Základy statistiky pro analýzu dat

Obsah kurzu

1. Den

Dopoledne: Základní pojmy, kategorizovaná data, testování hypotéz

Odpoledne: T-testy, neparametrické testy

2. Den

Dopoledne: ANOVA, třídění druhého stupně, kontingenční tabulka

Odpoledne: Korelační analýza

3. Den

Dopoledne: Regresní analýza

Odpoledne: Regresní analýza a opakování

Harmonogram kurzu

- **1. hodina 9:00-10:30**

  Pauza 10:30 – 10:45

- **2. hodina 10:45-12:15**

  Pauza 12:15-13:15

- **3. hodina 13:15-14:45**

  Pauza 14:45 – 15:00

- **4. hodina 15:00-16:30**

Typy proměnných

- **typy proměnných:**
  - číselné
  - kategorizované (nominální, ordinální)
  - textové
  - datum a čas

Popis nominálních proměnných

Rozložení četností variant znaku (pomocí tabulek četností).

Nejčastěji zastoupenou kategorii – modus (modálních kategorií někdy může být více než 1).

Variační poměr

$$v = 1 - \frac{n_{Mo}}{n}$$

Nominální rozptyl

$$nomvar = \sum((p_i(1 - p_i))$$

Normalizovaný nominální rozptyl

$$norm.nomvar = K * nomvar/(K - 1)$$

Popis ordinálních proměnných

Rozložení četností variant znaku (pomocí tabulek četností).

Nejčastěji zastoupenou kategorii – modus (modálních kategorií někdy může být více než 1).

Medián (mediánovou kategorii)

Variační rozpětí

Ordinální rozptyl

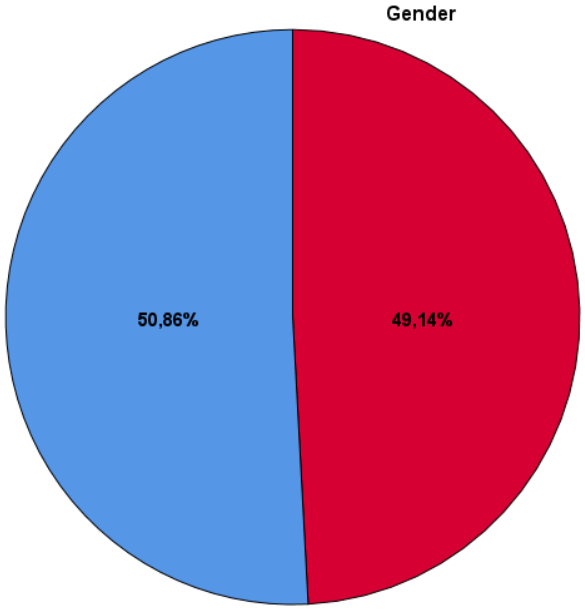
$$dorvar = 2\sum((P_i(1-P_i)$$

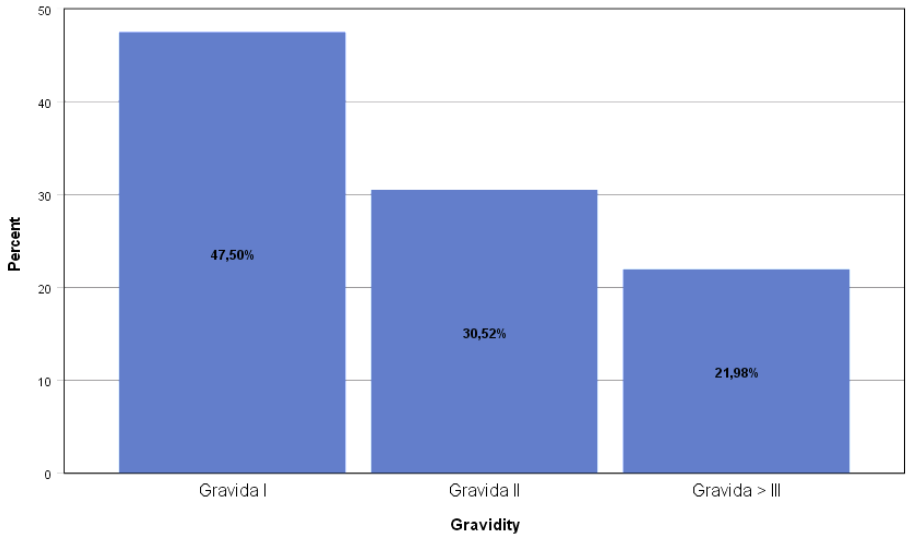
Normalizovaný ordinálni rozptyl

$$norm.dorvar = 2 * dorvar/(K-1)$$

Grafy

Pro znázornění rozložení četností se využívají i grafy znázorňující četnosti hodnot proměnných. Nejznámějšími variantami jsou koláčový a sloupcový graf.





