CZECH TECHNICAL UNIVERSITY IN PRAGUE

Faculty of Electrical Engineering
Department of Cybernetics

Opakování základních pojmů statistiky

Petr Pošík

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

Úvod do statistiky

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

Pravděpodobnost a statistika

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

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

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

Pravděpodobnost a statistika

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

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

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

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

Povaha statistiky

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

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

Samuel Johnson

Statistika: účel a členění

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

Statistika jako matematická disciplína:

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

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

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

Deskriptivní statistika se zabývá numerickým nebo grafickým popisem získaných dat

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

Analýza dat

Z `http://www.quora.com/What-is-the-difference-between-statistics-and-machine-learning`:

- **Machine Learning** is **AI people doing data analysis**.
- **Data Mining** is **database people doing data analysis**.
- **Applied Statistics** is **statisticians doing data analysis** (mathematical statistics is mathematicians doing statistics, a meta-activity analogous to theoretical computer scientists investigating, say, PAC and convergence - which sometimes gets put into machine learning).
- **Infographics** is **Graphic Designers doing data analysis**.
- **Data Journalism** is **Journalists doing data analysis**.
- **Econometrics** is **Economists doing data analysis** (and here you can win a Nobel Prize).
- **Psychometrics** is **Psychologists doing data analysis**.
- **Chemometrics** and **Cheminformatics are Chemists doing data analysis**.
- **Bioinformatics** is **Biologists doing data analysis**.

:-)

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

Základní pojmy

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

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

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

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

Proč výběr?

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

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

Druhy veličin

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

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

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

Druhy veličin

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

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

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

Druhy veličin

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

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

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

Druhy veličin

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

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

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

Druhy veličin

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

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

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

Šetření vs. experiment

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

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

Experiment Klinické studie, laboratorní testy

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

Randomizovaný experiment

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

Vzorek zkoumaných subjektů

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

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

- **Randomizace nemožná**, např. souvislost pohlaví s výší platů: pohlaví nelze přiřadit náhodně.
- **Randomizace možná, ale nepraktická**, např. studie kriminality ve městech a na vesnici: lidé se nepřestěhují jen kvůli průzkumu.
- **Randomizace možná, praktická, přesto se nedělá**, např. studie výhod předškolních vzdělávacích programů poskytovaných zdarma, kdy poptávka převyšuje nabídku: náhodné přidělení míst je férový postup, ale lidé odmítají připustit, že generátor náhodných čísel udělá jejich práci lépe než oni.
- **Etické problémy:** Experimenty na lidech a na zvířatech ano či ne? Správné otázky by měly znít: *Experimentujeme s rozmyslem nebo hazardujeme? Provádíme experimenty, abychom se z nich dozvěděli maximum, nebo jsou naše experimenty chabé, poskytují špatné informace a poškozují lidi?*

Etika: ilustrace

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

- *Student*: „Měl jste nějakou kontrolní skupinu?“

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

Etika: ilustrace

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

- *Student*: „Měl jste nějakou kontrolní skupinu?“
- *Chirurg*: „Myslíte tím, zda jsem svou novou metodou operoval jen polovinu pacientů?“

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

Etika: ilustrace

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

- *Student*: „Měl jste nějakou kontrolní skupinu?“
- *Chirurg*: „Myslíte tím, zda jsem svou novou metodou operoval jen polovinu pacientů?“
- *Student*: „Ano, to je přesně to, co mám na mysli.“

Lékař praštil pěstí do stolu a zahřměl:

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

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

Etika: ilustrace

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

- *Student*: „Měl jste nějakou kontrolní skupinu?“
- *Chirurg*: „Myslíte tím, zda jsem svou novou metodou operoval jen polovinu pacientů?“
- *Student*: „Ano, to je přesně to, co mám na mysli.“

Lékař praštil pěstí do stolu a zahřměl:

- *Chirurg*: „Samozřejmě, že ne! Tím bych odsoudil polovinu z nich k smrti!“
- *Student*: „A kterou polovinu?“

Kontrolní skupina

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

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

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

Statistika

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

Statistika je

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

Značení:

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

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

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

Náhodný výběr

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

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

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

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

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

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

Histogram a empirické rozdělení

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

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

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

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

Histogram a empirické rozdělení

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

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

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

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

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

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

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

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

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

Statistické odhady

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

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

Ale ona mi dala jenom odhad.

Anonym

Odhady

Cílem odhadování je

- najít takovou statistiku (funkci náhodného výběru), jejíž hodnoty *co nejlépe* odpovídají odhadovanému parametru populace.

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

Odhady

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

Cílem odhadování je

- najít takovou statistiku (funkci náhodného výběru), jejíž hodnoty *co nejlépe* odpovídají odhadovanému parametru populace.

Žádoucí vlastnosti:

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

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

Odhady

Cílem odhadování je

- najít takovou statistiku (funkci náhodného výběru), jejíž hodnoty *co nejlépe* odpovídají odhadovanému parametru populace.

Žádoucí vlastnosti:

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

Rozlišujeme odhady

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

Výběrový průměr

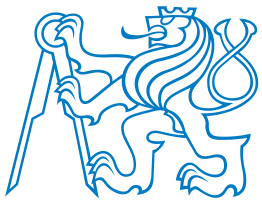
Výběrový průměr $\bar{X}$ je statistika (náhodná veličina) definovaná jako aritmetický průměr náhodného výběru:

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

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

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

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy



Výběrový průměr

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

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

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

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

Platí:

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

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

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

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

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

Centrální limitní věta

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

Centrální limitní věta

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

Podobná věta platí i pro jiná rozdělení alespoň asymptoticky.

