integration

Jak spočítat $\int_{x \in A} g(x) dx$?

Slabý zákon velkých čísel (weak law of large numbers)

Věta

Nechť $X_1, \ldots, X_n$ jsou n.n.v. se stř. hodnotou μ a rozptylem σ^2. Označme $S_n = (X_1 + \cdots + X_n)/n$. Pak pro každé $\varepsilon > 0$ platí

$$\lim_{n\to\infty} P(|S_n - \mu| > \varepsilon) = 0.$$

Říkáme, že posloupnost S_n konverguje k μ v pravděpodobnosti (in probability).

Centrální Limitní věta

Centrální Limitní věta

Věta

Nechť $X_1, \ldots, X_n$ jsou n.n.v. se střední hodnotou μ a rozptylem σ^2. Označme $Y_n = ((X_1 + \cdots + X_n) - n\mu)/(\sqrt{n} \cdot \sigma)$. Pak $Y_n \xrightarrow{d} N(0,1)$. Neboli, pokud F_n je distribuční funkce Y_n, tak

$$\lim_{n\to\infty} F_n(x) = \Phi(x) \quad \textit{for every } x \in \mathbb{R}.$$

Říkáme, že posloupnost Y_n konverguje k $N(0,1)$ v distribuci (in distribution).

Momentová vytvořující funkce

Definice

Pro náhodnou veličinu X označíme

$$M_X(t) = \mathbb{E}\left(e^{tX}\right).$$

Funkci $M_X(t)$ nazýváme momentová vytvořující funkce (moment generating function).

- $M_{Bern(p)}(t) = p \cdot e^t + (1-p)$.
- $M_X(t) = \sum_{n=0}^{\infty} \mathbb{E}\left(X^n\right) \frac{t^n}{n!}$.
- $M_{X+Y}(t) = M_X(t) M_Y(t)$, jsou-li X, Y n.n.v.
- $M_{Bin(n,p)} = (pe^t + 1 - p)^n$
- $M_{N(0,1)} = e^{t^2/2}$
- $M_{Exp(\lambda)} = \frac{1}{1-t/\lambda}$
- Pokud $M_X(t) = M_Y(t)$ na intervalu $(-a, a)$ pro nějaké $a > 0$, tak je $X = Y$ s.j.

NMAI059 Pravděpodobnost a statistika 1
8. přednáška

Robert Šámal

Přehled

Co už známe

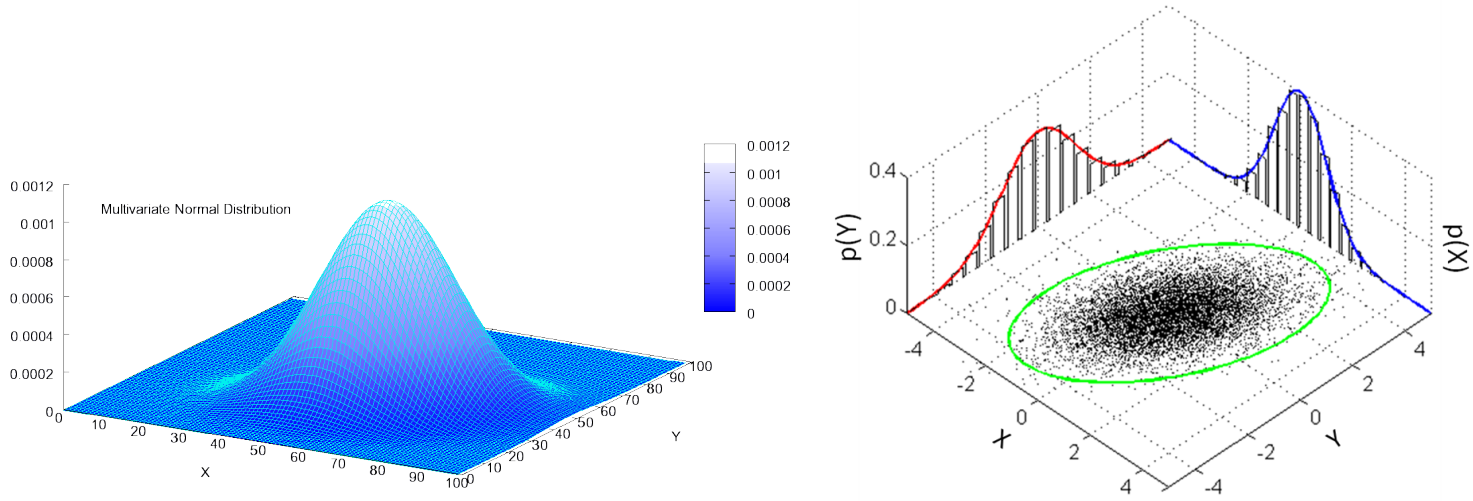
- sdružená distribuční funkce

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

- sdružená hustota: $f_{X,Y} \geq 0$ taková, že

$$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s,t)dtds.$$

- důležitý příklad: vícerozměrné normální rozdělení



Obrázek od editorů Wikipedie Piotrg a Bscan.

Podmiňování

Definice (zúžení náhodné veličiny na množinu)

X je n.v. na $(\Omega, \mathcal{F}, P)$, $B \in \mathcal{F}$, t.ž. $P(B) > 0$.

$$F_{X|B}(x) := P(X \leq x \mid B)$$

K tomu přísluší hustotní funkce $f_{X|B}$.

- pokud $B = \{X \in S\}$, tak

$$f_{X|B}(x) = \begin{cases} \frac{f_X(x)}{P(X \in S)} & \text{pokud } x \in S \\ 0 & \text{jinak} \end{cases}$$

Věta o rozkladu hustoty

Věta (věta o rozkladu hustoty)

Nechť X je spojitá n.v., nechť $B_1, B_2, \ldots$ je rozklad Ω. Pak

$$F_X(x) = \sum_i P(B_i) F_{X|B_i}(x) \quad \textit{a}$$

$$f_X(x) = \sum_i P(B_i) f_{X|B_i}(x).$$

Důkaz: věta o úplné pravděpodobnosti. (Spec. případ byl na cvičení – dva algoritmy.)

Marginální hustota

Věta

$$f_X(x) = \int_{y \in \mathbb{R}} f_{X,Y}(x, y) dy$$

$$f_Y(y) = \int_{x \in \mathbb{R}} f_{X,Y}(x, y) dx$$

Podmíněná hustota

Definice

Pro spojité n.v. X, Y definujeme podmíněnou hustotu (conditional pdf) předpisem

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

pokud je $f_Y(y) > 0$, jinak ji nedefinujeme.

- připomeňme, že $f_Y(y) = \int_{x \in \mathbb{R}} f_{X,Y}(x,y)dx$
- pro fixované y je $f_{X|Y}(x|y)$ hustota

Podmíněná, sdružená a marginální hustota

Věta

$$f_{X,Y}(x,y) = f_Y(y) f_{X|Y}(x|y)$$

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x|y) f_Y(y) dy$$

Součet spojitých n.v.

Věta

Nechť spojité X, Y jsou n.n.v. Pak $Z = X + Y$ je také spojitá n.v. a její hustotu dostaneme jako konvoluci funkcí f_X, f_Y, neboli

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) dx.$$

Ukázka konvoluce

Podmíněná hustota a střední hodnota

- $\mathbb{E}\big(X \mid B\big) := \int_{-\infty}^{\infty} x \cdot f_{X|B}(x)dx$
- $\mathbb{E}\big(g(X)|B\big) = \int_{-\infty}^{\infty} g(x) f_{X|B}(x)dx$

Věta (o úplné střední hodnotě)

Nechť X je spojitá n.v. Pokud $B_1, B_2, \ldots$ je rozklad, tak

$$\mathbb{E}\big(X\big) = \sum_i P(B_i)\mathbb{E}\big(X \mid B_i\big).$$

Důkaz: pomocí rozkladu hustoty.

Podmíněná hustota a střední hodnota

- $f_{X|Y}(x|y) := \frac{f_{X,Y}(x,y)}{f_Y(y)}$ je hustota n.v. X, pokud $Y = y$
- $\mathbb{E}\big(X \mid Y = y\big) := \int_{-\infty}^{\infty} x \cdot f_{X|Y}(x,y)dx$ je střední hodnota této veličiny
- $\mathbb{E}\big(g(X)|Y = y\big) = \int_{-\infty}^{\infty} g(x) \cdot f_{X|Y}(x,y)dx$
- Analogie věty o úplné střední hodnotě:

$$\mathbb{E}\big(X\big) = \int_{-\infty}^{\infty} \mathbb{E}\big(X \mid Y = y\big) f_Y(y)dy$$

- $\mathbb{E}\big(X\big) = \mathbb{E}(\mathbb{E}\big(X \mid Y\big))$

Přehled

Kovariance

Definice

Pro n.v. X, Y definujeme jejich kovarianci (covariance) předpisem

$$cov(X,Y) = \mathbb{E}\big((X - \mathbb{E}X)(Y - \mathbb{E}Y)\big).$$

Věta

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

- $var(X) = cov(X,X)$
- $cov(X, aY + bZ + c) = a\,cov(X,Y) + b\,cov(X,Z)$
- $cov(X,Y) = 0$ pokud X,Y jsou nezávislé
- ale nejen tehdy

Korelace

Definice

Korelace náhodných veličin X, Y je definována předpisem

$$\varrho(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X)\,var(Y)}}.$$

- je to přenormovaná kovariance
- $-1 \leq \varrho(X,Y) \leq 1$ (cvič.)
- Korelace neznamená příčinnou souvislost! (Např., korelace je symetrická, kauzalita nikoli!)
- Naopak, nekorelace neznamená nezávislost. (Př: X libovolná, $Y = +X$ nebo $Y = -X$, obojí se stejnou pravděpodobností.)

Rozptyl součtu

Věta

Nechť $X = \sum_{i=1}^{n} X_i$. *Pak*

$$var(X) = \sum_{i=1}^{n}\sum_{j=1}^{n} cov(X_i, X_j) = \sum_{i=1}^{n} var(X_i) + \sum_{i\neq j} cov(X_i, X_j).$$

Spec. jsou-li $X_1, \ldots, X_n$ *nezávislé, pak*

$$var(X) = \sum_{i=1}^{n} var(X_i).$$

Přehled

Cauchyho nerovnost

Věta

Nechť X, Y mají konečnou střední hodnotu a rozptyl. Pak

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

- Důsledek pro korelaci: $-1 \leq \varrho(X, Y) \leq 1$

Jensenova nerovnost

Věta

Nechť X má konečnou střední hodnotu a nechť g je konvexní reálná funkce. Pak

$$\mathbb{E}\big(g(X)\big) \geq g(\mathbb{E}\big(X\big)).$$

(Pro konkávní platí opačná nerovnost.)

Markovova nerovnost

Věta

Nechť náhodná veličina X splňuje $X \geq 0$. Pak

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

Čebyševova (Chebyshev) nerovnost

Věta

Nechť X má konečnou střední hodnotu μ a rozptyl σ^2. Pak

$$P(|X - \mu| \geq a \cdot \sigma) \leq \frac{1}{a^2}.$$

Chernoffova (Černovova) nerovnost

Věta

Nechť $X = \sum_{i=1}^{n} X_i$, *kde* X_i *jsou n.n.v. nabývající hodnot* ± 1 *s pravděpodobností* $1/2$. *Pak pro* $t > 0$ *platí*

$$P(X \leq -t) = P(X \geq t) \leq e^{-t^2/2\sigma^2},$$

kde $\sigma = \sigma_X = \sqrt{n}$.

Bez dk.

Přehled

Silný zákon velkých čísel (strong law of large numbers)

Věta

Nechť $X_1, \ldots, X_n$ jsou stejně rozdělené n.n.v. se stř. hodnotou μ a rozptylem σ^2. Označme $S_n = (X_1 + \cdots + X_n)/n$ tzv. výběrový průměr (sample mean). Pak platí

$$\lim_{n\to\infty} S_n = \mu \quad \textit{skoro jistě (tj. s pravděpodobností 1).}$$

Říkáme, že posloupnost S_n konverguje k μ skoro jistě (almost surely).

Monte Carlo integration

Jak spočítat $\int_{x \in A} g(x) dx$?
Speciálně:

$$g(x) = \begin{cases} 1 & \text{pro } x \in S \\ 0 & \text{jinak} \end{cases}$$

. . . obsah kruhu

Slabý zákon velkých čísel (weak law of large numbers)

Věta

Nechť $X_1, \ldots, X_n$ jsou stejně rozdělené n.n.v. se stř. hodnotou μ a rozptylem σ^2. Označme $S_n = (X_1 + \cdots + X_n)/n$. Pak pro každé $\varepsilon > 0$ platí

$$\lim_{n\to\infty} P(|S_n - \mu| > \varepsilon) = 0.$$

Říkáme, že posloupnost S_n konverguje k μ v pravděpodobnosti (in probability).

Centrální Limitní věta

Centrální Limitní věta

Věta

Nechť $X_1, \ldots, X_n$ jsou stejně rozdělené n.n.v. se střední hodnotou μ a rozptylem σ^2. Označme
$Y_n = ((X_1 + \cdots + X_n) - n\mu)/(\sqrt{n} \cdot \sigma)$.
Pak $Y_n \xrightarrow{d} N(0,1)$. Neboli, pokud F_n je distribuční funkce Y_n, tak

$$\lim_{n\to\infty} F_n(x) = \Phi(x) \quad \textit{pro každé } x \in \mathbb{R}.$$

Říkáme, že posloupnost Y_n konverguje k $N(0,1)$ v distribuci (in distribution).

Momentová vytvořující funkce

Definice

Pro náhodnou veličinu X označíme

$$M_X(t) = \mathbb{E}\left(e^{tX}\right).$$

Funkci $M_X(t)$ nazýváme momentová vytvořující funkce (moment generating function).