Popis kardinálních proměnných

- **Artimetický průměr**

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

Vlastnosti:

Součet odchylek hodnot souboru od průměru je roven nule:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

Přičteme-li k hodnotám souboru konstantu a; pak průměr nového souboru:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i + a) = \bar{x} + a.$$

Násobíme-li hodnoty souboru číslem b; násobí se průměr také b:

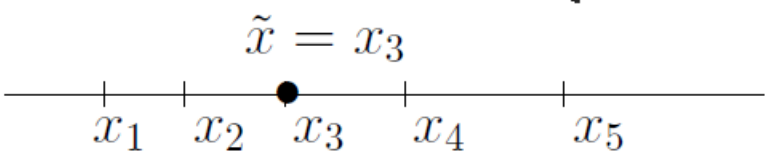
$$\frac{1}{n}\sum_{i=1}^{n} bx_i = b\bar{x}.$$

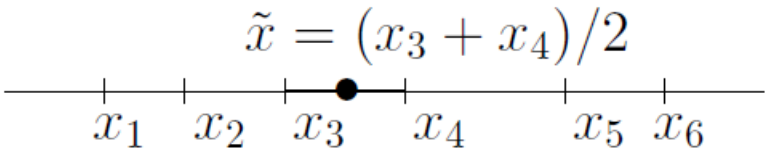
Charakteristiky polohy

Medián $\tilde{x}$

- **Průměr datového souboru je citlivý na hrubé chyby, kdy jedna chybná hodnota může výrazně změnit hodnotu průměru. Proto někdy používáme robustní charakteristiky, které jsou méně citlivé na zadání chybné hodnoty.**

$$\tilde{x} = \begin{cases} x_{(m)}, pro\ n = 2m-1, \\ \frac{1}{2}\left(x_{(m)} + x_{(m+1)}\right), pro\ n = 2m. \end{cases}$$





Modus $\hat{x}$

Hodnota souboru s největší četností.

Charakteristiky polohy

Kvantily, kvartily, decily, percentily

Kvantil datového souboru rozděluje soubor na dvě části. V jedné jsou hodnoty souboru, které jsou menší či nejvýše rovny kvantilu a ve druhé jsou hodnoty větší než kvantil.

***0<p<1, p*-kvantil resp. *100*p%* kvantil**

*Hodnota, pro kterou je přibližne 100*p% hodnot ze souboru menších a 100*(1-p)% hodnot větších.*

Speciální názvy mají kvantily:

$\tilde{x}_{50}$ je medián (median);

$\tilde{x}_{25}$ dolní kvartil (lower quartile);

$\tilde{x}_{75}$ horní kvartil (upper quartile).

Jako ***mezikvartilové rozpětí IQR*** (interquartile range) se definuje rozdíl: $IQR = \tilde{x}_{75} - \tilde{x}_{25}$

Nejjemnější používané rozdělení souboru je pomocí percentilů $(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_{99})$.

Charakteristiky polohy

Závěr

Modus snadno se najde, má ale minimální vypovídací hodnotu.

Medián určuje střed souboru a je méně citlivý na chyby.

Průměr zohledňuje všechny hodnoty, ale je citlivý na chyby.

Charakteristiky rozptýlenosti

Rozpětí datového souboru (range)

$$R = x_{max} - x_{min}$$

Snadno se spočítá, ale její hodnota je citlivá na zavlečené chyby. Vychází pouze ze dvou hodnot.

Rozptyl (dispersion,variance)

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Směrodatná odchylka (standard deviation)

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Jiné charakteristiky

Koeficient šikmosti (skewness)

$$A_3 = \frac{1}{ns^3}\sum_{i=1}^{n}(x_i - \overline{x})^3$$

Pro data, která jsou rozložena symetricky kolem průměru je A3 = 0. Hodnoty A3 blízké nule odpovídají rozdělení, které se blíží symetrickému. Je-li A3 > 0, pak je rozložení dat sešikmené vpravo, menší hodnoty než průměr jsou k němu více nahuštěny než hodnoty větší. Pro A3 < 0 je rozdělení sešikmené vlevo, větší hodnoty jsou více nahuštěny k průměru než hodnoty nižší.