Centrální limitní věta

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

Podobná věta platí i pro jiná rozdělení alespoň asymptoticky.

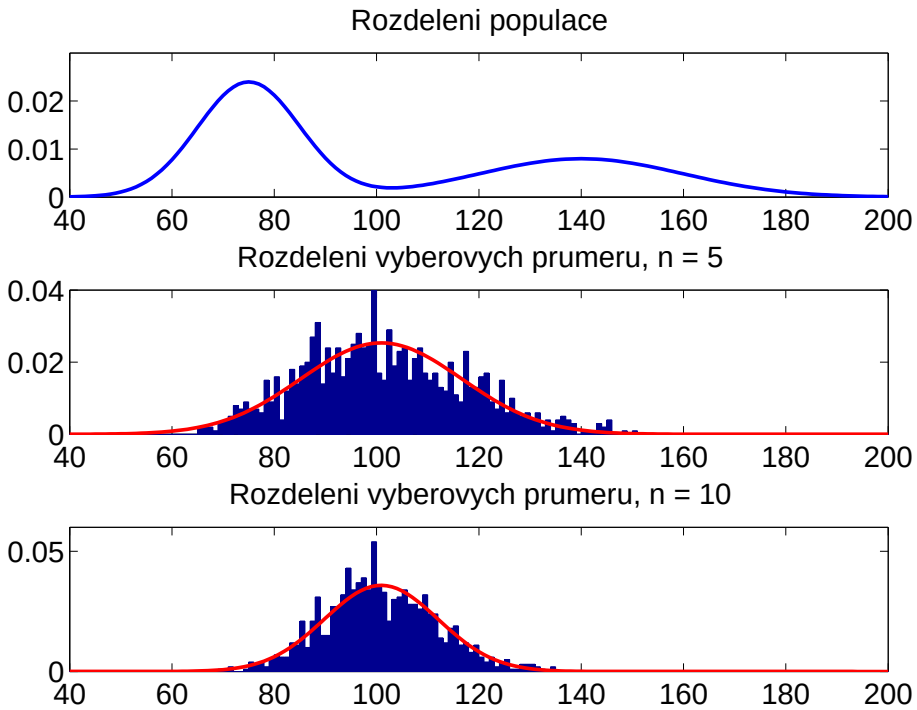
Centrální limitní věta: Nechť X_j, $j \in \mathbb{N}$, jsou nezávislé stejně rozdělené náhodné veličiny se střední hodnotou $\mathrm{E}\,X$ a směrodatnou odchylkou $\sigma_X \neq 0$. Pak normované náhodné veličiny

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

konvergují k normovanému normálnímu rozdělení v následujícím smyslu:

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

Ilustrace:



Výběrový rozptyl

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

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

a jeho **realizace**:

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

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

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

Výběrový rozptyl

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

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

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

a jeho **realizace**:

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

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

POZOR: Odhad rozptylu pomocí rozptylu empirického rozdělení

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

který je realizací odhadu

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

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

Rozdělení výběrového rozptylu

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

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

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

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

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

Rozdělení výběrového rozptylu

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

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

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

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

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

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

Vlastnosti:

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

Rozdělení χ^2

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

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

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

Vlastnosti:

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

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

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

Výběrový rozptyl z normálního rozdělení

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

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

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

Z vlastností rozdělení χ^2 plyne

- pro střední hodnotu výběrového rozptylu

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

$$\mathrm{E}\,S_X^2 = \mathrm{D}\,X\,, \qquad \text{což potvrzuje nestrannost odhadu.}$$

Výběrový rozptyl z normálního rozdělení

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

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

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

Z vlastností rozdělení χ^2 plyne

- pro střední hodnotu výběrového rozptylu

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

- pro rozptyl výběrového rozptylu

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

Výběrový rozptyl z normálního rozdělení

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

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

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

Z vlastností rozdělení χ^2 plyne

- pro střední hodnotu výběrového rozptylu

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

- pro rozptyl výběrového rozptylu

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

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

Výběrová směrodatná odchylka

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

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

a její **realizace**:

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

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

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

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

Výběrová směrodatná odchylka

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

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

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

a její **realizace**:

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

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

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

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

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

je taktéž *vychýlený* odhad směrodatné odchylky!

Výběrový medián

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

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

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

Míry polohy

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

Míry polohy

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

Míry polohy

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

Míry polohy

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

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

- Modus je

Míry polohy

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

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

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

Míry polohy

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

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

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

Míry polohy

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

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

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

Míry polohy

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

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

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

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

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

Míry polohy

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

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

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

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

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

Míry polohy

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

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

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

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

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

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

Místo závěru

Úvod do statistiky
- Pravděpodobnost a statistika
- Účel
- Analýza dat
- Základní pojmy
- Druhy veličin
- Šetření vs. exp.
- Randomizace
- Statistika
- Náhodný výběr
- Histogram a empirické rozdělení
- Odhady
- Výběrový průměr
- Cent. lim. věta
- Výběrový rozptyl
- Rozdělení výběrového rozptylu
- Chí-kvadrát
- Výběrová sm.odch.
- Výběrový medián
- Míry polohy

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

Benjamin Disraeli