- $M_{Bern(p)}(t) = p \cdot e^t + (1-p)$.
- $M_X(t) = \sum_{n=0}^{\infty} \mathbb{E}\left(X^n\right) \frac{t^n}{n!}$.
- $M_{X+Y}(t) = M_X(t) M_Y(t)$, jsou-li X, Y n.n.v.
- $M_{Bin(n,p)} = (pe^t + 1 - p)^n$
- $M_{N(0,1)} = e^{t^2/2}$
- $M_{Exp(\lambda)} = \frac{1}{1-t/\lambda}$
- Pokud $M_X(t) = M_Y(t)$ na intervalu $(-a, a)$ pro nějaké $a > 0$, tak je $X = Y$ s.j.

NMAI059 Pravděpodobnost a statistika 1
9. přednáška

Robert Šámal

Nerovnosti, které známe z minula

- Markov:

$$X \geq 0 \Rightarrow P(X \geq a\mathbb{E}(X)) \leq \frac{1}{a}$$

- Čebyšev/Chebyshev

$$P(|X - \mathbb{E}(X)| \geq a\sigma_X) \leq \frac{1}{a^2}$$

- Chernoff ($\sigma_X = \sqrt{n}$)

$$X = \sum_{i=1}^{n} X_i, X_i = \pm 1 \Rightarrow P(|X - \mathbb{E}(X)| > a\sigma_X) \leq 2e^{-a^2/2}$$

Přehled

Slabý zákon velkých čísel (weak law of large numbers)

Věta

Nechť $X_1, \ldots, X_n$ jsou stejně rozdělené n.n.v. se stř. hodnotou μ a rozptylem σ^2. Označme $S_n = (X_1 + \cdots + X_n)/n$. Pak pro každé $\varepsilon > 0$ platí

$$\lim_{n\to\infty} P(|S_n - \mu| \geq \varepsilon) = 0.$$

Říkáme, že posloupnost S_n konverguje k μ v pravděpodobnosti (in probability), píšeme $S_n \xrightarrow{P} \mu$).

SZVČ $\rightarrow$ Centrální Limitní věta

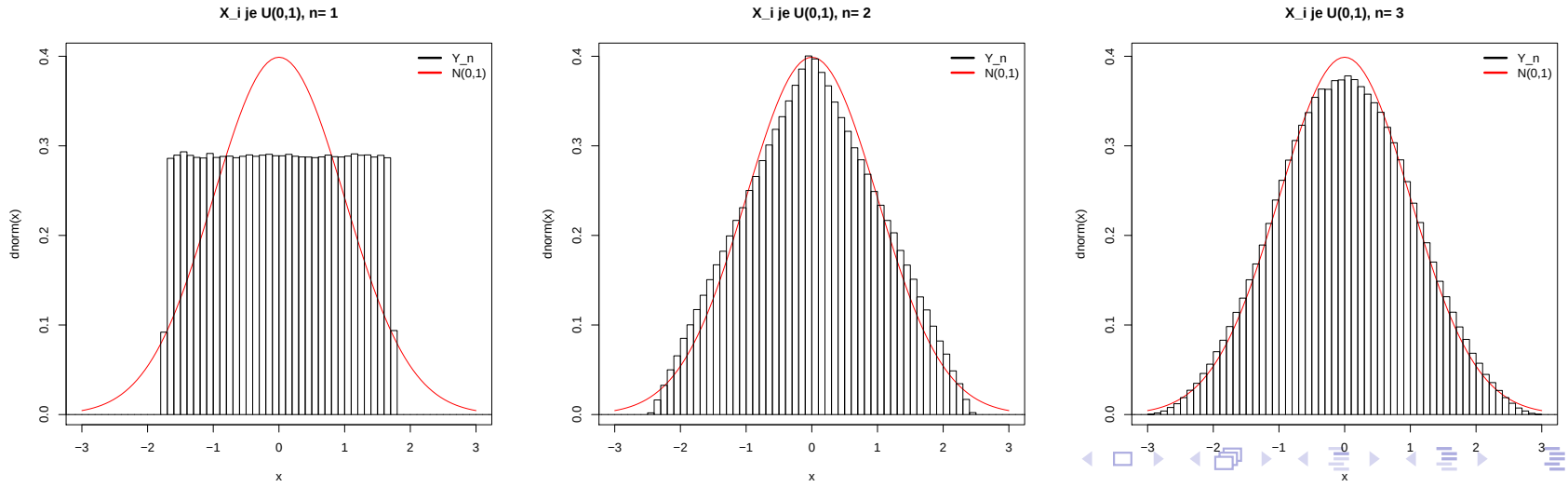
Centrální Limitní věta

Věta

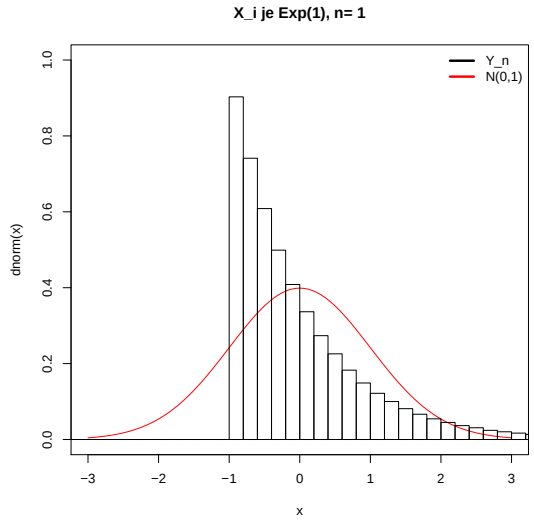
Nechť $X_1, \ldots, X_n$ jsou stejně rozdělené n.n.v. se střední hodnotou μ a rozptylem σ^2. Označme
$Y_n = ((X_1 + \cdots + X_n) - n\mu)/(\sqrt{n} \cdot \sigma)$.
Pak $Y_n \xrightarrow{d} N(0,1)$. Neboli, pokud F_n je distribuční funkce Y_n, tak

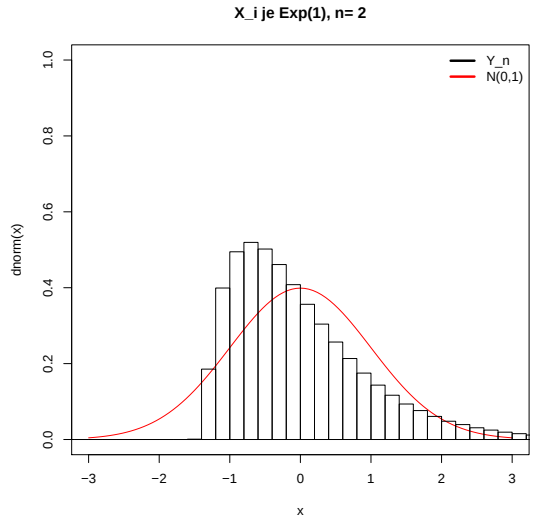
$$\lim_{n\to\infty} F_n(x) = \Phi(x) \quad \textit{pro každé } x \in \mathbb{R}.$$

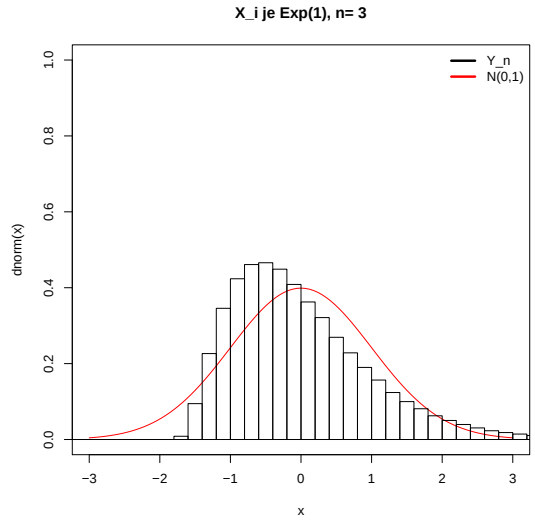
Říkáme, že posloupnost Y_n konverguje k $N(0,1)$ v distribuci (in distribution).

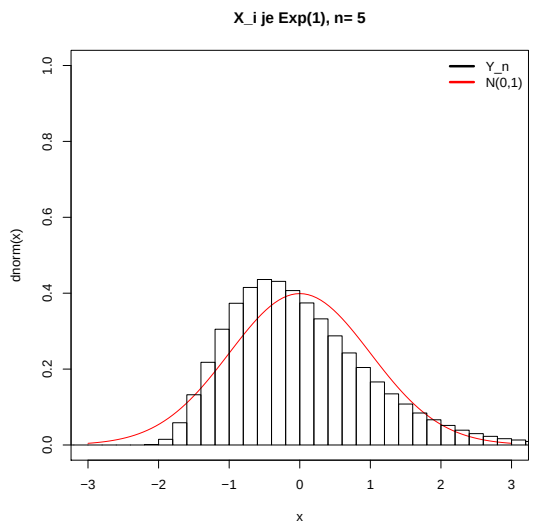


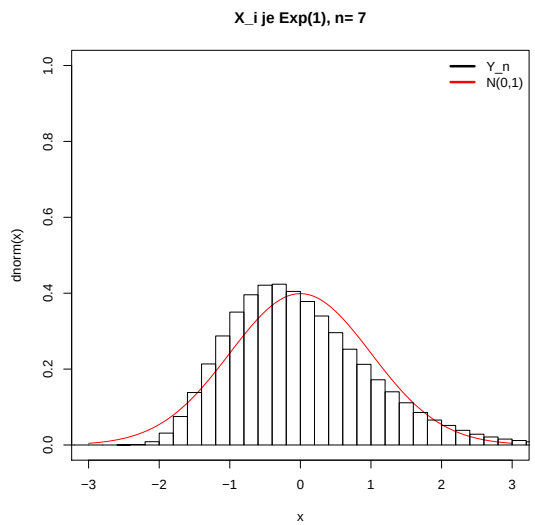
CLV další ukázka

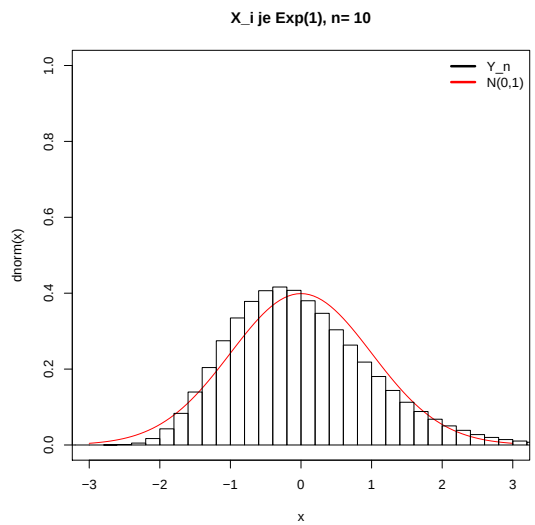












Bonus: Momentová vytvořující funkce

Definice

Pro náhodnou veličinu X označíme

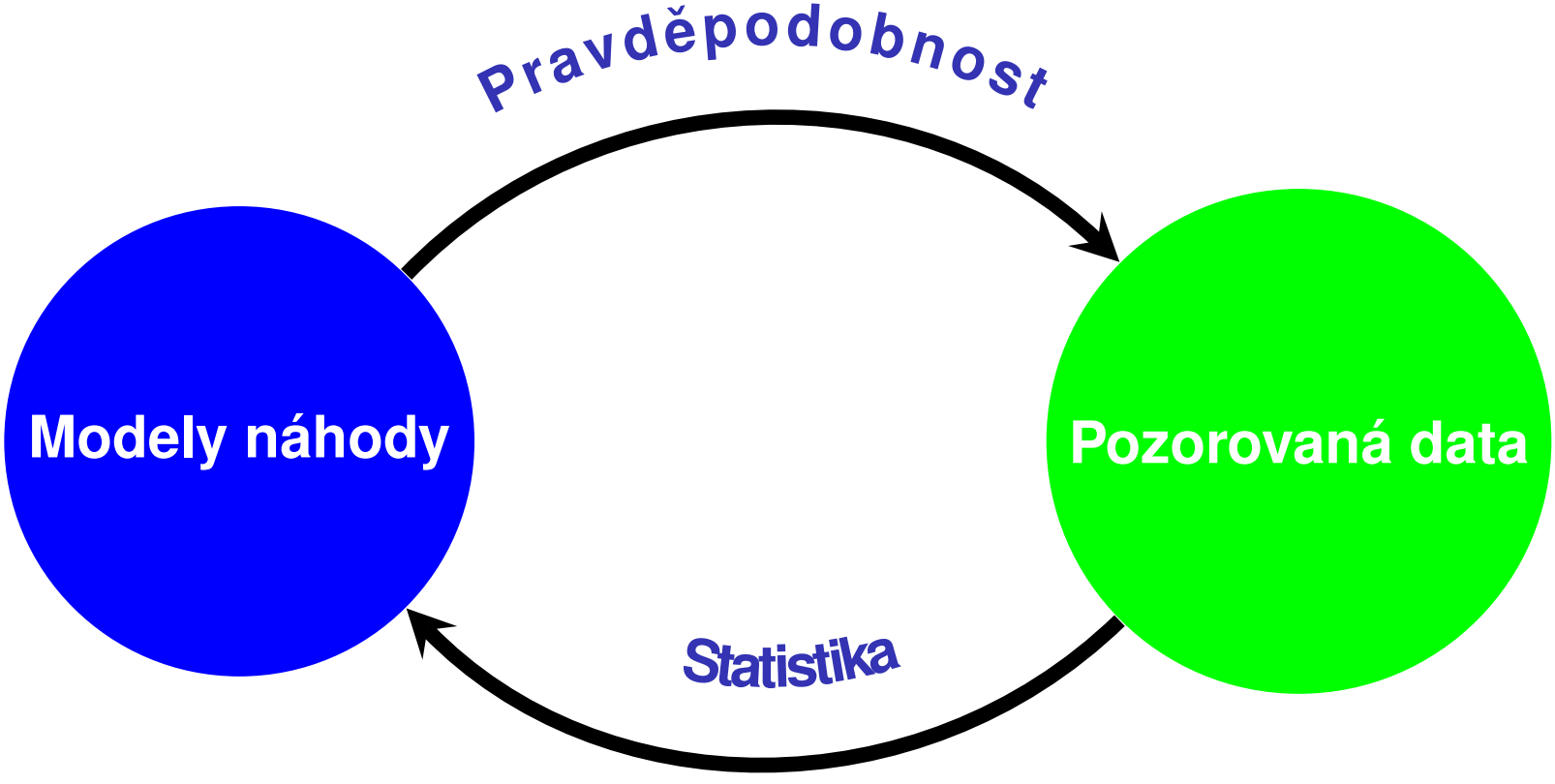
$$M_X(t) = \mathbb{E}\left(e^{tX}\right).$$

Funkci $M_X(t)$ nazýváme momentová vytvořující funkce (moment generating function).

