

Základní statistické pojmy

- Populace, jedinec, vzorek, znak

Typy proměnných

- Kvalitativní proměnné (nominální, dichotomické, ordinální) – modality
- Kvantitativní proměnné (diskrétní, kontinuální) - třídy

Statistika

- A) Deskriptivní statistika
- B) Inferenční statistika

Tabulky

- **Absolutní četnost**
 - o n_i - absolutní četnost i -tého znaku
 - o N – celková četnost všech znaků
- **Relativní četnost**
 - o $f_i = \frac{n_i}{N}$
- **Tabulka kumulativních četností**

Grafy

- Sloupcový graf (absolutních četností, relativních četností)
- Kruhový diagram

A) Grafy kvantitativních diskretních proměnných

- Sloupcový graf
- Kruhový graf
- Křivky kumulativních četností

B) Grafy kvantitativních kontinuálních proměnných

- **Histogramy**
 - o Index třídy – centrální hodnota třídy
 - o Interval třídy - $Interval\ třídy = \frac{Max.\ hodnota - Min.\ hodnota}{Počet\ tříd}$
 - o Výpočet tříd v histogramu
 - Sturgovo pravidlo: $počet\ tříd = 1 + (3.3 \log_{10} n)$ $= 1 + (3 * \text{LOG}(n))$
 - Yulovo pravidlo: $počet\ tříd = 2.5 \sqrt[4]{n}$ $= 2,5 * n^{(0,25)}$
- **Polygony kumulativních četností**

Charakteristiky centrální polohy

- **Průměr** – pro kvantitativní data
- **Medián** – pro kvantitativní a kvalitativní ordinální
- **Modus** – kvantitativní i kvalitativní data

Průměr pro kontinuální proměnné

- Aritmetický průměr
 - $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{\sum_{i=1}^n x_i}{n}$
- Vážený průměr
 - $\bar{x} = \frac{\sum_{i=1}^d w_i x_i}{\sum_{i=1}^d w_i}$
 - w_i – váha i-té hodnoty

Průměr pro diskrétní kvantitativní proměnné

- $\bar{x} = \frac{\sum_{i=1}^p n_i m_i}{\sum_{i=1}^p n_i}$
- *centrum třídy* = $m_i = \frac{x_i + x_{i+1}}{2}$

Třídy	Centra	Počet	Vážený počet
$[x_1, x_2]$	m_1	n_1	$n_1 \cdot m_1$
$[x_2, x_3]$	m_2	n_2	$n_2 \cdot m_2$
...	...	...	
$[x_p, x_{p+1}]$	m_p	n_p	$n_p \cdot m_p$
		$\sum_{i=1}^p n_i$	$\sum_{i=1}^p n_i \cdot m_i$

Medián

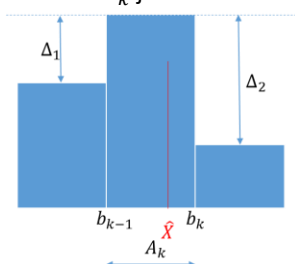
- pro lichý počet jedinců (n)
 - $med = x_{(\frac{n+1}{2})}$
 - Je roven prostřední hodnotě
- Pro sudý počet jedinců (n)
 - $med = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$
 - Je roven průměru dvou prostředních hodnot

Modus

- Nejčastěji se vyskytující hodnota

Korekce modu u diskrétních proměnných

- $Mod_{corr.} = b_{k-1} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) A_k$
 - A_k je velikost intervalu, který obsahuje hodnotu modu



Rozpětí

$$- \quad x_{rozpětí} = x_{max} - x_{min}$$

Odchylka od průměru

$$- \quad OP = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

Odchylka od mediánu

$$- \quad OM = \frac{\sum_{i=1}^n |x_i - med|}{n}$$

Variance / rozptyl

$$- \quad s_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

Variance v případě četností

$$- \quad s_x^2 = \frac{\sum_{i=1}^k f_i (x_i - \bar{x})^2}{(\sum_{i=1}^k f_i) - 1}$$

- k : počet tříd
- f_i : četnost třídy i
- x_i : centrální hodnota třídy i
- $\bar{x}$: celkový průměr (viz výše)

$$\blacksquare \quad \bar{x} = \frac{\sum_{i=1}^k x_k \cdot f_k}{\sum_{i=1}^k f_k}$$

Třída (i)	Četnost třídy (f_i)	Centrum třídy (x_i)	Celk. průměr ($\bar{x}$)	$f_i * (x_i - \bar{x})^2$
i_1	f_1	x_1	$\bar{x}$	$f_1 * (x_1 - \bar{x})^2$
i_2	f_2	x_2	$\bar{x}$	$f_2 * (x_2 - \bar{x})^2$
...	...	...	...	...
i_k	f_k	x_k	$\bar{x}$	$f_k * (x_k - \bar{x})^2$
	$\left(\sum_{i=1}^k f_i\right) - 1$			$\sum_{i=1}^k f_i (x_i - \bar{x})^2$