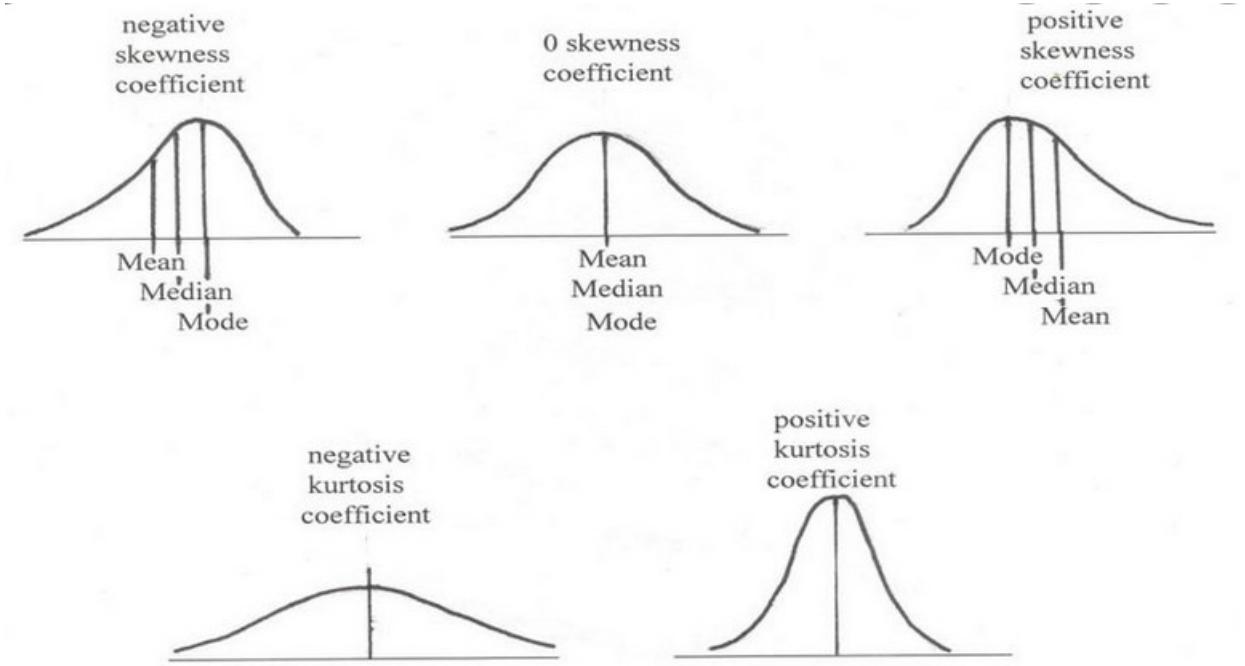
Koeficient špičatosti (kurtosis)

$$A_4 = \frac{1}{ns^4}\sum_{i=1}^{n}(x_i - \overline{x})^4 - 3$$

Je-li A4 blízké nule, říkáme, že se jedná o soubor s normální špičatostí.

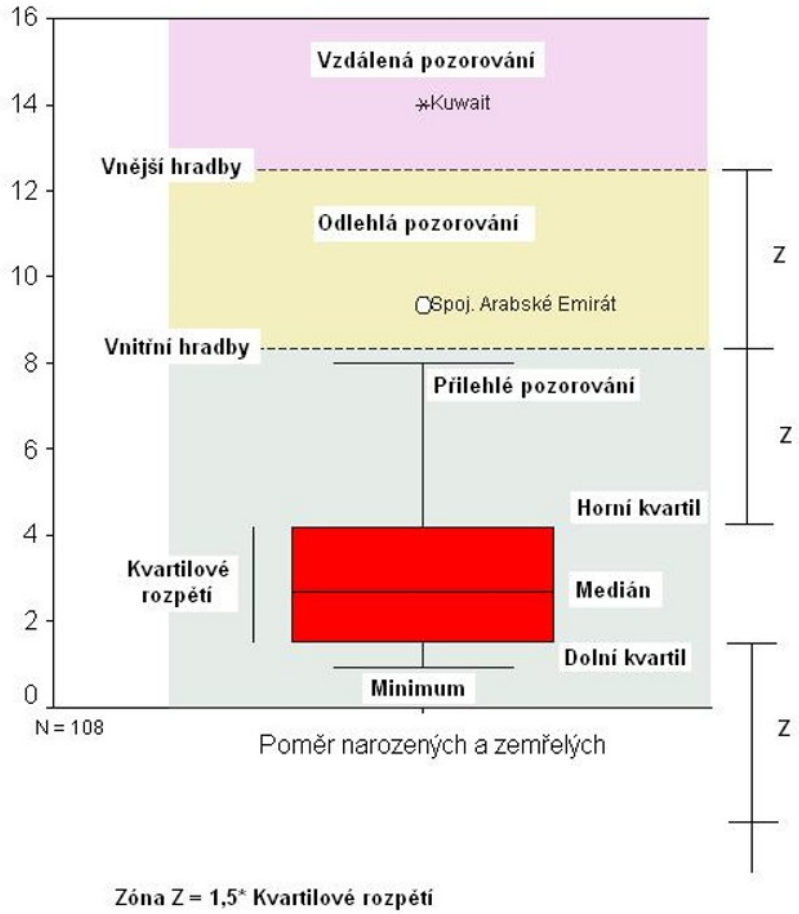
Při A4 < 0 mluvíme o souborech plochých a při A4 > 0 mluvíme o souborech špičatých.