- $M_X(t) = \sum_{n=0}^{\infty} \mathbb{E}\left(X^n\right) \frac{t^n}{n!}$.
- $M_{Bern(p)}(t) = p \cdot e^t + (1-p)$.
- $M_{X+Y}(t) = M_X(t) M_Y(t)$, jsou-li X, Y n.n.v.
- $M_{Bin(n,p)} = (pe^t + 1 - p)^n$
- $M_{N(0,1)} = e^{t^2/2}$
- $M_{Exp(\lambda)} = \frac{1}{1-t/\lambda}$
- Pokud $M_X(t) = M_Y(t)$ na intervalu $(-a, a)$ pro nějaké $a > 0$, tak je $X = Y$ s.j.

Přehled

Plán přednášky



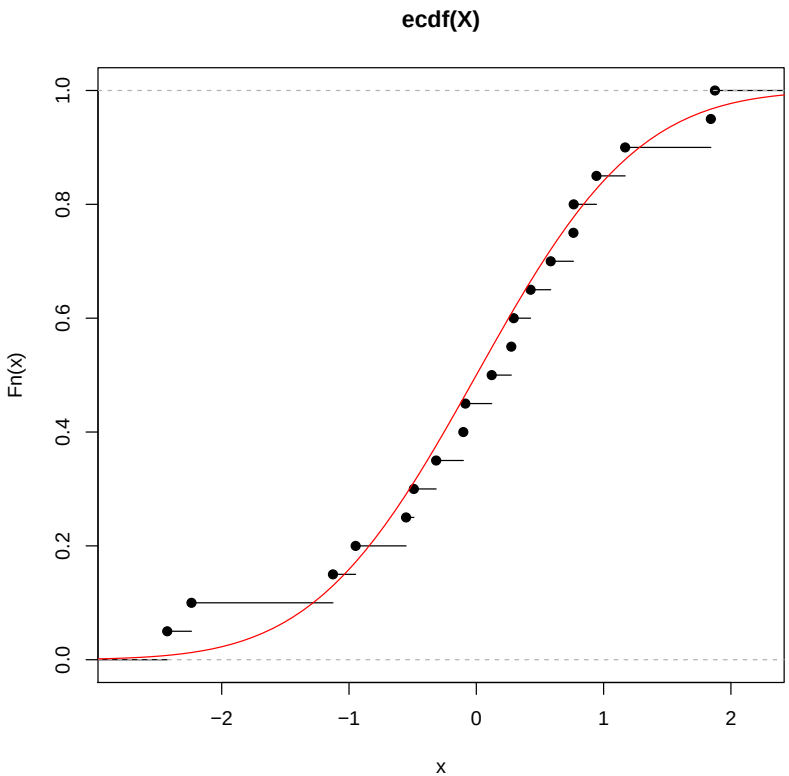
1. ilustrace – počet leváků

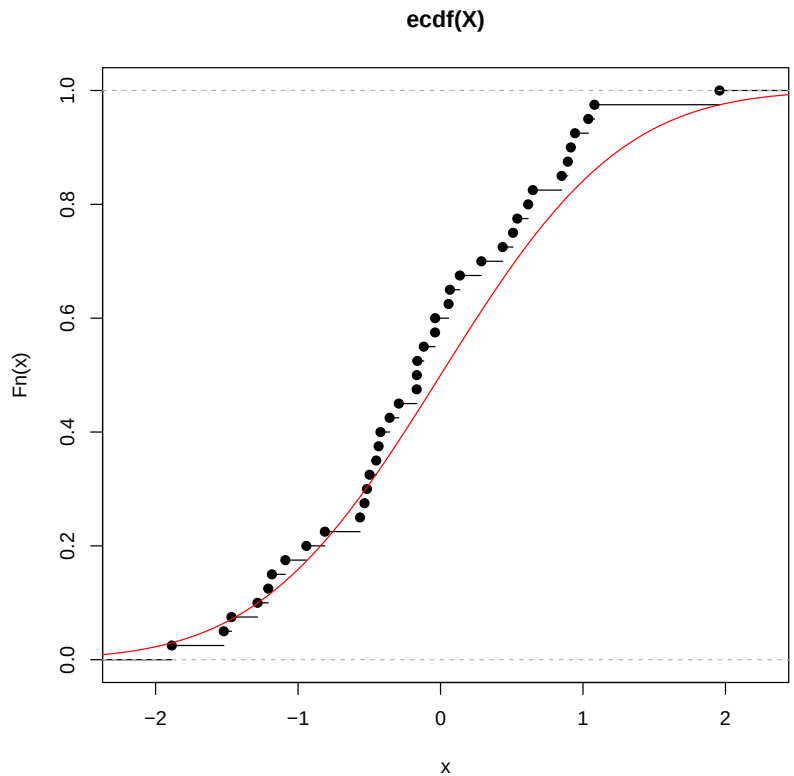
2. ilustrace – doba běhu programu

- $X_1, \ldots, X_n \sim F$ n.n.v., F je jejich distribuční funkce
- **Definice:** *Empirická distribuční funkce (empirical CDF)* je definována

$$\widehat{F}_n(x) = \frac{\sum_{i=1}^{n} I(X_i \leq x)}{n},$$

kde $I(X_i \leq x) = 1$ pokud $X_i \leq x$ a 0 jinak.





Empirická distribuční funkce – vlastnosti

Věta

Pro pevné x platí

- $\mathbb{E}\big(\widehat{F}_n(x)\big) = F(x)$
- $var(\widehat{F}_n(x)) = \frac{F(x)(1-F(x))}{n}$
- $\widehat{F}_n(x)$ *konverguje k* $F(x)$ *v pravděpodobnosti, píšeme* $\widehat{F}_n(x) \xrightarrow{P} F(x)$.

Důkaz.

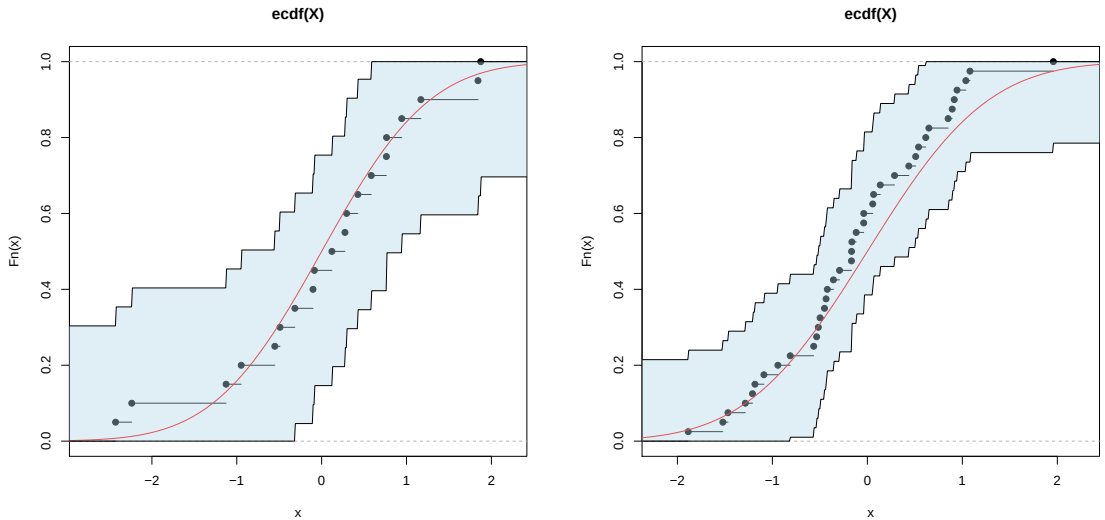
Slabý zákon velkých čísel. □

Empirická distribuční funkce – Dvoretzky-Kiefer-Wolfowitz (DKW)

Věta

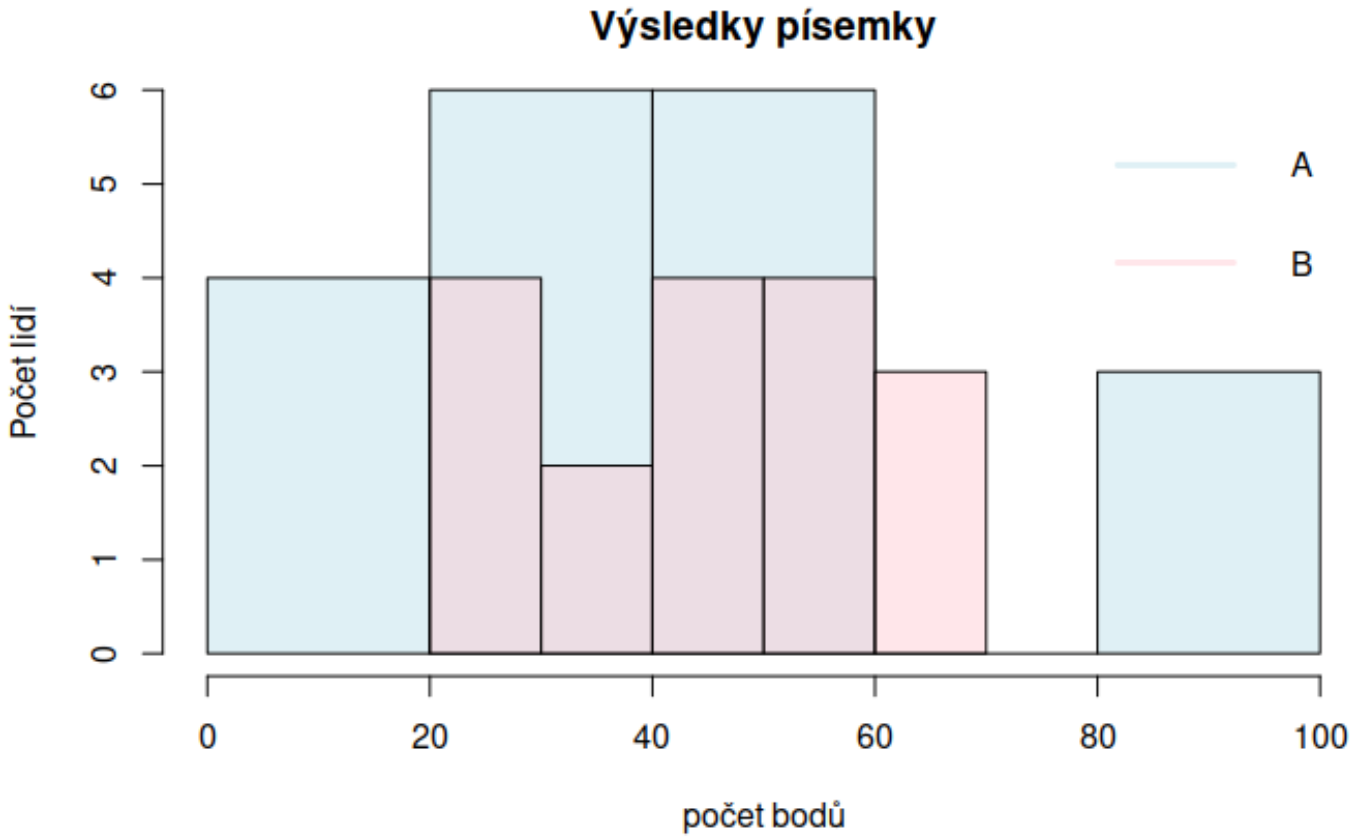
Nechť $X_1, \ldots, X_n \sim F$ *jsou n.n.v.,* $\widehat{F}_n$ *jejich empirická distribuční funkce. Nechť* $\mathbb{E}(X_i)$ *je konečná. Zvolme* $\alpha \in (0,1)$ *a označme* $\varepsilon = \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}}$. *Pak platí*

$$P(\widehat{F}_n(x) - \varepsilon \leq F(x) \leq \widehat{F}_n(x) + \varepsilon) \geq 1 - \alpha.$$



