

STATISTIKA

ČESKÁ ZEMĚDĚLSKÁ UNIVERIZTA
APLIKOVANÁ STATISTIKA

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

$$U = \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}}$$

$$\bar{\Delta}y = \frac{\sum_{t=2}^T \Delta y_t}{T-1}$$

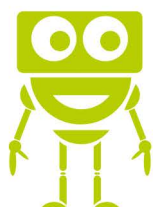
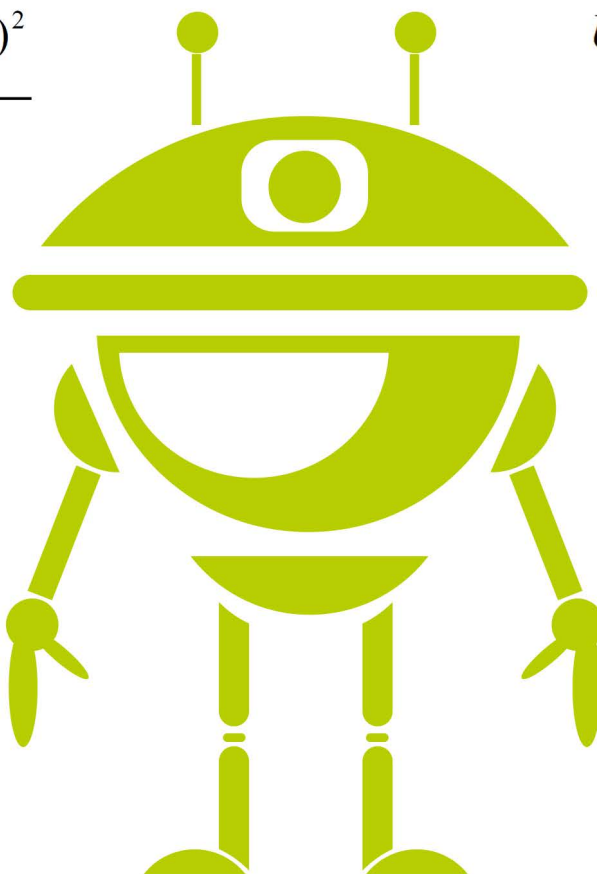
$$\sigma = \sqrt{D(X)}$$

$$E(X) = \pi$$

$$P(x) = P(X=x)$$

$$y = \beta_0 + \beta_1 x + \varepsilon$$

$$W_\alpha = \{F; F \geq F_{1-\alpha}\}$$



**STATISTICKY
NEKLASICKY**

Adriana Řeháčková
www.statistickyneklasicky.cz

APLIKOVANÁ STATISTIKA

ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA

Ukážeme si, že statistika není žádná nuda a že má opravdu smysl! Cílem kurzu je, abys dokázal základní až lehce pokročilé statistické metody správně použít při řešení příkladů.

OBSAH KURZU:

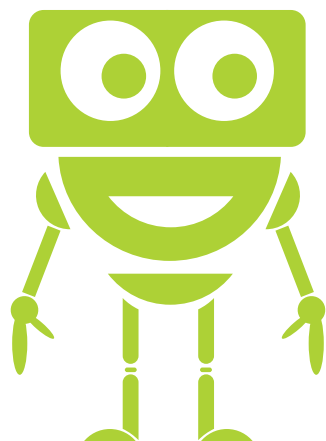
- 0(Jednovýběrové neparametrické testy
- 1(Dvouvýběrové neparametrické testy
- 2(Neparametrické testy pro 3 a více výběrů
- 3(Jednoduchá lineární regrese
- 4(Vícenásobná lineární regrese
- 5(Demografie

O TOMTO MATERIÁLU:

Žádná část tohoto materiálu nesmí být nijak použita či reprodukována bez písemného svolení autora.