Šikmost a špičatosť



Grafická znázornění

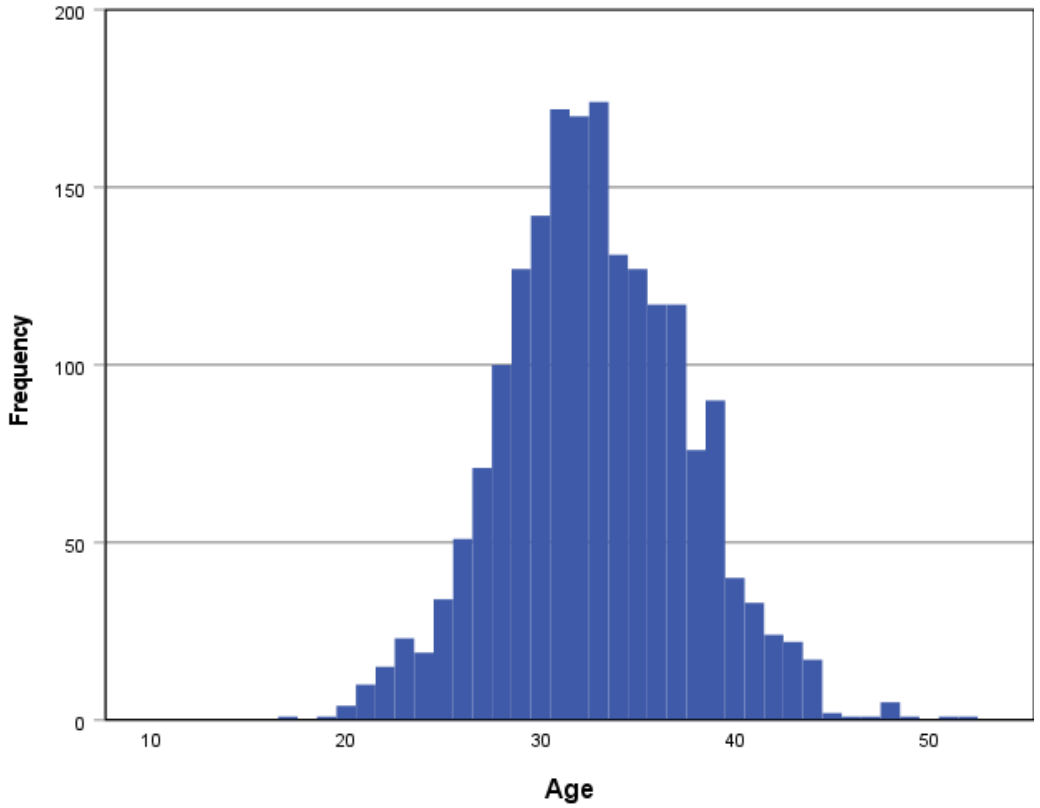
Krabicový graf (boxplot)