Směrodatná odchylka

$$- \quad s_x = \sqrt{s_x^2}$$

Variační koeficient

$$- \quad c_v = \frac{s_x}{|\bar{x}|}$$
$$- \quad c_v = \frac{100 * s_x}{|\bar{x}|} \text{ (v \%)}$$

Kvantily pro kontinuální proměnné

- pořadové číslo z p -procentního kvantilu u seřazených hodnot je hodnota
 - o $\frac{n \cdot p}{100} \leq z \leq \left(\frac{n \cdot p}{100} + 1\right)$

Kvantily pro diskrétní kontinuální proměnné

- $Q = L + \left[\frac{(nq) - \sum n_i(inf)}{n_i(kvartil)} \right] \cdot c$
 - o L : spodní hranice třídy kvartilu
 - o n : celková četnost
 - o q : $\frac{1}{4}$ pro 1. kvartil; $\frac{1}{2}$ pro Q_2 , $\frac{3}{4}$ pro Q_3
 - o $\sum n_i(inf)$: součet absolutních četností tříd, které jsou před třídou kvartilu
 - o $n_i(kvartil)$: absolutní četnost třídy kvartilu
 - o c : velikost třídy

Distribuce

- Indikátory distribuce – šikmost, strmost
- Pearsonův indikátor šikmosti
 - o $S_k = \frac{3(\bar{x} - med)}{s_x}$
 - $\bar{x}$: průměr distribuce
 - med : medián distribuce
 - s_x : směrodatná odchylka

Korelace

- Pearsonův, spearmanův korelační koeficient
- Korelace mezi proměnnou X a Y
 - o $r_{xy} = \frac{cov(X,Y)}{s_x \cdot s_y} = \frac{\sum_{i=1}^k (X_i - \bar{X})(Y_i - \bar{Y})}{s_x \cdot s_y}$
 - $cov(X, Y)$: kovariance X a Y
 - s_x : směrodatná odchylka proměnné X
 - s_y : směrodatná odchylka proměnné Y
- Pearsonův korelační koeficient
 - o $r_{xy} = \frac{n \sum XY - \sum X \sum Y}{\sqrt{[n \sum X^2 - (\sum X)^2][n \sum Y^2 - (\sum Y)^2]}}$
 - n : velikost vzorku
 - X : hodnoty proměnné X
 - Y : hodnoty proměnné Y
- Koeficient determinace
 - o $r_{xy}^2 = r_{xy}^2$
 - o $r_{xy}^2 * 100$ [v %]
- Korelační matice

Lineární regrese

- Nezávislá proměnná=prediktor
- Závislá proměnná=kritérium
- $Y' = bX + a$
 - o X je hodnoty použité jako prediktor
 - o Y' jsou hodnoty závislé proměnné
 - o b : regresní koeficient (svažitost a směr linie)
 - $b = \frac{\sum XY - (\sum X \sum Y / n)}{\sum X^2 - [(\sum X)^2 / n]}$
 - o a : regresní konstanta (bod, ve kterém linie protíná osu Y)
 - $a = \frac{\sum Y - b \sum X}{n}$

Pravděpodobnostní rozdělení

- Normální distribuce
- Z-skóre
 - o $Z = \frac{(x - \bar{x})}{s_k}$
 - x : hodnoty distribuce
 - $\bar{x}$: průměr distribuce
 - s_k : směrodatná odchylka distribuce
- Z-skóre – porovnání souborů
 - o $Z_c = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_{\bar{x}_1}^2}{n_1} + \frac{s_{\bar{x}_2}^2}{n_2}}}$
 - o Kritické hodnoty se hledají v tabulce normální distribuce