Copyright ©Statistickyneklasicky 2021

Autor materiálu: Adriana Řeháčková
Sazba a grafické úpravy: Adriana Řeháčková



STATISTICKY NEKLASICKY

NEPARAMETRICKÉ TESTY

Pojďme se společně pustit do testů, které **nevyžadují normální rozdělení** a jejich výpočty vycházejí z pořadových čísel, z toho tedy vyplývá, že jsou odolné vůči odlehlým hodnotám. Tyto testy jsou vhodné i pro **ordinální proměnné** a pro **výběry o menším rozsahu**. Možná si tedy říkáte, proč vůbec řešit nějaké parametrické testy, když máme univerzálnější řešení, které lze použít mnohem častěji. Odpověď je jednoduchá, tyto testy nejsou tak silné a přesné, jako parametrické testy.

V případě jednoho výběru používáme nejčastěji **jednovýběrový Wilcoxonův test**. Jedná se o neparametrickou variantu t-testu, která spočívá na principu pořadí a spočívá na testování mediánu (prostřední hodnoty souboru). Chceme tedy testovat, zda se **medián** rovná nějakému číslu **c**.

Pokud je počet hodnot menší nebo roven 20, použijeme jednoduchý výpočet, kde od hodnot odečteme předpokládaný medián a těmto hodnotám přiřadíme pořadí a následně sečteme pořadí hodnot menších než 0 (záporné) a větších než 0 (kladné). Pořadí určíme v absolutním vyjádření. V případě, že je počet hodnot větší než 20, je vzorec komplikovanější.

Pozor, že oproti parametrickým testům, je porovnání s kritickou hodnotou definováno **opačně** a my se ptáme, zda je výsledek testu menší, než kritická hodnota. Pokud ano, tak **zamítáme nulovou hypotézu** na dané hladině významnosti.

DVOUVÝBĚROVÉ NEPARAMETRICKÉ TESTY

Platí zde opět rozdělení na závislé a nezávislé výběry spočívající na stejném principu jako v případě parametrických testů.

Wilcoxonův dvouvýběrový test pro závislé výběry

Jedná se o neparametrickou alternativu párového t-testu a v tomto případě nulová hypotéza předpokládá, že rozdíly mezi dvěma soubory jsou symetricky rozložené kolem nuly, alternativní hypotéza předpokládá, že nejsou takto rozložené.

Princip výpočtu spočívá ve výpočtu rozdílů hodnot mezi dvěma soubory, těmto rozdílům opět přiřadíme v absolutní hodnotě pořadí a princip je obdobný **jednovýběrovému Wilcoxonovu testu**. Tento výpočet opět platí pro $n < 20$.

Mann-Whitney U test pro nezávislé výběry

Neparametrický test pro dva nezávislé výběry. Tento test nevyžaduje, aby v obou skupinách byl stejný počet pozorování a princip jeho výpočtu je následující. Všechny jednotky uspořádáme vzestupně (za oba soubory dohromady) a přiřadíme jim pořadí. Následně sečteme pořadí patřící jednomu a druhému souboru (R_A a R_B). Následně můžeme spočítat testovací statistiky. Jako výsledek testového kritéria potom vybereme menší z testovacích statistik, tedy buď U_A nebo U_B .

Kruskalův-Wallisův test pro více výběrů

Jedná se o test, který je zobecněním Mann-Whitneyho testu a užívá se pro **3 a více výběrů**. Jedná se určitou neparametrickou variantu ANOVA testu, tento test lze využít pro výběry, které **nemají normální rozdělení**.

Postup opět spočívá na udělení pořadí všem hodnotám (za všechny skupiny) podle velikosti a následném součtu pořadí u jednotlivých skupin. Testová statistika vypadá následovně.



JEDNOVÝBĚROVÉ NEPARAMETRICKÉ TESTY

- 1) Při sledování sklizně brambor bylo pořízeno z 10 vzorků následující množství ztrát (v gramech). Posuďte na 5% hladině významnosti zda jsou v souboru odlehlá pozorování.

Číslo vzorku	1	2	3	4	5	6	7	8	9	10
Množství ztrát (v g)	23,5	18,5	19,6	18,7	21,4	22,4	19,1	13,9	25,2	5,2

- 2) Pomocí neparametrického testu otestujte, zda střední hodnota množství ztrát je 18 gramů.