Zóna Z = 1,5* Kvartilové rozpětí

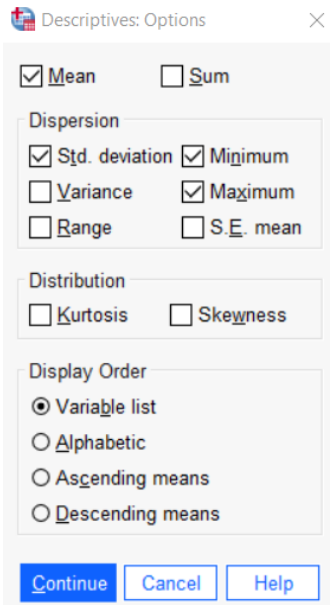
Grafické znázornění

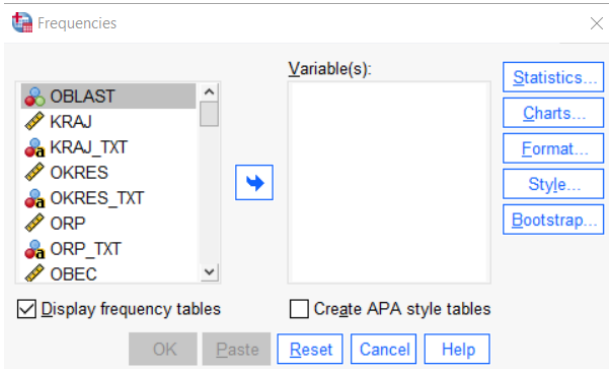
Histogram



Statistické tabelace a přehledy

- **Procedury**
  - Descriptive Statistics – Descriptives
    - Základní popisné statistiky
  - Descriptive Statistics – Frequencies
    - Tabulky četností pro kategorizované proměnné
  - Compare Means – Means
    - Tabulky statistik ve skupinách





Výběrová šetření: populace

Populace = základní soubor

Příklady: osoby nad 18 let žijící v ČR / ženy po porodu / pacienti, kteří se v posledním roce léčili v dané nemocnici / osoby trpící astmatem / děti do 10 let trpící atopickým ekzémem …

- **určení populace**
  - výčtem prvků
  - zadáním jejich vlastností
- **rozsah populace: konečný x nekonečný**
- **parametr: zvolená číselná charakteristika populace**

  Příklad: průměrný věk, podíl mužů, průměrné hodnocení spokojenosti s léčbou …

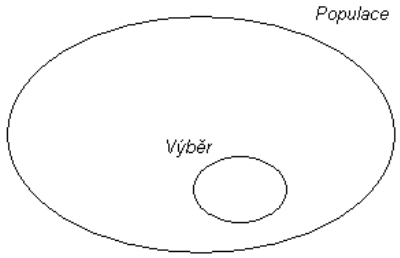
Výběrová šetření: základní pojmy

- **úplné šetření** – **sledujeme znaky u všech jednotek populace**
  (vyskytuje se pouze výjimečně, například sčítání lidu)
- **výběrové šetření** – **znaky sledujeme pouze u vybraných jednotek**

  ⇒ **výběr**

  ekonomicky a časově přijatelnější

  Příklad:
  - ***výzkum zdravotního stavu a životního stylu obyvatel Č***
  - ***zjišťování spokojenosti pacientů v nemocnici …***



- **rozsah výběru** – **počet vybraných jednotek**
- **reprezentativní výběr** – **odráží strukturu celého zkoumaného souboru**