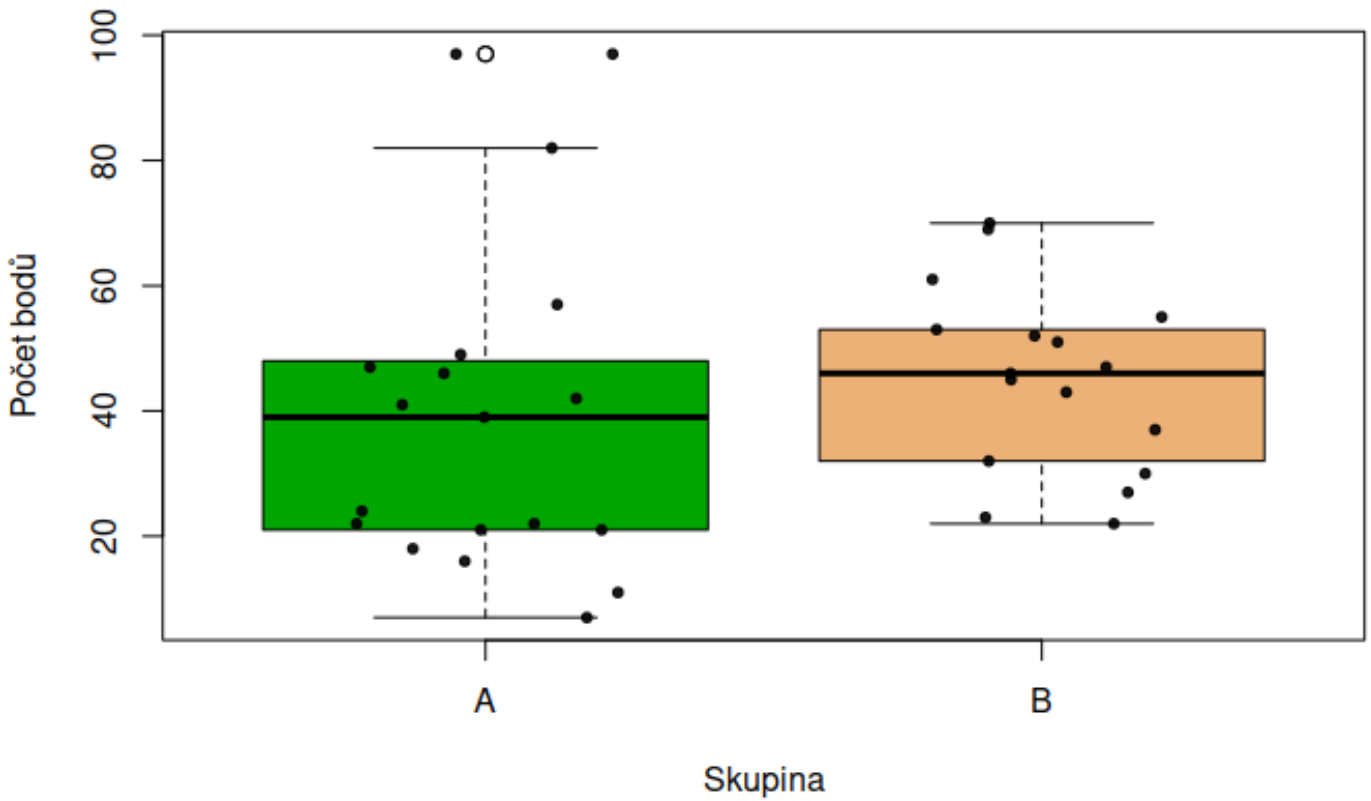
Intro – explorační analýza dat (exploratory data analysis)

- posbíráme data (a dáme pozor na systémové chyby – nezávislost, nezaujatost, . . .)
- různé tabulky (třeba v Excelu a spol.)
- vhodné obrázky: histogram, krabicový diagram (boxplot), atd.



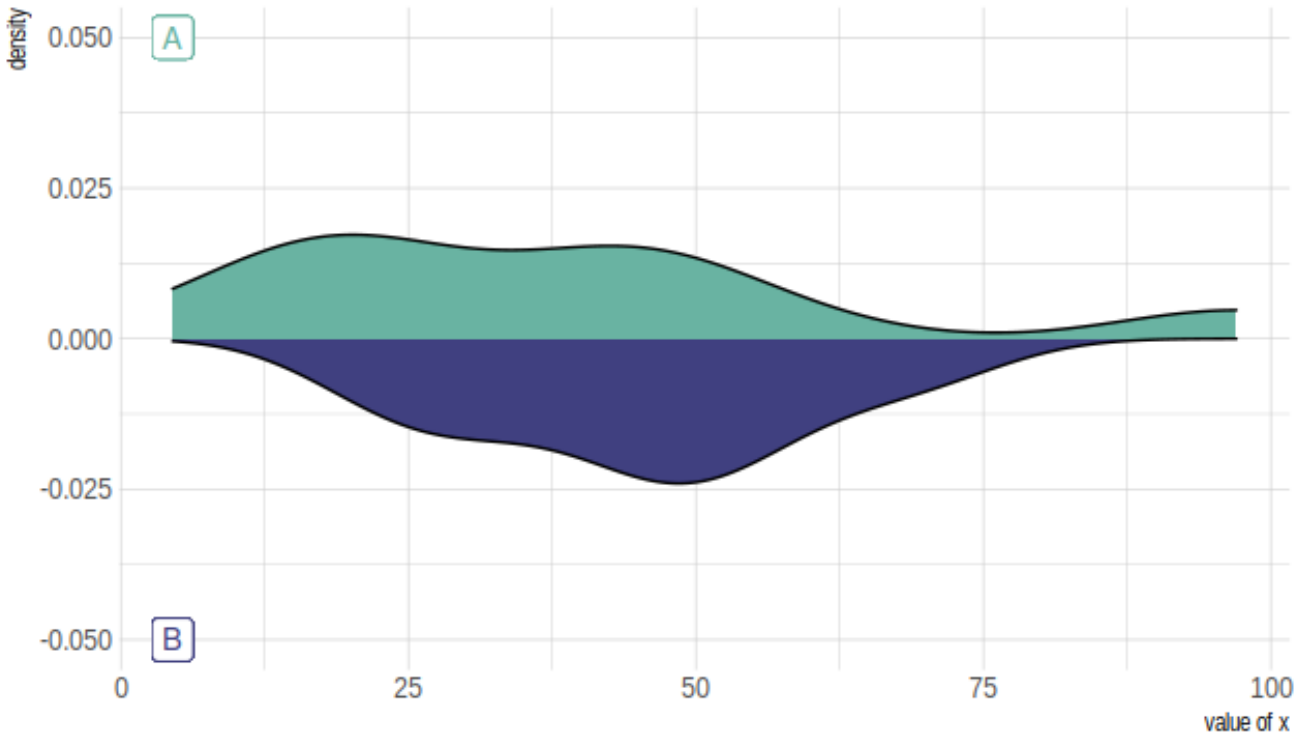
Intro – explorační analýza dat (exploratory data analysis)

- posbíráme data (a dáme pozor na systémové chyby – nezávislost, nezaujatost, . . .)
- různé tabulky (třeba v Excelu a spol.)
- vhodné obrázky: histogram, krabicový diagram (boxplot), atd.



Intro – explorační analýza dat (exploratory data analysis)

- posbíráme data (a dáme pozor na systémové chyby – nezávislost, nezaujatost, . . .)
- různé tabulky (třeba v Excelu a spol.)
- vhodné obrázky: histogram, krabicový diagram (boxplot), atd.



Náhodný výběr

- s vracením
- bez vracení

$\Omega = \{$všechny n-tice obyvatel ČR$\}$
Pro $\omega = (\omega_1, \ldots, \omega_n)$ zvolíme $X_i = I(\omega_i$ je levák$)$.

Statistika – přehled

- nezávislá měření – hodnoty n.n.v. $X_1, \ldots, X_n \sim F$
  náhodný výběr s distribuční funkcí F s rozsahem n

- neparametrické modely: povolujeme velkou třídu F

- parametrické modely: $F \in \{F_\vartheta : \vartheta \in \Theta\}$
- příklady:
  - $Pois(\lambda)$ (parametr $\vartheta = \lambda$, $\Theta = \mathbb{R}^+$)
  - $U(a, b)$ (parametr $\vartheta = (a, b)$, $\Theta = \mathbb{R}^2$)
  - $N(\mu, \sigma^2)$ (parametr $\vartheta = (\mu, \sigma)$, $\Theta = \mathbb{R} \times \mathbb{R}^+$)

- „Všechny modely jsou špatné, ale některé jsou užitečné.“
  (George Box)

Zkoumané úlohy – cíle konfirmační analýzy (confirmatory data analysis)

- bodové odhady
- intervalové odhady
- testování hypotéz
- (lineární) regrese

- *statistika* – libovolná funkce náhodného výběru, tj. např. aritmetický průměr, medián, maximum, atd.
  Tj. $T = T(X_1, \ldots, X_n)$.

Další typy zkoumaných problémů

- Je zkoumaný lék účinný?
- Je naše mince, kostka spravedlivá?
- Předpokládáme, že výška člověka je normálně rozdělená. Jaké je μ, σ?
- Předpokládáme, že výška člověka je normálně rozdělená. V jakém vztahu je průměrná výška mužů a žen? Praváků a leváků?
- Jak závisí náklon šikmé věže v Pise na čase?

Zkoumané úlohy – předpoklady

- Vždy předpokládáme, že máme nezávislá měření – hodnoty n.n.v. $X_1, \ldots, X_n \sim F$
- O F předpokádáme, že patří do nějakého *modelu* – množiny vhodných distr. funkcí.
- parametrické/neparametrické modely

Zkoumané úlohy – cíle

- bodové odhady
- intervalové odhady
- testování hypotéz
- (lineární) regrese

Přehled

Výběrový průměr a rozptyl

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\bar{S}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

$$\widehat{S}_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

Cíle

Definice

Odhad $\hat{\Theta}_n = \hat{\Theta}_n(X_1, \ldots, X_n)$ *parametru* ϑ *je*

- *nestranný (unbiased) – pokud* $\vartheta = \mathbb{E}\big(\hat{\Theta}_n\big)$
- *asymptoticky nestranný (asymptotically unbiased) – pokud* $\vartheta = \lim_{n\to\infty} \mathbb{E}\big(\hat{\Theta}_n\big)$
- *vychýlení (bias)* $bias_\vartheta(\hat{\Theta}_n) = \mathbb{E}\big(\hat{\Theta}_n\big) - \vartheta$
- *střední kvadratická chyba (mean squared error, MSE) je* $\mathbb{E}\big((\hat{\Theta} - \vartheta)^2\big)$

Věta

$MSE = bias_\vartheta(\hat{\Theta}_n)^2 + var_\vartheta(\hat{\Theta}_n)$

Parametry výběrového momentu a rozptylu

Věta

1. $\bar{X}_n$ *je konzistentní nestranný odhad* μ
2. $\bar{S}_n$ *je konzistentní asymptoticky nestranný odhad* μ
3. $\widehat{S}_n$ *je konzistentní nestranný odhad* μ

Metoda momentů

- $m_r(\vartheta) := \mathbb{E}\left(X^r\right)$ pro $X \sim F_\vartheta$ … r-tý moment
- $\widehat{m_r(\vartheta)} := \frac{1}{n}\sum_{i=1}^{n} X_i^r$ pro náhodný výběr $X_1, \ldots, X_n$ z F_ϑ … r-tý výběrový moment

Věta

$\widehat{m_r(\vartheta)}$ je nestranný konzistentní odhad pro $m_r(\vartheta)$

- Odhad metodou momentů je řešení soustavy rovnic

$$m_r(\vartheta) = \widehat{m_r(\vartheta)} \qquad r = 1, \ldots, k.$$

Metoda momentů – příklady

Metoda maximální věrohodnosti (maximal likelihood, ML)

- náh. výběr $X = (X_1, \ldots, X_n)$ z modelu s parametrem ϑ
- možný výsledek $x = (x_1, \ldots, x_n)$
- ... sdružená pravděpodobnostní funkce $p_X(x; \vartheta)$
- ... sdružená hustota $f_X(x; \vartheta)$
- *věrohodnost (likelihood)* $L(x; \vartheta)$ značí p_X nebo f_X
- normálně: máme pevné ϑ, a $L(x; \vartheta)$ je funkce x
- teď: máme pevné x a $L(x; \vartheta)$ je funkce ϑ

Metoda MV (ML):
volíme takové ϑ, pro které je $L(x; \vartheta)$ maximální

Metoda maximální věrohodnosti (maximal likelihood, ML)

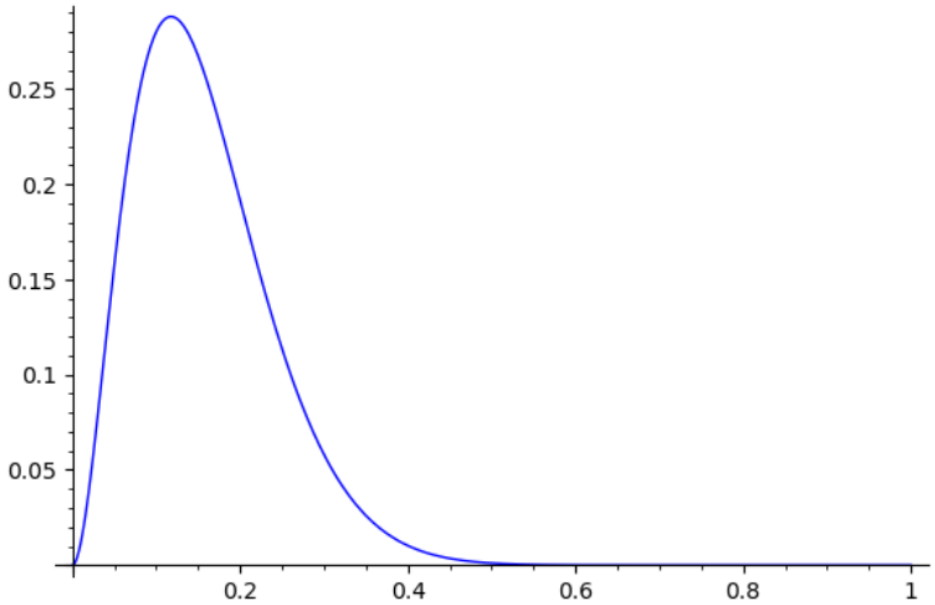
- **Metoda MV (ML):**
  volíme takové ϑ, pro které je $L(x;\vartheta)$ maximální
- definujeme také $\ell(x;\vartheta) = \log(L(x;\vartheta))$
- díky nezávislosti je

$L(x;\vartheta) =$

$\ell(x;\vartheta) =$

ML – leváci

```
plot(binomial(17,2)*p^2*(1-p)^15, [0,1])
```



Přehled

Intervalové odhady

- místo jednoho čísla s nejistým významem vypočítáme z dat interval $[\hat{\Theta}^-, \hat{\Theta}^+]$

Definice

Nechť $\hat{\Theta}^-$, $\hat{\Theta}^+$ *jsou n.v. které závisí na náhodném výběru* $X = (X_1, \ldots, X_n)$. *Tyto n.v. určují intervalový odhad, též konfidenční interval o spolehlivosti* $1 - \alpha$ *(*$1 - \alpha$ *confidence interval), pokud*

$$P(\hat{\Theta}^- \leq \vartheta \leq \hat{\Theta}^+) \geq 1 - \alpha.$$

Intervalové odhady normální náhodné veličiny

NMAI059 Pravděpodobnost a statistika 1
10. přednáška

Robert Šámal

Přehled

Náhodný výběr

- bez vracení
  $\Omega = \{$všechny n-tice obyvatel ČR$\}$
  Pro $\omega = (\omega_1, \ldots, \omega_n)$ zvolíme $X_i = I(\omega_i$ je levák$)$.
- s vracením
  $\Omega = \{$všechny n-tice obyvatel ČR, mohou se opakovat$\}$
  Pro $\omega = (\omega_1, \ldots, \omega_n)$ zvolíme $X_i = I(\omega_i$ je levák$)$.
- varianty (stratifikovaný výběr)
  Chceme adekvátně reprezentovat různé podmnožiny (dané věkem, bydlištěm, ...).
  Nebudeme dále zkoumat.