Číslo vzorku	1	2	3	4	5	6	7	8	9	10
Množství ztrát (v g)	23,5	18,5	19,6	18,7	21,4	22,4	19,1	13,9	25,2	5,2



3) Liší se výsledky studentů studujících na ekonomické a technické škole?
Řešte pomocí vhodného neparametrického testu.

Ekonomická škola	Technická škola
60	96
95	74
47	58
60	86
85	60
90	76
75	62
83	84
71	77
75	80
95	84
86	87

4) 9 vzorků bylo postupně analyzováno 2 metodami. Zjistěte, zda existuje významný rozdíl v měření mezi metodou 1 a metodou 2. Použijte vhodný neparametrický test.

vzorek	metoda 1	metoda 2
1	2.62	2.51
2	2.6	2.41
3	2.41	2.39
4	2.85	2.9
5	2.91	2.88
6	2.7	2.73
7	2.68	2.54
8	2.93	2.89
9	2.85	2.75



DVOUVÝBĚROVÉ NEPARAMETRICKÉ TESTY

5) Výzkumník chce otestovat účinnost nového léku proti bolesti hlavy. Získá 20 dobrovolníků, náhodně je rozdělí do dvou skupin po 10 osobách: jedna skupina si domů odnese placebo, druhá testovaný lék (ani účastníci, ani výzkumník nevědí, kdo je ve které skupině). Účastníci studie si mají vzít lék ve chvíli, kdy je začne bolet hlava a zaznamenat, jak dlouho poté bolest trvala (kolik minut). Řešte pomocí vhodného **neparametrické testu**.

Skupina s placebem	Skupina s testovaným lékem
95	75
85	60
100	30
120	65
80	100
90	70
85	40
80	55
75	65
120	110



6) Ve 3 lokalitách bylo náhodně osloveno celkem 15 respondentů, u nichž bylo zjišťováno, kolikrát navštíví hypermarket potravin v rámci jednoho měsíce. Pomocí vhodného testu rozhodněte, zda je počet návštěv hypermarketu ovlivněn lokalitou hypermarketu. Test proveďte na 5% hladině významnosti. Řešte pomocí vhodného **neparametrického testu**.

Lokalita	Počet návštěv
A	2,3,5,4,1
B	3,4,6,8,3
C	4,5,7,1,9



REGRESE A KORELACE

Zabýváme se dvojicí veličin:

Y - vysvětlovaná, závisle proměnná

X - vysvětlující, nezávisle proměnná

Hledáme vysvětlení chování Y při dané hodnotě X.

Koeficient determinace (Index determinace):

$$R^2 = I^2 = \frac{S_T}{S_y}$$

kolik % variability **y** v daném souboru jsme vysvětlili zvoleným regresním modelem.

Korelační koeficient:

$$r_{yx} = r_{xy} = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \sqrt{n \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2}} = \frac{\overline{xy} - \bar{x} \bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)}} = \frac{s_{xy}}{s_x s_y}$$

měří sílu (intenzitu) lineární závislosti, nabývá hodnot z intervalu <-1, 1>

1) V tabulce jsou uvedeny roční náklady na údržbu (v dolarech) a cena domu (v tis. Dolarů).

Náklady	835	63	240	1005	184	213	313	658	195	545
Cena	136	24	52	143	42	43	67	106	61	99

- Modelujte závislost nákladů na údržbu na ceně regresní přímkou.
- Zhodnoťte kvalitu modelu pomocí koeficientu determinace.
- Interpretujte věcně hodnotu regresního koeficientu b_1 .
- Odhadněte střední hodnotu nákladů u domů za 80 tis. dolarů



2) Sledujeme závislost mezi výškou kmene (m) a objemem dřeva (m³) u 8 náhodně vybraných stromů:

Výška	14	8	12	19	13	9	21	10
objem	4,2	3	4	4,9	4,4	3,2	5,3	3,9

Vypočítejte Pearsonův korelační koeficient (zaokrouhlete na čtyři desetinná místa).