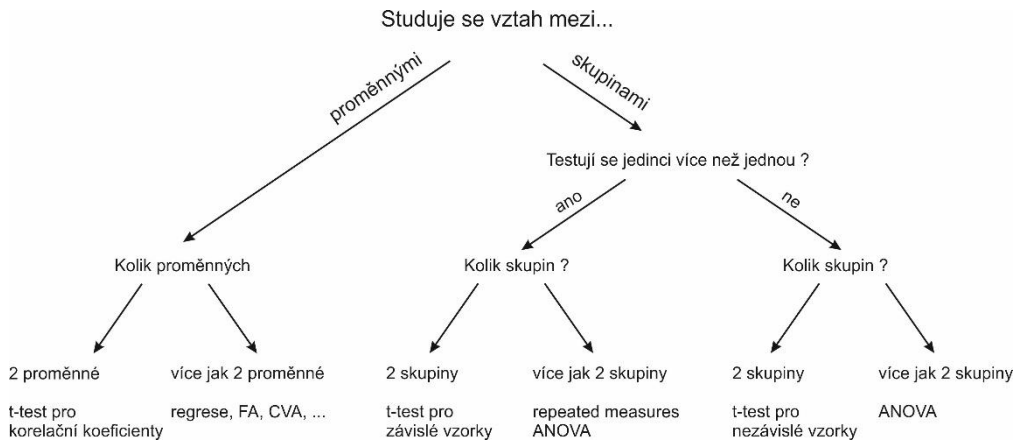
Testování hypotéz

- Nezávislé a závislé vzorky
 - o Nezávislé: Dva vzorky jsou nezávislé, jestliže každý jedinec náleží jen jedné skupině.
 - o Závislé (párové): Máme jednu skupinu jedinců, kteří jsou měřeni víckrát (před a po).
- Hypotézy
 - o H_0 : nulová hypotéza ($\mu_A = \mu_B$)
 - o H_1 : alternativní hypotéza nesměrová/bilaterální/nondirectional hypothesis ($\mu_A \neq \mu_B$)
 - o H_1 : alternativní hypotéza směrová/unilaterální/directional hypothesis ($\mu_A > \mu_B$ nebo $\mu_A < \mu_B$)
- Typy chyb (I, II)
- Signifikance, úroveň spolehlivosti
- Kritická hodnota

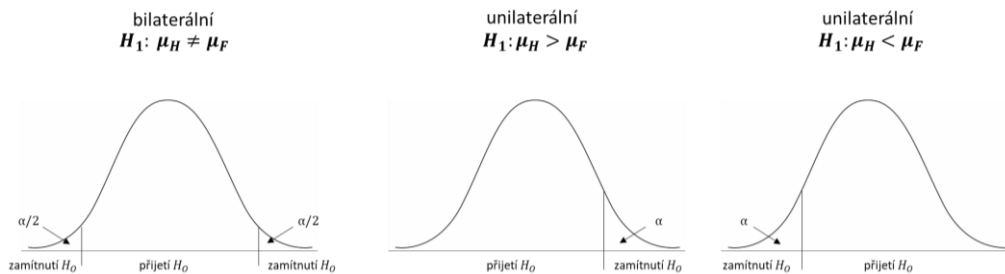
Testování hypotéz - metoda

- 1) Určení H_0 a H_1
- 2) Výběr vhodného statistického testu
- 3) Vypočtení hodnoty testu statistiky
- 4) Určení úrovně spolehlivosti, hodnoty α
- 5) Určení kritické hodnoty - záleží na distribuci testu, N, α , H_1
- 6) Rozhodnutí - porovnání vypočtené hodnoty testu s kritickou hodnotou => přijetí či zamítnutí H_0

Výběr vhodného statistického testu



Určení kritické hodnoty



Parametrické statistické testy

- Nepárový t-test

- $$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \cdot \frac{n_1+n_2}{n_1n_2}}}$$
 - $\bar{X}_1, \bar{X}_2$: průměry skupiny 1 a 2
 - n_1, n_2 : počet členů skupiny 1 a 2
 - s_1^2, s_2^2 : variance skupiny 1 a 2
- určení kritické hodnoty: $df_{t-test} = n_1 + n_2 - 2$

- Párový t-test

- $$t = \frac{\sum D}{\sqrt{\frac{n \sum D^2 - (\sum D)^2}{n-1}}}$$
 - $\sum D$: součet všech rozdílů mezi skupinami
 - $\sum D^2$: součet čtvercových rozdílů mezi skupinami
 - n : počet párových pozorování
- určení kritické hodnoty: $df_{t-test} = n - 1$

- One-way / ANOVA

- $$F = \frac{MS_{\text{mezi skupinami}}}{MS_{\text{uvnitř skupin}}}$$
- Mezi kterými skupinami je rozdíl? Nutno použít Multiple t-test (upravit α !) nebo Post-hoc (Bonferroni, Tukey HSD, Sheffe, Fisher LSD)

- t-test pro korelační koeficient

- $r_{xy} = r_{\text{vypočtená}} \Rightarrow$ stačí vypočítat korelační koeficient
- Určení kritické hodnoty: $df = n - 2$, kde $n = \text{počet párů}$

Test normality

- Hypotézy:
 - o H_0 : Hodnoty jsou normálně distribuovány
 - o H_1 : Hodnoty nejsou normálně distribuovány
- Kolmogorov-Smirnov, Shapiro-Wilk test
- Levene's test – pro rovnost variance (ANOVA)

Parametrické a neparametrické testy

Porovnáváme	Parametrický	Neparametrický
2 nezávislé skupiny	nepárový t-test	Mann-Whitney U test (také nazývaný Mann-Whitney-Wilcoxon, Wilcoxon rank-sum test, či Wilcoxon-Mann-Whitney test)
2 a více nezávislých skupin	One-way/simple ANOVA <i>Multiple t-test (upravit α)</i> <i>TukeyHSD</i> <i>Bonferroni</i>	Kruskal-Wallis (porovnává mediány) <i>Multiple Test of Wilcoxon / Mann-Whitney pairwise</i> <i>Pairwise Wilcoxon Rank Sum Tests</i> <i>Dwass-Steel test</i> <i>Nemenyi</i> <i>Dunn</i>