Statistika – model

- nezávislá měření – hodnoty n.n.v. $X_1, \ldots, X_n \sim F$
  náhodný výběr s distribuční funkcí F s rozsahem n

- neparametrické modely: povolujeme velkou třídu F

- parametrické modely: $F \in \{F_\vartheta : \vartheta \in \Theta\}$
- příklady:
  - $Pois(\lambda)$ (parametr $\vartheta = \lambda$, $\Theta = \mathbb{R}^+$)
  - $U(a, b)$ (parametr $\vartheta = (a, b)$, $\Theta = \mathbb{R}^2$)
  - $N(\mu, \sigma^2)$ (parametr $\vartheta = (\mu, \sigma)$, $\Theta = \mathbb{R} \times \mathbb{R}^+$)

- „Všechny modely jsou špatné, ale některé jsou užitečné.“
  (George Box)

Zkoumané úlohy – cíle konfirmační analýzy (confirmatory data analysis)

- bodové odhady
- intervalové odhady
- testování hypotéz
- (lineární) regrese

- *statistika* – libovolná funkce náhodného výběru, tj. např. aritmetický průměr, medián, maximum, atd.
  Tj. $T = T(X_1, \ldots, X_n)$.

Přehled

Výběrový průměr a rozptyl

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\bar{S}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

$$\widehat{S}_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

Odhad

Definice

Odhad je libovolná statistika.

Vlastnosti bodových odhadů

Definice

Odhad $\hat{\Theta}_n = \hat{\Theta}_n(X_1, \ldots, X_n)$ *parametru* ϑ *je*

- *nestranný (unbiased) – pokud* $\vartheta = \mathbb{E}\big(\hat{\Theta}_n\big)$ *(pro každé* ϑ*)*
- *asymptoticky nestranný (asymptotically unbiased) – pokud* $\vartheta = \lim_{n\to\infty} \mathbb{E}\big(\hat{\Theta}_n\big)$
- *konzistentní (consistent) – pokud* $\hat{\Theta}_n \xrightarrow{P} \vartheta$*.*
- *vychýlení (bias)* $bias_\vartheta(\hat{\Theta}_n) := \mathbb{E}\big(\hat{\Theta}_n\big) - \vartheta$
- *střední kvadratická chyba (mean squared error, MSE) je* $MSE := \mathbb{E}\big((\hat{\Theta}_n - \vartheta)^2\big)$

Věta

$MSE = bias_\vartheta(\hat{\Theta}_n)^2 + var_\vartheta(\hat{\Theta}_n)$

Parametry výběrového momentu a rozptylu

Věta

1. $\bar{X}_n$ *je konzistentní nestranný odhad* μ
2. $\bar{S}_n$ *je konzistentní asymptoticky nestranný odhad* μ
3. $\widehat{S}_n$ *je konzistentní nestranný odhad* μ

Metoda momentů

- $m_r(\vartheta) := \mathbb{E}\left(X^r\right)$ pro $X \sim F_\vartheta$ … r-tý moment
- $\widehat{m_r(\vartheta)} := \frac{1}{n}\sum_{i=1}^{n} X_i^r$ pro náhodný výběr $X_1, \ldots, X_n$ z F_ϑ … r-tý výběrový moment

Věta

$\widehat{m_r(\vartheta)}$ je nestranný konzistentní odhad pro $m_r(\vartheta)$

- Odhad metodou momentů je řešení soustavy rovnic

$$m_r(\vartheta) = \widehat{m_r(\vartheta)} \qquad r = 1, \ldots, k.$$

Metoda momentů – příklady

Metoda maximální věrohodnosti (maximal likelihood, ML)

- náh. výběr $X = (X_1, \ldots, X_n)$ z modelu s parametrem ϑ
- možný výsledek $x = (x_1, \ldots, x_n)$
- ... sdružená pravděpodobnostní funkce $p_X(x; \vartheta)$
- ... sdružená hustota $f_X(x; \vartheta)$
- *věrohodnost (likelihood)* $L(x; \vartheta)$ značí p_X nebo f_X
- normálně: máme pevné ϑ, a $L(x; \vartheta)$ je funkce x
- teď: máme pevné x a $L(x; \vartheta)$ je funkce ϑ

Metoda MV (ML):
volíme takové ϑ, pro které je $L(x; \vartheta)$ maximální

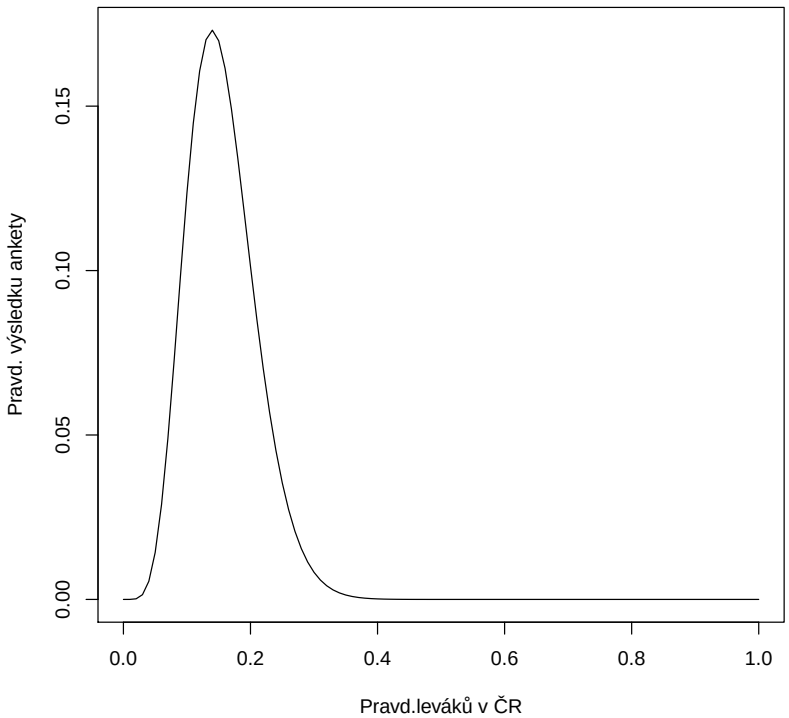
Metoda maximální věrohodnosti (maximal likelihood, ML)

- **Metoda MV (ML):**
  volíme takové ϑ, pro které je $L(x;\vartheta)$ maximální
- definujeme také $\ell(x;\vartheta) = \log(L(x;\vartheta))$
- díky nezávislosti je

$L(x;\vartheta) =$

$\ell(x;\vartheta) =$

ML – leváci


Pravd. výsledku ankety
0.00
0.05
0.10
0.15
0.0
0.2
0.4
0.6
0.8
1.0
Pravd.leváků v ČR

Přehled

Intervalové odhady

- místo jednoho čísla s nejistým významem vypočítáme z dat interval $[\hat{\Theta}^-, \hat{\Theta}^+]$

Definice

Nechť $\hat{\Theta}^-$, $\hat{\Theta}^+$ *jsou n.v. které závisí na náhodném výběru* $X = (X_1, \ldots, X_n)$. *Tyto n.v. určují intervalový odhad, též konfidenční interval o spolehlivosti* $1 - \alpha$ *(*$1 - \alpha$ *confidence interval), pokud*

$$P(\hat{\Theta}^- \leq \vartheta \leq \hat{\Theta}^+) \geq 1 - \alpha.$$

Intervalové odhady normální náhodné veličiny

Věta

$X_1, \ldots, X_n$ je náhodný výběr z $N(\vartheta, \sigma^2)$.
σ známe, ϑ chceme určit, $\alpha \in (0,1)$.
Nechť $\Phi(z_{\alpha/2}) = 1 - \alpha/2$. $\hat{\Theta}_n = \bar{X}_n$.

$$C_n := [\hat{\Theta}_n - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \hat{\Theta}_n + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}]$$

Pak $P(C_n \ni \vartheta) = 1 - \alpha$.

Důkaz.

Intervalové odhady pomocí CLV

Věta

$X_1, \ldots, X_n$ je náhodný výběr z rozdělení se střední hodnotou ϑ, rozptylem σ^2.
σ známe, ϑ chceme určit, $\alpha \in (0,1)$.
Nechť $\Phi(z_{\alpha/2}) = 1 - \alpha/2$.

$$C_n := [\hat{\Theta}_n - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \hat{\Theta}_n + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}]$$

Pak $P(C_n \ni \vartheta)$ se limitně blíží $1 - \alpha$.

Studentovo rozdělení

- $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ ... výběrový průměr
- $\widehat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ ... výběrový rozptyl

- Nechť $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$
- Pak $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$
- *Studentovo t-rozdělení s $n-1$ stupni volnosti* je rozdělení n.v.

$$\frac{\bar{X}_n - \mu}{\widehat{S}_n/\sqrt{n}}$$

- Distribuční funkce Ψ_{n-1} (v tabulkách ...)

Int. odhady normální n.v. pomocí Studentova t

Věta

$X_1, \ldots, X_n$ je náhodný výběr z $N(\vartheta, \sigma^2)$.
ϑ chceme určit, σ neznáme; $\alpha \in (0,1)$. Nechť
$\Psi_{n-1}(z_{\alpha/2}) = 1 - \alpha/2$. $\hat{\Theta}_n = \bar{X}_n$, $\widehat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$

$$C_n := [\hat{\Theta}_n - z_{\alpha/2}\frac{\widehat{S}_n}{\sqrt{n}}, \hat{\Theta}_n + z_{\alpha/2}\frac{\widehat{S}_n}{\sqrt{n}}]$$

Pak $P(C_n \ni \vartheta) = 1 - \alpha$.

NMAI059 Pravděpodobnost a statistika 1
11. přednáška

Robert Šámal

Přehled

Statistika – bodové odhady (point estimation)

Statistika – intervalové odhady

Testování hypotéz

Výběrový průměr a rozptyl

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\bar{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

$$\widehat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

Metoda maximální věrohodnosti (maximal likelihood, ML)

- **Metoda MV (ML):**
  volíme takové ϑ, pro které je $L(x;\vartheta)$ maximální
- definujeme také $\ell(x;\vartheta) = \log(L(x;\vartheta))$
- díky nezávislosti je

$$L(x;\vartheta) = L(x_1;\vartheta)\dots L(x_n;\vartheta)$$
$$\ell(x;\vartheta) = \ell(x_1;\vartheta) + \dots + \ell(x_n;\vartheta)$$

Bin(20,p)	**0.2**	**0.3**	**0.4**	**0.45**	**0.5**	**0.55**	**0.6**
7	0.0545	0.1643	0.1659	0.1221	0.0739	0.0366	0.0146
8	0.0222	0.1144	0.1797	0.1623	0.1201	0.0727	0.0355
9	0.0074	0.0654	0.1597	0.1771	0.1602	0.1185	0.071
10	0.002	0.0308	0.1171	0.1593	0.1762	0.1593	0.1171
11	0.0005	0.012	0.071	0.1185	0.1602	0.1771	0.1597
12	0.0001	0.0039	0.0355	0.0727	0.1201	0.1623	0.1797
13	0	0.001	0.0146	0.0366	0.0739	0.1221	0.1659
14	0	0.0002	0.0049	0.015	0.037	0.0746	0.1244

ML – další ilustrace

Přehled

Intervalové odhady (interval estimation)

- místo jednoho čísla s nejistým významem vypočítáme z dat interval $[\hat{\Theta}^-, \hat{\Theta}^+]$

Definice

Nechť $\hat{\Theta}^-$, $\hat{\Theta}^+$ jsou n.v. které závisí na náhodném výběru $X = (X_1, \ldots, X_n)$ z distribuce F_ϑ. Tyto n.v. určují intervalový odhad, též konfidenční interval o spolehlivosti $1 - \alpha$ ($1 - \alpha$ confidence interval), pokud

$$P(\hat{\Theta}^- \leq \vartheta \leq \hat{\Theta}^+) \geq 1 - \alpha.$$

- tohle jsou tzv. oboustranné odhady
- jednostranný odhad: $[\hat{\Theta}^-, \infty)$ nebo $(-\infty, \hat{\Theta}^-]$

Intervalové odhady normální náhodné veličiny

Věta

$X_1, \ldots, X_n$ je náhodný výběr z $N(\vartheta, \sigma^2)$.
σ známe, ϑ chceme určit, $\alpha \in (0,1)$.
Nechť $\Phi(z_{\alpha/2}) = 1 - \alpha/2$. Zvolíme $\hat{\Theta}_n := \bar{X}_n$.

$$C_n := [\hat{\Theta}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \hat{\Theta}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

Pak $P(C_n \ni \vartheta) = 1 - \alpha$.

Důkaz.

Intervalové odhady pomocí CLV

Věta

$X_1, \ldots, X_n$ je náhodný výběr z rozdělení se střední hodnotou ϑ, rozptylem σ^2.
σ známe, ϑ chceme určit, $\alpha \in (0,1)$.
Nechť $\Phi(z_{\alpha/2}) = 1 - \alpha/2$. Zvolíme $\hat{\Theta}_n := \bar{X}_n$.