3) Vypočítejte hodnotu korelačního koeficientu, pokud znáte sdružené regresní rovnice:

$$y' = 16,9 - 1,12x \quad x' = 4,7 - 0,67y$$

4) Vypočítejte lineární rovnici pro tento vztah (závisle proměnná je y).

průměr x = 9, průměr y = 19, směrodatná odchylka x = 1,8, směrodatná odchylka y = 2,6,
korelační koeficient = 0,74.



5) Tabulka obsahuje údaje o stáří, počtu najetých km a ceně 20 ojetých aut značky Octavia Combi.

- sestavte dvojnásobný reg. model pro závislost ceny na stáří a počtu najetých km.
- Vypočtete koeficient úplné vícenásobné korelace.
- Z kolika % zle změny v ceně vysvětlit nalezeným reg. modelem.
- použijte reg. model k odhadu ceny auta starého 6 let, které má najeto 60 tis.km

Stáří	Najeto	Cena
(roky)	(tis. km)	(tis. Kč)
3	106	167
4	134	139
4	51	159
5	102	135
5	125	139
6	104	139
6	49	139
6	74	145
7	156	109
7	147	119
7	59	129
7	83	135
8	137	99
8	91	99
8	114	109
8	97	119
9	298	63
9	165	69
9	172	76
9	145	77

	Průměr	Směr. odchylka
stari	6,7500	1,83174
najeto	120,45	55,611
Cena_v_tis	118,2500	29,93831

Matice párových koeficientů korelace

		stari	najeto	Cena_v_tis
stari	Pearson Correlation	1	,486*	-,908**
najeto	Pearson Correlation	,486*	1	-,728**
Cena_v_tis	Pearson Correlation	-,908**	-,728**	1



VÍCENÁSOBNÁ REGRESE – Převzaté zadání

Hotelová společnost vlastní 10 hotelů. Popište závislost celkových měsíčních tržeb (v mil Kč) na měsíčních tržbách vyprodukovaných stravovacími úseky (v mil Kč) a na počet obsazených lůžek v konkrétním měsíci.

Měsíční tržby celkem (mil. Kč) y	Tržba za stravování (mil Kč) x_1	Počet obsazených lůžek x_2
12	2	150
8	1,2	94
17	8,3	254
21,3	8,4	399
10	3	95
12,5	4,8	149
25	11,5	247
38,6	14,2	400
88	16,1	952
10	2,5	120

- a) Odhadněte parametry vícenásobné lineární regresní funkce popisující závislost celkových měsíčních tržeb na měsíčních tržbách vyprodukovaných stravovacími úseky (v mil Kč) a na počtu obsazených lůžek v konkrétním měsíci
- b) Charakterizujte těsnost závislosti
- c) Zjistěte, jaké celk. měsíční tržby může společnost očekávat u objektu kde obvyklé měs. tržby vyprodukované stravovacími úseky jsou 5 mil. Kč a obvyklý počet obsazených lůžek je 165

$$n = 10; \sum y = 242,4; \sum x_1 = 72; \sum x_2 = 2860; \sum y * x_1 = 2721,04; \sum y * x_2 = 124772,2; \sum x_1 * x_2 = 31020,5; \sum x_1^2 = 776,28; \sum x_2^2 = 1427992; r_{x_1x_2} = 0,8315; r_{yx_1} = 0,8354; r_{yx_2} = 0,9760$$

$$b_0 n + b_1 \sum x_{1i} + b_2 \sum x_{2i} = \sum y_i$$

$$b_0 * 10 + b_1 * 72 + b_2 * 2860 = 242,4$$

$$b_0 \sum x_1 + b_1 \sum x_{1i}^2 + b_2 \sum x_{1i}x_{2i} = \sum x_{1i}y_i$$

$$b_0 * 72 + b_1 * 776,28 + b_2 * 31020,5 = 2721,04$$

$$b_0 \sum x_2 + b_1 \sum x_{1i}x_{2i} + b_2 \sum x_{2i}^2 = \sum x_{2i}y_i$$

$$b_0 * 2860 + b_1 * 31020,5 + b_2 * 1427992 = 124772,2$$

$$b_0 = -2,5648$$

$$b_1 = 0,3507$$

$$b_2 = 0,0849$$

b) Charakterizujte těsnost závislosti

$$r_{yx_1.x_2} = \frac{r_{yx_1} - r_{yx_2} * r_{x_1x_2}}{\sqrt{(1 - r_{yx_2}^2) * (1 - r_{x_1x_2}^2)}}$$