Výběrová šetření: princip

Princip výběrových šetření

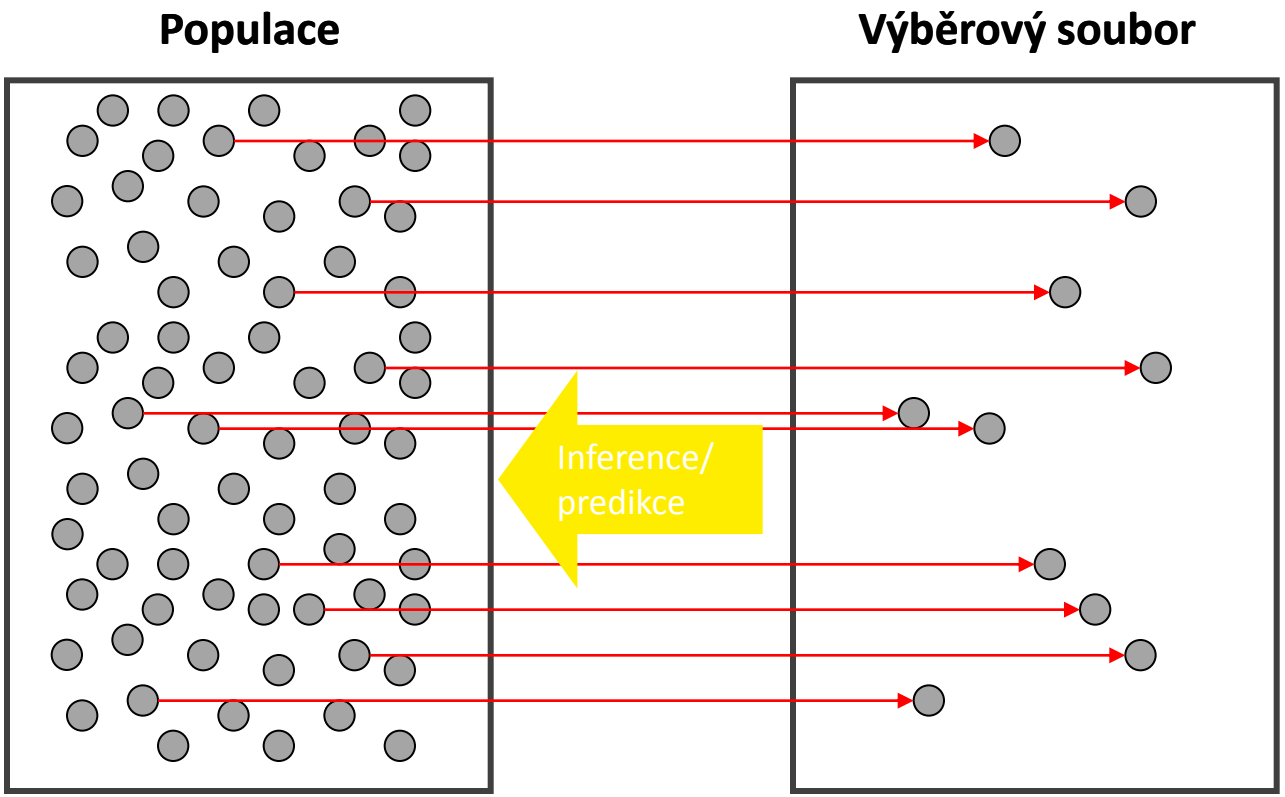
- u části populace (výběru) zjišťujeme zvolené charakteristiky
- za určitých předpokladů můžeme výsledky zobecnit na celou populaci
- závěry jsou vždy provázeny určitou nejistotou

⇒ snaha o vyjádření míry nejistoty pomocí statistických metod

Pozn.: Můžeme charakterizovat pouze míru nejistoty, která vyplývá ze způsobu realizace výběru (velikost a metoda výběru). Nezahrnuje nejistoty typu: nekvalitní data, přesnost měření…

Výběrová šetření: schéma

Reprezentativní výběrový soubor



Výběrová šetření: zajištění reprezentativnosti

Snaha o zajištění reprezentativnosti výběru:

kvótní výběr

- **určení kvót: expertní odhady nebo na základě jiného výzkumu**
- **vždy pouze pro několik základních znaků**
- **problém s vyjádřením přesnosti odhadů a odvozením intervalů spolehlivosti**

náhodný výběr

- **výběr by měl s velkou pravděpodobností odrážet vlastnosti celé populace**
- **nezávislost na subjektivním odhadu**
- **míru nejistoty lze vyhodnotit pomocí zákonů statistiky**
- **v mezinárodních srovnáních obvykle vyžadovaný standard**
- **problém odmítání účasti ve výzkumu (nonresponse)**
- **optimální metoda výběru: prostý náhodný výběr (v praxi ale může být obtížně realizovatelný)**

Statistická indukce (statistické usuzování)

- **souhrn metod pro zkoumání výběrového souboru s využitím aparátu teorie pravděpodobnosti**
- **Na základě těchto metod můžeme usuzovat (formulovat závěry) o základním souboru.**

  1. Teorie odhadu – odhadování parametrů rozdělení

  2. Testování statistických hypotéz – testujeme hypotézy o shodě parametrů rozdělení či shodě rozdělení

- **pravděpodobnost = míra očekávání výskytu náhodného jevu**

Statistické testování hypotéz

- **je rozhodovací postup na základě kterého odmítáme nebo neodmítáme statistickou hypotézu**
- **jeho výsledkem je zamítnutí nebo nezamítnutí zvoleného matematického modelu (statistické hypotézy)**

<u>Statistická hypotéza</u>

- **formální výrok**
- **tvrzení o parametrech rozdělení nebo jeho tvaru**

- **statistická hypotéza je výrok, který musí splňovat tři podmínky, aby závěry mohly být korektně použity:**

1. **je *relevantní* vzhledem k analýze dat a interpretaci**
2. **je *prověřitelný* – existují data a statistické postupy o určení jeho platnosti**
3. **je formulován *nezávisle na datech***