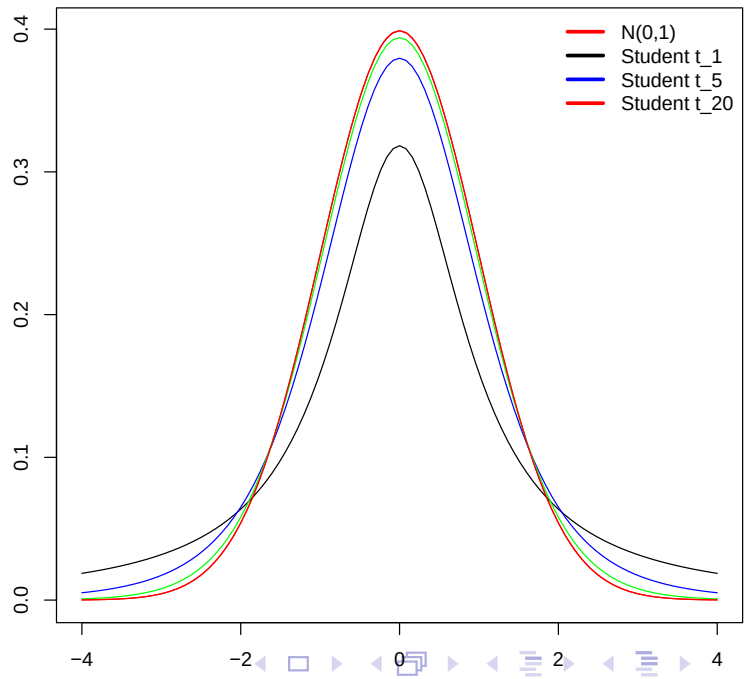
$$C_n := [\hat{\Theta}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \hat{\Theta}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$$

Pak $\lim_{x\to\infty} P(C_n \ni \vartheta) = 1 - \alpha$.

Studentovo rozdělení

- $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ ... výběrový průměr
- $\widehat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ ... výběrový rozptyl

- Nechť $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$
- Pak $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$
- *Studentovo t-rozdělení s $n-1$ stupni volnosti* je rozdělení n.v. $\dfrac{\bar{X}_n - \mu}{\widehat{S}_n/\sqrt{n}}$
- Distribuční funkci budeme značit Ψ_{n-1}
  Je v tabulkách,
  v R: **pt**(x,n−1)



Int. odhady normální n.v. pomocí Studentova t

Věta

$X_1, \ldots, X_n$ je náhodný výběr z $N(\vartheta, \sigma^2)$.
ϑ chceme určit, σ neznáme; $\alpha \in (0,1)$. Nechť
$\Psi_{n-1}(z_{\alpha/2}) = 1 - \alpha/2$. $\hat{\Theta}_n = \bar{X}_n$, $\widehat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$

$$C_n := [\hat{\Theta}_n - z_{\alpha/2}\frac{\widehat{S}_n}{\sqrt{n}}, \quad \hat{\Theta}_n + z_{\alpha/2}\frac{\widehat{S}_n}{\sqrt{n}}]$$

Pak $P(C_n \ni \vartheta) = 1 - \alpha$.

Přehled

Úvod do testování hypotéz

- Je naše mince spravedlivá?
- Je naše kostka spravedlivá?
- Má vylepšený program kratší dobu běhu než původní?
- Je léčba nemoci metodou X dobrá? (Lepší než placebo, lepší než metoda Y, . . .)
- Jsou leváci lepší boxeři?

- dvě hypotézy: H_0, H_1
- H_0 – *nulová hypotéza (null hypothesis)* – značí defaultní, konzervativní model (léčba, mince je spravedlivá)
- H_1 – *alternativní hypotéza (alternative hypothesis)* – značí alternativní model „pozoruhodnost“

Testování hypotéz – ilustrace

- Chceme testovat, zda je mince spravedlivá.
- Hodíme n-krát mincí, orel padne S-krát.
- Pokud je $|S - n/2|$ moc velké, tak mince není spravedlivá.

Testování hypotéz – ilustrace

- Chceme testovat, zda je mince spravedlivá.
- H_0: je spravedlivá
- H_1: není spravedlivá („Vědci objevili, že v kasinu byla použita falešná mince.“)
- Výsledky: zamítneme H_0/nezamítneme H_0
- Chyba 1. druhu: chybné zamítnutí. Zamítneme H_0, i když platí. Trapas.
- Chyba 2. druhu: chybné přijetí. Nezamítneme H_0, ale ona neplatí. Promarněná příležitost.
- Potřebujeme určit k takové, že budeme zamítat H_0 pokud $|S - n/2| > k$.

Testování hypotéz – obecný postup

- Vybereme vhodný statistický model.
- Volíme *hladinu významnosti (significance level)* α: pravd. chybného zamítnutí H_0. Typicky $\alpha = 0.05$.
- Určíme *testovou statistiku* $S = h(X_1, \ldots, X_n)$, kterou budeme určovat z naměřených dat.
- Určíme *kritický obor (rejection region)* – množinu W.
- Naměříme hodnoty $x_1, \ldots, x_n$ náh. veličin $X_1, \ldots, X_n$.
- Rozhodovací pravidlo: zamítneme H_0 pokud $h(x_1, \ldots, x_n) \in W$.
- $\alpha = P(h(X) \in W; H_0)$
- $\beta = P(h(X) \notin W; H_1)$ … *síla testu*

- často α nevolíme předem, ale spočítáme p*-hodnotu:* minimální α, pro které bychom H_0 zamítli.

Testování hypotéz – příklad

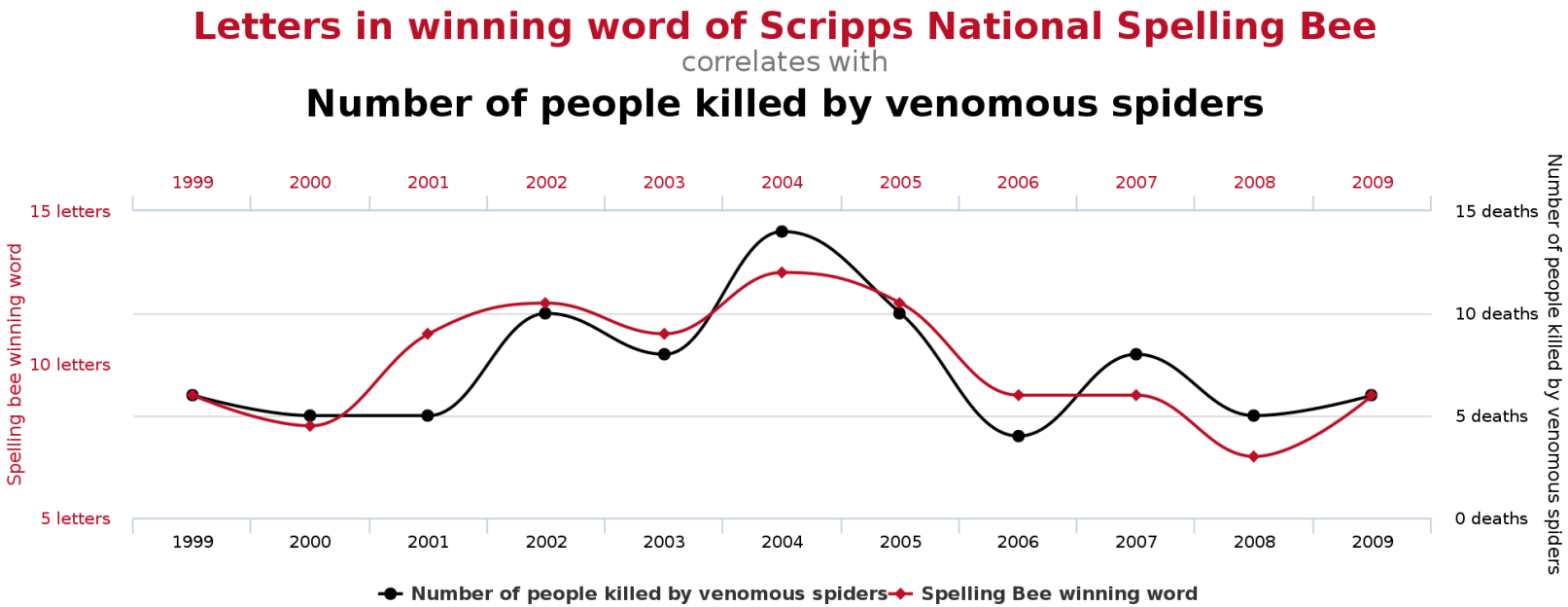
- $X_1, \ldots, X_n$ náhodný výběr z $N(\vartheta, \sigma^2)$
- σ^2 známe
- $H_0 : \vartheta = 0 \qquad H_1 : \vartheta \neq 0$

Testování hypotéz – příklad

- $X_1, \ldots, X_{n_1}$ náhodný výběr z $Ber(\vartheta_X)$
- $Y_1, \ldots, Y_{n_2}$ náhodný výběr z $Ber(\vartheta_Y)$
- $H_0 : \vartheta_X = \vartheta_Y \qquad H_1 : \vartheta_X \neq \vartheta_Y$

p-hacking

- napřed získáme data, pak v nich hledáme zajímavosti
- když máme dost dat, tak tam nějaké budou „shodou okolností“
- *reprodukovatelnost* – po explorační analýze dat uděláme nezávislý sběr dat a ten analyzujeme konfirmačně
- nebo dopředu náhodně rozdělíme data na část pro tvorbu hypotéz a část pro jejich potvrzení . . . jednoduchý případ křížové validace (cross validation)



NMAI059 Pravděpodobnost a statistika 1
12. přednáška

Robert Šámal

Statistika – Co už víme

- základní nastavení: uvažujeme náhodný výběr $X_1, \ldots, X_n$ z distribuce F_ϑ — popisuje proces měření, jak mohlo měření proběhnout
- naměříme data – konkrétní čísla, tzv. realizaci náhodného výběru $x_1, \ldots, x_n$ — jak naše měření skutečně proběhlo

1. bodové odhady: máme určit co nejlepší číslo, odhad pro parametr ϑ, nebo nějakou jeho funkci $g(\vartheta)$.
2. intervalové odhady: máme určit interval, ve kterém parametr ϑ pravděpodobně leží
3. testování hypotéz

Přehled

Testování hypotéz – ilustrace

- Chceme testovat, zda je mince spravedlivá.
- H_0: je spravedlivá *očekávaný stav světa*
- H_1: není spravedlivá *překvapivé zjištění* („Vědci objevili, že v kasinu byla použita falešná mince.“)
- Výsledky: zamítneme H_0/nezamítneme H_0
- Chyba 1. druhu: chybné zamítnutí. Zamítneme H_0, i když platí. Trapas.
- Chyba 2. druhu: chybné přijetí. Nezamítneme H_0, ale ona neplatí. Promarněná příležitost.
- Potřebujeme určit k takové, že budeme zamítat H_0 pokud $|S - n/2| > k$.

Testování hypotéz – obecný postup

- Vybereme vhodný statistický model.
- Volíme *hladinu významnosti (significance level)* α: pravd. chybného zamítnutí H_0. Typicky $\alpha = 0.05$.
- Určíme *testovou statistiku* $T = h(X_1, \ldots, X_n)$, kterou budeme určovat z naměřených dat.
- Určíme *kritický obor (rejection region)* – množinu W.
- Naměříme hodnoty $x_1, \ldots, x_n$ náh. veličin $X_1, \ldots, X_n$.
- Rozhodovací pravidlo: zamítneme H_0 pokud $h(x_1, \ldots, x_n) \in W$.
- $\alpha = P(h(X) \in W; H_0)$
- $\beta = P(h(X) \notin W; H_1)$ … $1 - \beta$ je tzv. *síla testu*

- často α nevolíme předem, ale spočítáme tzv. p*-hodnotu:* minimální α, pro které bychom H_0 zamítli.

Testování hypotéz – příklad

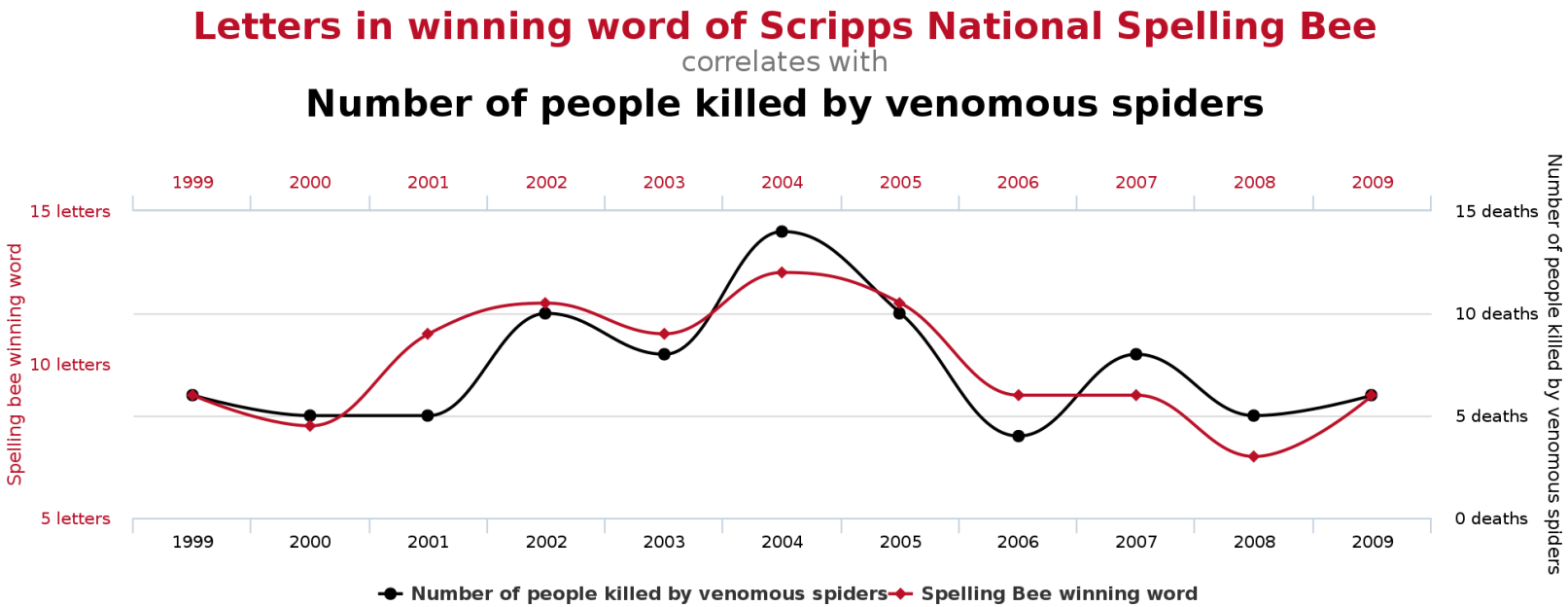
- $X_1, \ldots, X_n$ náhodný výběr z $N(\vartheta, \sigma^2)$
- σ^2 známe, μ dáno
- $H_0 : \vartheta = \mu \qquad H_1 : \vartheta \neq \mu$