$$r_{yx_2.x_1} = \frac{r_{yx_2} - r_{yx_1} * r_{x_1x_2}}{\sqrt{(1 - r_{yx_1}^2) * (1 - r_{x_1x_2}^2)}}$$

$$r_{yx_1.x_2} = \frac{0,8354 - 0,9760 * 0,8315}{\sqrt{(1 - 0,9760^2) * (1 - 0,8315^2)}} = 0,19719$$

$$r_{yx_2.x_1} = \frac{0,9760 - 0,8354 * 0,8315}{\sqrt{(1 - 0,8354^2) * (1 - 0,8315^2)}} = 0,9215$$

Parciální koeficient $r_{yx_1.x_2}$ udává sílu závislosti mezi celkovou výší měsíční tržby a měsíční tržbou za stravování při konstantním počtu obsazených lůžek – závislost výše celkových tržeb na výši tržeb za stravování je přímá a slabá.

Úplný korelační koeficient

$$R_{yx_1x_2} = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2 * r_{yx_1} * r_{yx_2} * r_{x_1x_2}}{1 - r_{x_1x_2}^2}}$$

$$R_{yx_1x_2} = \sqrt{\frac{0,8354^2 + 0,9760^2 - 2 * 0,8354 * 0,9760 * 0,8315}{1 - 0,8315^2}} = 0,976944$$



Závislost mezi výší celkovou měsíční tržbou, měsíční tržbo za stravování a počtem obsazených lůžek je velmi silná.

c) (stravovací úsek tržby 5 mil., lůžka 165)

$$\hat{y}_i = b_0 + b_1x_1 + b_2x_2$$

$$\hat{y}_i = -2,5648 + 0,3507x_1 + 0,0849x_2$$

$$\hat{y}_i = -2,5648 + 0,3507 \cdot 5 + 0,0849 \cdot 165$$

$$\hat{y}_i = 13,1972 \text{ mil. Kč}$$

Společnost může očekávat celkové měsíční tržby ve výši 13,1972 mil. Kč při měsíčních tržbách za stravování ve výši 5 mil, Kč a při počtu obsazených lůžek v daném měsíci 165.

Vysvětlete rozdíl mezi párovým korelačním koeficientem a parciálním korelačním koeficientem:

Parciální (dílčí): $R_{yx_1x_2}$ měří sílu závislosti mezi y a x_1 za předpokladu, že ostatní proměnné jsou konstantní (nemají vliv).

Párový: r_{yx_1} párový korelační koeficient měří sílu závislosti mezi proměnnou x_1 a y .

Popište, co vyjadřuje vícenásobný koeficient determinace u vícenásobné závislosti:

Vyjadřuje kolika procenty je ovlivněna změna závislé proměnné Y nezávisle proměnnými $x_1, x_2, \dots, x_n$

Vícenásobný koeficient determinace nabývá jakých hodnot?

Udává se v % a říká nám % vysvětlené variability. Leží v intervalu od 0 do 1.



Index hospodářského zatížení:

$$IHZ = \frac{I.e.g. + II.e.g. + III.e.g.}{II.e.g.}$$

Vážený index hospodářského zatížení:

$$IHZ = \frac{0,7 * I.e.g. + II.e.g. + 0,7 * III.e.g.}{II.e.g.}$$

Index závislosti mladých (zelené zatížení):

$$IZM = \frac{I.e.g.}{II.e.g.}$$

Index závislosti starých (šedé zatížení):

$$IZS = \frac{III.e.g.}{II.e.g.}$$

Porovnejte podíl obyvatel v předproduktivním věku mezi dvěma městy A a B, když máte k dispozici následující údaje:

Ukazatel	Obec A	Obec B
Vážený index hospodářského zatížení	1,345	1,432
Počet obyvatel v produktivním věku	9550	8320
Počet obyvatel v postproduktivním věku	2312	2410

Index závislosti mladých v nejmenovaném kraji je 0,450. V kraji žije 678 940 obyvatel, z toho 154 340 obyvatel v předproduktivním věku. Vypočítejte počet obyvatel v postproduktivním věku v daném kraji.