Formulace statistických hypotéz

- **formulujeme 2 hypotézy, které jsou ve vzájemné opozici**
- **nulová hypotéza H_0**
  - je pevně daný formální výrok specifický pro každý test
  - vyjadřuje náš předpoklad, který chceme otestovat, konkrétní hodnota testovaného populačního parametru μ_0
  - často znamená hodnotu populačního parametru rovnou 0
- **alternativní hypotéza H_A nebo H_1**
  - obecně jakákoliv jiná hodnota populačního parametru μ_0, než je v H_0
  - prakticky volíme interval, který neobsahuje μ_0
  - 3 varianty podle věcné smysluplnosti
    - oboustranný test $\mu \neq \mu_0$
    - pravostranný test $\mu > \mu_0$
    - levostranný test $\mu < \mu_0$

Rozhodovací chyby

rozhodnutí H_0 vs H_A

situace rozhodování:

	přijímáme H_0	přijímáme H_A
platí H_0	OK	chyba I. druhu
platí H_A	chyba II. druhu	OK

princip:

a) stanovíme *maximální přípustnou pravděpodobnost chyby I.druhu* = α

b) volíme testovou statistiku (a případně design sběru dat) tak, aby *minimalizovala pravděpodobnost chyby II.druhu* = β

Testové kritérium (testová statistika)

- **shrnutí informace z náhodného výběru pomocí vhodné funkce výběrových hodnot**
- **testové kritérium je funkce specifická pro každý test**
- **samotný odhad parametru nestačí – vliv na testování mají i jiné vlastnosti dat než jen testovaný parametr**

princip:

Statistická funkce dat (tzv. *testová statistika*) T je konstruována tak, aby vyjadřovala *míru neshody dat s nulovou hypotézou*, neshody odhadu a zvolené testované hodnoty. Čím vyšší je hodnota testové statistiky, tím je platnost nulové hypotézy méně pravděpodobná, *data hypotézu nepotvrzují, ale vyvracejí ji*.

Testování

- stanovili jsme statistickou hypotézu
- vybrali test
- určili nulovou hypotézu konkrétnímu test
- určili jsme testovou statistiku
- vypočítali jsme testovou statistiku

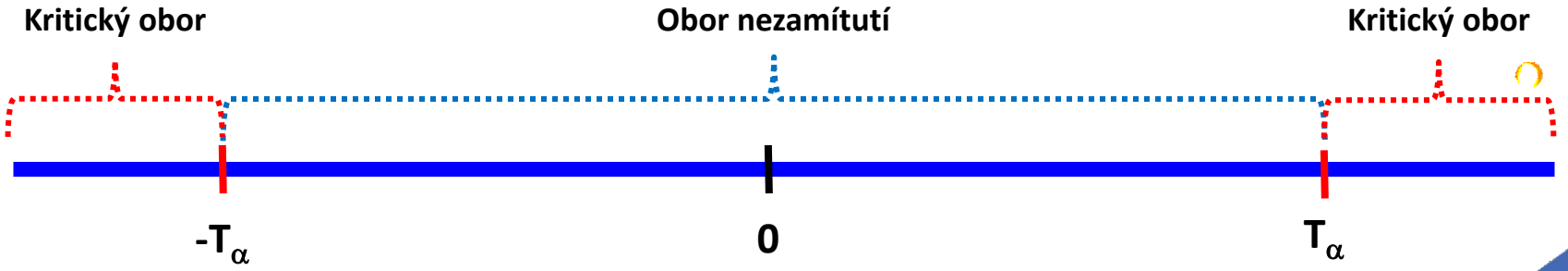
A co teď?

- porovnáme hodnotu testové statistiky s ***kritickou hodnotou***
- *kritická hodnota* odděluje ***kritický*** obor ***W*** (obor ***zamítnutí*** nulové hypotézy) od oboru „***přijetí***“ (obor ***nezamítnutí*** nulové hypotézy)
- pokud testová statistika patří do kritického oboru = **zamítáme** H_0
- pokud testová statistika patři do oboru přijetí = **nezamítáme** H_0

Jak stanovit kritickou hodnotu?