Testování hypotéz – příklad dvojvýběrového testu

- $X_1, \ldots, X_{n_1}$ náhodný výběr z $Ber(\vartheta_X)$
- $Y_1, \ldots, Y_{n_2}$ náhodný výběr z $Ber(\vartheta_Y)$
- $H_0 : \vartheta_X = \vartheta_Y \qquad H_1 : \vartheta_X \neq \vartheta_Y$

p-hacking

- napřed získáme data, pak v nich hledáme zajímavosti
- když máme dost dat, tak tam nějaké budou „shodou okolností“
- *reprodukovatelnost* – po explorační analýze dat uděláme nezávislý sběr dat a ten analyzujeme konfirmačně
- nebo dopředu náhodně rozdělíme data na část pro tvorbu hypotéz a část pro jejich potvrzení . . . jednoduchý případ křížové validace (cross validation)



Přehled

χ^2_k – rozdělení χ-kvadrát

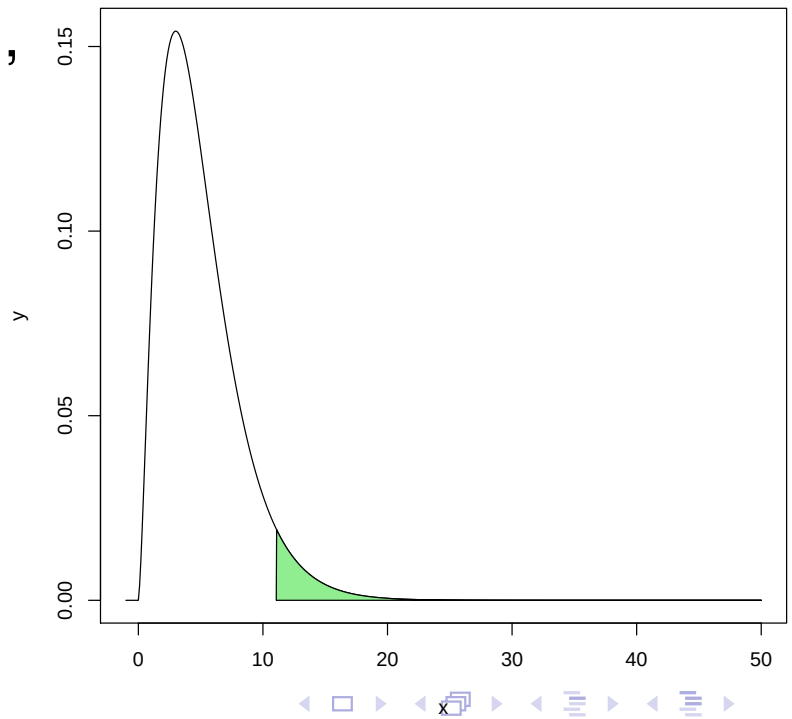
Definice

$Z_1, \ldots, Z_k \sim N(0,1)$ *n.n.v. Rozdělení náhodné veličiny*

$$Q = Z_1^2 + \cdots + Z_k^2$$

se nazývá χ-kvadrát s k stupni volnosti. (Opravdu k!)

- $\mathbb{E}(Q) = k$ (lehké)
- $var(Q) = 2k$ (pro info, netřeba pamatovat)
- hustota jde napsat vzorcem, jde najít např. na Wikipedii
- $Q \doteq N(k, 2k)$ pro velká k (CLV)



Multinomické a kategoriální rozdělení

Definice

Dána $p_1, \ldots, p_k \geq 0$ *tak, že* $p_1 + p_2 + \cdots + p_k = 1$.
n*-krát zopakuji pokus, kde může nastat jedna z* k *možností,* i*-tá má pravděpodobnost* p_i.
$X_i :=$ *kolikát nastala* i*-tá možnost* $(X_1, \ldots, X_k)$ *má multinomické rozdělení s parametry* $n, (p_1, \ldots, p_k)$.

- triviální případ: $X_i =$ počet hodů kostkou, kdy padlo i
- důležitý případ: $X_i =$ počet výskytů i-tého písmene, i-tého slovního druhu, ...
- $P(X_1 = x_1, \ldots, X_k = x_k) = \binom{n}{x_1, \ldots, x_k} p_1^{x_1} \ldots p_k^{x_k}$

Pearsonova χ^2 statistika

- $(X_1, \ldots, X_k)$ – multinomické rozdělení s parametry $n, (p_1, \ldots, p_k)$ jako minule
- $E_i := \mathbb{E}(X_i) = np_i$
- *Pearsonova χ^2 statistika* je funkce

$$T := \sum_{i=1}^{k} \frac{(X_i - E_i)^2}{E_i}$$

- **Věta** $T \xrightarrow{d} \chi^2_{k-1}$

Test dobré shody (goodness of fit)

- $(X_1, \ldots, X_k)$ – multinomické rozdělení s parametry $n, \vartheta = (\vartheta_1, \ldots, \vartheta_k)$ jako minule
- n známe, ϑ neznáme.
- Hypotéza H_0: $\vartheta = \vartheta^*$
- $E_i := n\vartheta_i^*$ pro všechna i
- Použijeme statistiku $\chi^2 = T := \sum_{i=1}^{k} \frac{(X_i - E_i)^2}{E_i}$
- Hypotézu H_0 zamítneme, pokud $T > \gamma$
- $\gamma := F_Q^{-1}(1-\alpha)$, kde $Q \sim \chi^2_{k-1}$
- $P(\text{chyba prvního druhu}) = P(T > \gamma; H_0) \to P(Q > \gamma) = \alpha$

Test dobré shody – příklad

- Házíme opakovaně kostkou. Jednotlivá čísla padla s četností 92, 120, 88, 98, 95, 107.
- Je kostka spravedlivá?

Další rozšíření

- Pro zkoumání rozdělení libovolné n.v. Y můžeme vybrat „přihrádky“ $B_1, \ldots, B_k$ (rozklad $\mathbb{R}$) a zkoumat, kolikrát je $Y \in B_i$
- Obdobný test pro nezávislost (diskrétních) náhodných veličin

Přehled

Lineární regrese – zadání

- data: (x_i, y_i) pro $i = 1, \ldots, n$
- cíl: $y = \vartheta_0 + \vartheta_1 x$

- měříme pomocí kvadratické odchylky

$$\sum_{i=1}^{n} \big(y_i - (\vartheta_0 + \vartheta_1 x_i)\big)^2$$

Lineární regrese – řešení

- Minimalizujeme výraz

$$\sum_{i=1}^{n} \big(y_i - (\vartheta_0 + \vartheta_1 x_i)\big)^2$$

- řešení: Optimální parametry jsou

$$\hat{\vartheta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \hat{\vartheta}_0 = \bar{y} - \vartheta_1 \bar{x},$$

kde $\bar{x} := (x_1 + \cdots + x_n)/n$, $\bar{y} := (y_1 + \cdots + y_n)/n$.

Lineární regrese – proč součet čtverců?

- Předpokládejme, že $x_1, \ldots, x_n$ jsou pevná, y_i je zvoleno jako hodnota náhodné veličiny

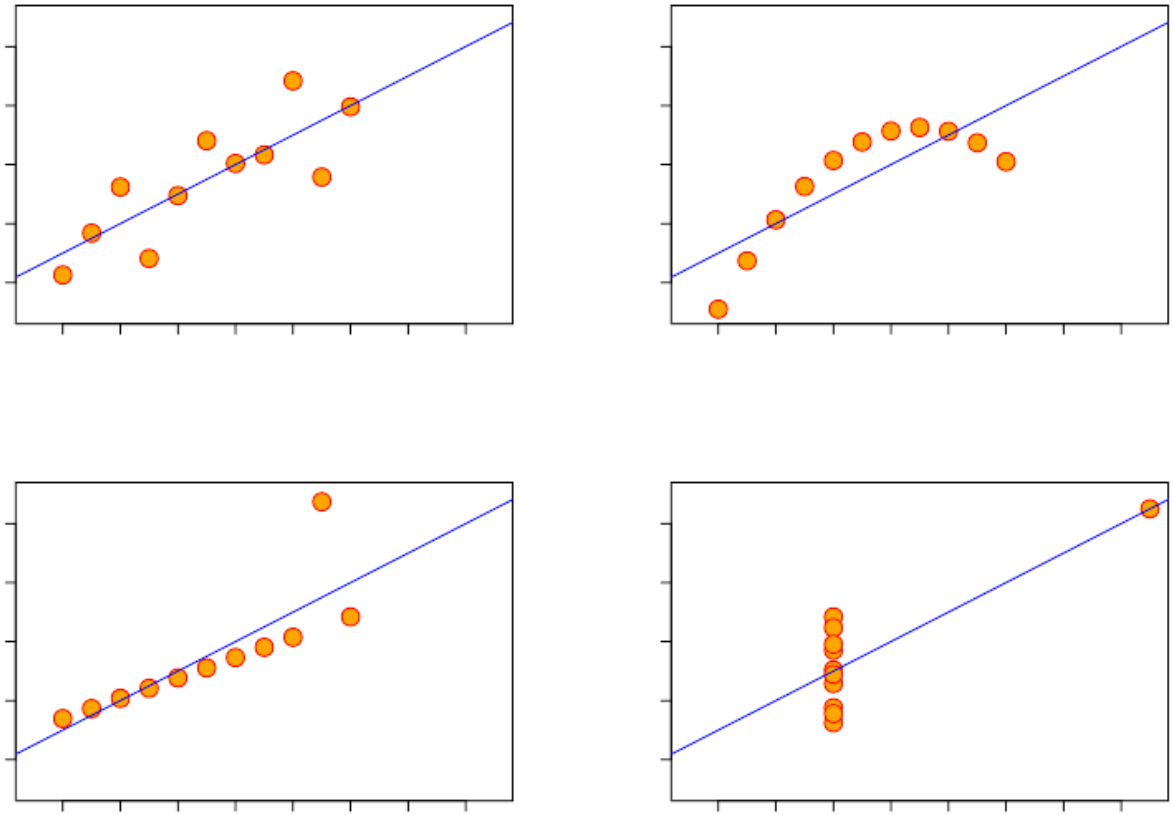
$$Y_i = \vartheta_0 + \vartheta_1 x_i + W_i$$

- $W_i \sim N(0, \sigma^2)$ pro všechna i; $W_1, \ldots, W_k$ nezávislé.
- metoda maximální věrohodnosti:

$$L(y; \vartheta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - \vartheta_0 - \vartheta_1 x_i)^2}{2\sigma^2}}$$

- $\ell(y; \vartheta) = \log L(y; \vartheta) = a + b \sum_{i=1}^{n} (y_i - \vartheta_0 - \vartheta_1 x_i)^2$

Limity regrese

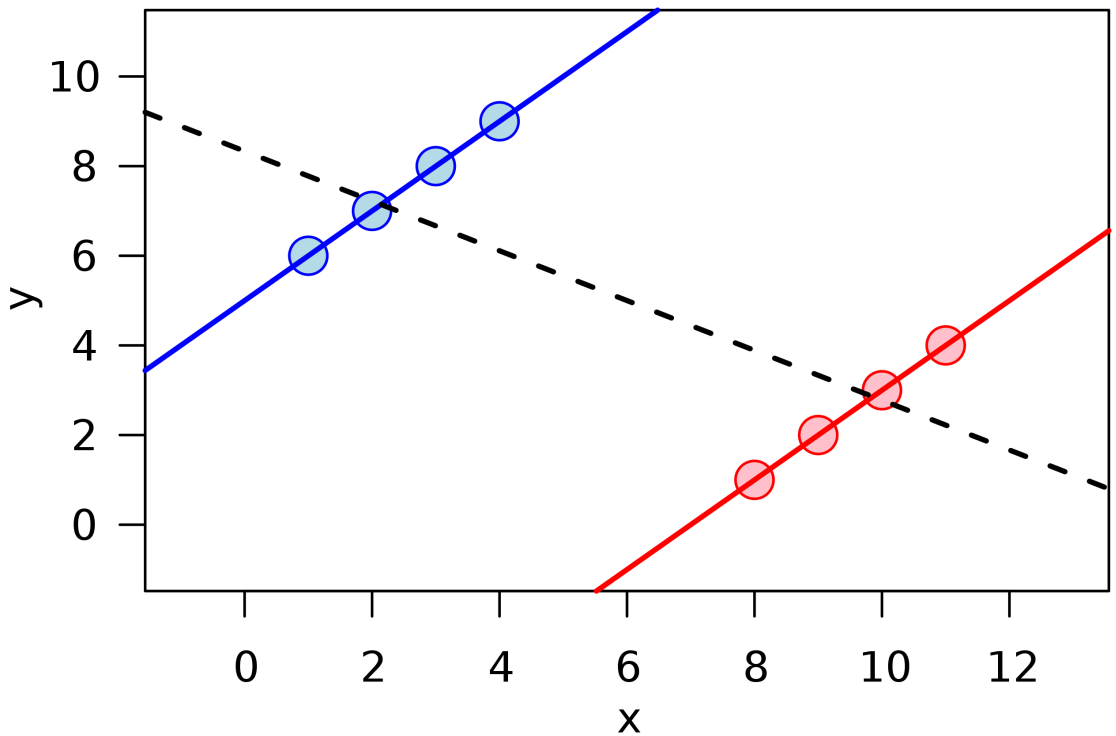


(data: Francis Anscombe 1973, obrázek: wikieditor Schutz)