DODATEK

TESTY DOBRÉ SHODY A NORMÁLNÍ ROZDĚLENÍ

Chí-kvadrát – test dobré shody

slouží ke statistickému testování shody mezi očekávanými a pozorovanými hodnotami. Jednoduše testujeme, zda se pozorované hodnoty shodují s teoretickými (očekávanými). Zda odchylka odhadnuté hodnoty od té skutečně naměřené vznikla jen náhodně nebo zda byl odhad špatný.

Chí-kvadrát testem vypočítaná hodnota se pak srovnává s kritickou hodnotou odpovídající zvolené hladině významnosti (nejčastěji 5%) při daném počtu stupňů volnosti.

Chceme například prokázat:

Jsou muži a ženy ve skupině zastoupeni rovnoměrně (tedy 50/50%)

Jsou výrobky dle jakosti zastoupeny v poměru 3:1:1 (60:20:20%)

Zda data mají normální rozdělení

- 1) Nechtě při průzkumu známek studentů bylo 50 studentů s jedničkou, 70 studentů s dvojkou, 200 studentů s trojkou a 30 studentů zkoušku neudělalo, dostalo 4. Zároveň se dle předešlých let očekává následující rozdělení: Jedničku bude mít 10% studentů, dvojkou 20%, trojkou 50% a čtyřku 20% studentů. Odpovídá průzkum očekávání?

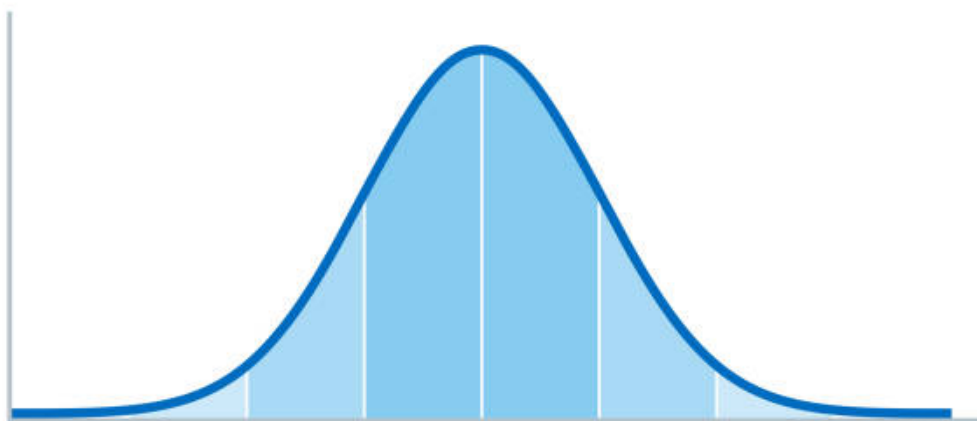
Normální rozdělení

Normální rozdělení je možné otestovat například Testem dobré shody, Davidovým testem a nebo použít Kolmogorovův-Smirnovův test. Pokud data mají normální rozdělení, potom jsou rozděleny dle Gaussovy křivky, kterou je možné vidět níže. Dále také platí, že v případě normality jsou data symetricky rozmístěná kolem střední hodnoty a platí, že: průměr = medián = modus.

V případě, že data normální rozdělení nemají, je nutné k vyřešení hypotéz používat **neparametrické testy**.

H0: data mají normální rozdělení

H1: data nemají normální rozdělení



2) Soubor obsahuje 40 hodnot. Pochází data z normálního rozdělení $N(10,16)$?

Třídy	Četnosti	Teor. Pravděp.
$x \leq 6$	5	0,1586553
$6 < x \leq 8$	7	0,1498823
$8 < x \leq 10$	9	0,1914625
$10 < x \leq 12$	8	0,1914625
$12 < x \leq 14$	7	0,1498823
$14 < x$	4	0,1586553