Kritická hodnota a kritický obor

- **při platnosti známe rozdělení testové statistiky T**
  - víme, jakých hodnot s jakou pravděpodobností T nabývá
- **lze spočítat kritickou hodnotu T_α v závislosti na zvoleném α**
- **volí se taková , aby pravděpodobnost, že T bude v kritickém oboru byla rovna α**
  - lze splnit nekonečně mnoho způsoby
  - jen jeden ale má nejmenší pravděpodobnost chyby druhého druhu β, tzv. stejnoměrně nejsilnější kritický obor
  - zvolený kritický obor má charakter intervalu definovaný T_α
- **oboustranný kritický obor $P(|T| > T_\alpha) = \alpha$**
- **pravostranný kritický obor $P(T > T_\alpha) = \alpha$**
- **levostranný kritický obor $P(T < -T_\alpha) = \alpha$**



Konečné rozhodnutí

- **zbývá stanovit α a tím je stanovena kritická hodnota T_α**
  - v praxi se nejčastěji volí hodnota 0,05, řidčeji 0,01
  - pouze zvyk, není vhodné brát hodnotu jako nepřekročitelné pravidlo
  - zavedl R. A. Fisher (1 z 20)
- **dva ekvivalentní způsoby rozhodování**
  a) porovnání vypočtené testové statistiky s *kritickou hodnotou*, která odpovídá hodnotě α, je-li statistika T *vyšší než kritická hodnota* T_α, zamítáme nulovou hypotézu H_0
  b) k vypočtené T zjistíme dosaženou *signifikanci* (*P-value*) a porovnáme ji s kritickou hladinou α, je-li signifikance $\leq \alpha$ zamítáme hypotézu H_0
     - signifikance je takové α, pro které $T = T_\alpha$
     - signifikance se dá interpretovat jako pravděpodobnost, že $T >= T_\alpha$ při platnosti H_0 , tedy., že odchylka dat od H_0 je způsobena pouze náhodou.
- **zamítnutí H_0 a přijetí H_A, znamená, že data neodpovídají H_0**

Interpretace rozhodnutí

- **zamítáme H_0 – přijímáme H_A**
  - říkáme pouze, že data neodpovídají H_0
  - hodnota testového kritéria závisí na velikosti rozdílu mezi odhadem a testovanou hodnotu, ale také na velikosti výběru
  - čím větší výběr tím menší rozdíl je statisticky významný
  - Vždy interpretujeme i *věcný* význam statisticky významného rozdílu.
- **nezamítáme H_0 – NEPŘÍJÍMÁME H_A**
  - testové kritérium nepadne do kritického oboru, respektive signifikance je větší než zvolená mez α
  - říkáme pouze, že data neodporují H_0
  - nemůžeme přijmout H_0, protože neznáme pravděpodobnost chyby II. druhu β
  - H_A může ve skutečnosti platit, ale nemáme dost dat, abychom to prokázali
  - používaný termín obor přijetí je zavádějící

Síla testu

- **síla testu je 1-β, kde β je pravděpodobnost chyby II. druhu**
  - vyjadžuje spolehlivost, se kterou správně zamítneme H_0, tedy, když platí H_A
- **síla závisí na skutečné hodnotě parametru v oboru H_A a poetu případů**
  - hodnotu parametru neznáme
    - kdybychom znali, bylo by celé testování zbytečné
- **závislot síly na parametru popisuje silofunkce**
  - teoreticky lze zkonstruovat
  - kritérium kvality testu, upřednosňujeme testy s větší silou, pro stejné velikosti souboru
  - lze použít pro odhad potřebné velikosti výběrového souboru
    - musíme si určit skutečnou hodnotu parametru, který chcene testem potvrdit, musíme znát i další součásti testového kritéria.

Praktické použití

Test Statistics

	Blízkost řeky
Chi-Square	375,684
df	1
Asymp. Sig.	,000

Independent Samples Test

		t-test for Equality of Means						
		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
							Lower	Upper
Medián ceny domu	Equal variances assumed	-3,996	504	0,09	-6,346157	1,587954	-9,465981	-3,226333

ROZHODOVACÍ PRAVIDLO:

significance ≤ α zamítáme nulovou hypotézu

significance > α není důvod zamítat nulovou hypotézu

(α je typicky = .05 nebo .01)

Postup statistického testování hypotéz

obecný postup:

1. **formulujeme statistickou hypotézu**
2. **pro její ověření vybereme příslušný statistický test**
3. **určíme jeho *nulovou hypotézu* H_0 *a k ní alternativní hypotézu* H_A**
4. **určíme *kritickou hladinu* α pro rozhodování**
5. **určíme *testovou statistiku (testové kritérium)* $T(X)$**
6. **dosadíme *data* X *do* $T(X)$**
7. **vyhodnotíme testové kritérium**

   a) *k α zjistíme T_α; $T(X) \geq T_\alpha$ => zamítneme H_0,*

   b) *vypočteme signifikanci; signifikanci $\leq \alpha$ => zamítneme H_0*

oba postupy a) a b) jsou ekvivalentní, není-li splněna podmínka, nezamítáme H_0