- nelineární regrese

Simpsonův paradox

Treatment / Stone size	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**



NMAI059 Pravděpodobnost a statistika 1
13. přednáška

Robert Šámal

Přehled

Situace

- Máme k dispozici dvě sady nezávislých náhodných veličin (náhodné výběry):
- $X_1, \ldots, X_n \sim F_X$ a $Y_1, \ldots, Y_m \sim F_Y$
- Chceme rozhodnout, zda platí $H_0 : F_X = F_Y$ nebo $H_1 : F_X \neq F_Y$
- Příklady: doba běhu programu před/po vylepšení, hladina cholesterolu u lidí co jedí/nejedí Zázračnou Superpotravu$^{\text{TM}}$, frekvenci krátkých slov v textu autora X a Y.
- Nevíme nic o vlastnostech F_X, F_Y (zejména nečekáme, že je normální)

Postup

- Zvolíme vhodnou statistiku, např.

$$T(X_1, \ldots, X_n, Y_1, \ldots, Y_m) = |\bar{X}_n - \bar{Y}_m|$$

- $t_{\text{obs}} := T(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$
- Za předpokladu H_0 jsou „všechny permutace stejné“: X_i i Y_j se generovaly ze stejného rozdělení.
- Náhodně zpermutujeme zadaných $m + n$ čísel a pro každou permutaci vyčíslíme T – dostaneme čísla $T_1, T_2, \ldots, T_{(m+n)!}$ (každé stejně pravděpodobné).
- Jako p-hodnotu vezmeme pravděpodobnost, že $T > t_{\text{obs}}$, neboli

$$p = \frac{1}{(m+n)!} \sum_j I(T_j > t_{\text{obs}}).$$

- To je pravděpodobnost chyby 1. druhu, neboli H_0 zamítneme, pokud je $p < \alpha$ (pro naši zvolenou hodnotu α, např. $\alpha = 0.05$).

Vylepšení

- Zkoušet všechny permutace může trvat moc dlouho. Vezmeme tedy jen vhodný počet B nezávisle náhodně vygenerovaných permutací a spočítáme jenom B hodnot $T_1, \ldots, T_B$.
- Jako p-hodnotu vezmeme odhad pravděpodobnost, že $T > t_{\text{obs}}$, neboli

$$\frac{1}{B} \sum_{j=1}^{B} I(T_j > t_{\text{obs}}).$$

- Pro dostatečně velké m, n dává podobné výsledky jako testy založené na CLV, vhodné je tedy zejména pro středně velké počty.

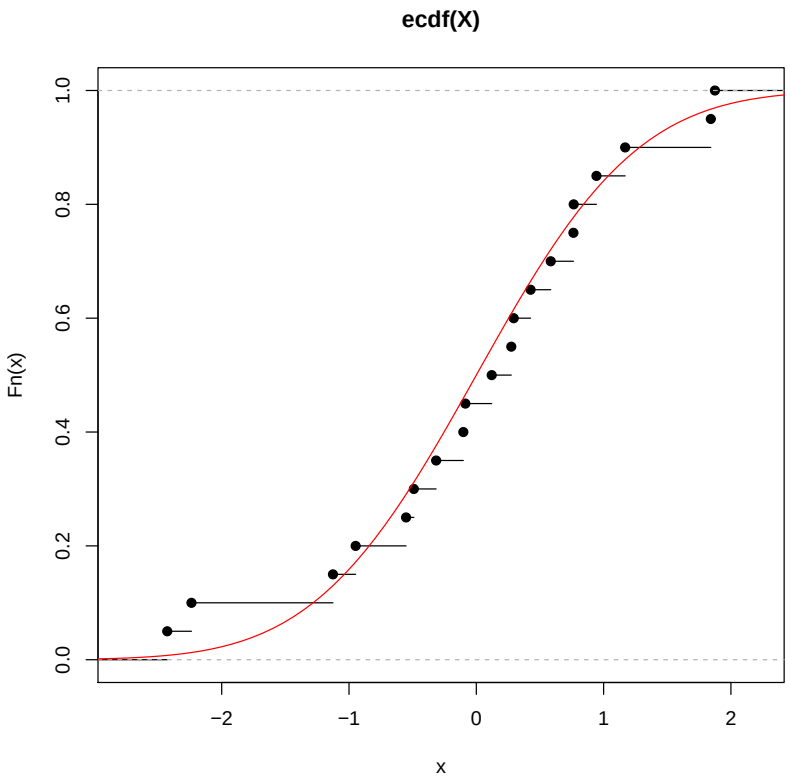
Přehled

Empirická distribuční funkce – připomenutí

- $X_1, \ldots, X_n \sim F$ n.n.v., F je jejich distribuční funkce
- **Definice:** *Empirická distribuční funkce (empirical CDF)* je definována

$$\widehat{F}_n(x) = \frac{\sum_{i=1}^{n} I(X_i \leq x)}{n},$$

kde $I(X_i \leq x) = 1$ pokud $X_i \leq x$ a 0 jinak.



Boostrap – základní idea

- z naměřených dat $X_1 = x_1, \ldots, X_n = x_n \sim F$ vytvoříme $\widehat{F}_n$
- další data můžeme samplovat z $\widehat{F}_n$
- to se dělá tak, že vybereme uniformně náhodné $i \in \{1, \ldots, n\}$ a řekneme x_i

Bootstrap – základní použití

- $T_n = g(X_1, \ldots, X_n)$ nějaká statistika (funkce dat)
- chceme odhadnout $var\, T_n$
- nasamplujeme $X_1^*, \ldots, X_n^* \sim \widehat{F}_n$ (viz minulá strana)
- spočteme $T_n^* = g(X_1^*, \ldots, g_n^*)$
- opakujeme B-krát, dostaneme $T_{n,1}^*, \ldots, T_{n,B}^*$
- odhad rozptylu:

$$\frac{1}{B}\sum_{b=1}^{B}\left(T_{n,b}^* - \frac{1}{B}\sum_{k=1}^{B} T_{n,k}^*\right)^2$$

Přehled

Srovnání dvou přístupů ke statistice

Frekventistický/klasický přístup

- Pravděpodobnost je dlouhodobá frekvence (z 6000 hodů kostkou padla šestka 1026-krát). Je to objektivní vlastnost reálného světa.
- Parametry jsou pevné, neznámé konstanty. Nelze o nich říkat smysluplné pravděpodobnostní výroky.
- Navrhujeme statistické procedury tak, aby měly žádané dlouhodobé vlastnosti. Např. 95 % z našich intervalových odhadů pokryje neznámý parametr.

Bayesovský přístup

- Pravděpodobnost popisuje, jak moc věříme nějakému jevu, jak moc jsme ochotní se vsadit. (Pravděpodobnost, že Thomas Bayes měl 18. prosince 1760 šálek čaje, je 90 %.)
- Můžeme vyslovovat pravděpodobnostní výroky i o parametrech (třebaže jsou to pevné konstanty).
- Spočítáme distribuci ϑ a z ní tvoříme bodové a intervalové odhady, atd.

Bayesovská metoda – základní popis

- neznámý parametr považujeme za náhodnou veličinu Θ
- zvolíme *apriorní distribuci (prior distribution)*, neboli hustotu pravděpodobnosti $f_\Theta(\vartheta)$ nezávislou na datech.
- zvolíme statistický model $f_{X|\Theta}(x|\vartheta)$, který popisuje, co naměříme (s jakou pravděpodobností), v závislosti na hodnotě parametru
- poté, co pozorujeme hodnotu $X = x$, spočítáme *posteriorní distribuci (posterior distribution)* $f_{\Theta|X}(\vartheta|x)$
- z té pak odvodíme, co potřebujeme např. najdeme a, b, aby $\int_a^b f_{\Theta|X}(\vartheta|x)d\vartheta \geq 1 - \alpha$

- $\vartheta = \theta$ malá théta, Θ je velká théta

Bayesova věta

Věta (Bayesova pro diskrétní náhodné veličiny)

X, Θ jsou diskrétní n.v.

$$p_{\Theta|X}(\vartheta|x) = \frac{p_{X|\Theta}(x|\vartheta)p_{\Theta}(\vartheta)}{\sum_{\vartheta' \in Im\Theta} p_{X|\Theta}(x|\vartheta')p_{\Theta}(\vartheta')}.$$

(sčítance s $p_{\Theta}(\vartheta') = 0$ považujeme za 0).

Věta (Bayesova pro spojité náhodné veličiny)

X, Θ jsou spojité n.v., které mají hustotu f_X, f_{Θ} i sdruženou hustotu $f_{X,\Theta}$

$$f_{\Theta|X}(\vartheta|x) = \frac{f_{X|\Theta}(x|\vartheta)f_{\Theta}(\vartheta)}{\int_{\vartheta' \in \mathbb{R}} f_{X|\Theta}(x|\vartheta')f_{\Theta}(\vartheta')d\vartheta'}.$$

(sčítance s $f_{\Theta}(\vartheta') = 0$ považujeme za 0).

Bayesovské bodové odhady – MAP a LMS

MAP – Maximum A-Posteriori

Volíme $\hat{\vartheta}$ tak, aby maximalizovalo

- $p_{\Theta|X}(\vartheta|x)$ v diskrétním případě
- $f_{\Theta|X}(\vartheta|x)$ ve spojitém případě
- Podobné metodě ML v klasickém přístupu, pokud bychom volili „flat prior“ – uniformní $p_{\Theta}(\vartheta)$.

LMS – Least Mean Square

Též metoda podmíněné střední hodnoty.

- Volíme $\hat{\vartheta} = \mathbb{E}\big(\Theta \mid X = x\big)$.
- Nestranný bodový odhad, má nejmenší možnou hodnotu LMS: $\mathbb{E}\big((\Theta - \hat{\vartheta})^2 | X = x\big)$.

Příklad 1

Bayesovský klasifikátor spamů:

- vytvoříme seznam podezřelých slov (money, win, pharmacy, . . .)
- N.v. X_i popisuje, zda email obsahuje podezřelé slovo w_i.
- N.v. Θ popisuje, zda email je spam $\Theta = 1$ nebo ne $\Theta = 0$.
- Z předchozích emailů získáme odhady $p_{X|\Theta}$ a p_Θ
- Použijeme Bayesovu větu na výpočet $p_{\Theta|X}$

Příklad 2

Romeo a Julie se mají sejít přesně v poledne. Julie ale přijde pozdě o dobu popsanou náhodnou veličinou $X \sim U(0, \vartheta)$. Parametr ϑ modelujeme náhodnou veličinou $\Theta \sim U(0, 1)$. Co z naměřené hodnoty $X = x$ usoudíme o ϑ?

Příklad 3

Pozorujeme náhodné veličiny $X = (X_1, \ldots, X_n)$, předpokládáme $X_i \sim N(\vartheta, \sigma_i^2)$ a ϑ je hodnota náhodné veličiny $\Theta \sim N(x_0, \sigma_0)$. Co z naměřených hodnot $X = x = (x_1, \ldots, x_n)$ usoudíme o ϑ?

Příklad 4

Házíme mincí, pravděpodobnost, že padne panna je ϑ. Z n hodů padla panna v $X = k$ případech. Pokud naše apriorní distribuce byla $U(0, 1)$, jaká bude distribuce posteriorní?

Přehled

Základní metoda (inverse transformation method)

Věta

Nechť F je funkce „typu distribuční funkce“: neklesající zprava spojitá funkce s $\lim_{x\to-\infty} F(x) = 0$ a $\lim_{x\to+\infty} F(x) = 1$.
Nechť Q je odpovídající kvantilová funkce.
Nechť $U \sim U(0,1)$ a $X = Q(U)$.
Pak X má distribuční funkci F.

- Funguje dobře, když umíme vyčíslit Q, třeba pro exponenciální nebo geometrické rozdělení.
- Gamma rozdělení je součet několika exponenciálních – tak ho tak i vygenerujeme.

Varianta základní metody pro diskrétní proměnné

- Chceme n.v. X, která nabývá hodnot $x_1, x_2, \ldots$ s pravděpodobnostmi $p_1, p_2, \ldots$. ($\sum_i p_i = 1$).
- Vygenerujeme $U \sim U(0,1)$.
- Najdeme i takové, že $p_1 + \cdots + p_{i-1} < U < p_1 + \cdots + p_i$.
- Položíme $X := x_i$.

- Funguje hezky když máme vzorec pro $p_1 + \cdots + p_i$ (např. geometrické rozdělení).
- Binomické rozdělení je lepší simulovat jako součet n nezávislých Bernoulliových veličin.
- Na další (Poisson) jsou speciální triky).

Zamítací metoda (rejection sampling)

- Chceme vygenerovat n.v. s hustotou f.
- Umíme vygenerovat n.v. s hustotou g (která je „podobná“).
- $\frac{f(y)}{g(y)} \leq c$ pro nějakou konstantu c.
- Postup
  1. Vygenerujeme Y s hustotou g, a $U \sim U(0,1)$.
  2. Pokud $U \leq \frac{f(Y)}{cg(Y)}$, tak $X := Y$.
  3. Jinak hodnotu Y, U zamítneme a opakujeme od bodu 1.
- Zdůvodnění: vygenerovat náhodnou hodnotu X s hustotou f je totéž, jako vygenerovat náhodný bod pod grafem funkce f, jehož vodorovná (x-ová) souřadnice je X (a svislá je uniformně náhodná mezi 0 a X).

Navazující přednášky

- Pravděpodobnost a statistika 2 - NMAI073
- Úvod do aproximačních a pravděpodobnostních algoritmů - NDMI084
- Úvod do strojového učení v Pythonu|systému R - NPFL129|NPFL054
- a mnoho magisterských přednášek