

https://dl1.cuni.cz/pluginfile.php/1089461/mod_resource/content/1/skripta_Base_24_Analyzy.pdf

PŘEDNÁŠKA 9. 3.

V knihovně je průzkum "Jaké knihy si půjčujete na dovču?". Odpovědí 800 náhodně vybraných, roztržďeně na pohlaví dle tabulky; na hladině významnosti 5 % zda závisí na pohlaví, příp. sílu závislosti

Pohlaví	Co si berete na dovču?					n _i
	Populárně naučné	Průvodci	Romány	Časopisy	Nic	
žena	20 15,3	82 96,9	126 193,8	156 112,2	126 91,8	510
muž	4	70	178	20	18	290
n. _j	24	152	304	176	144	800

Metody zkoumání závislosti

- 1) Analýza rozptylu > podmíněné průměry
- 2) Test nezávislosti kategorií znaků > OK
- 3) Regresní analýza > ne
- 4) Korelační > ne

1) H₀: odpověď závisí na pohlaví čtenáře

H₁: non H₀

2) G = vzorec

3) W = vzorec

W = (G; G > X 2 0,95 (4))

W = (G; G > 9,488) - dle tabulky

4) G = 158,052

5) G náleží W --> zamítáme H₀, přijímáme H₁

Na hladině významnosti 5 % že odpověď na otázku závisí na pohlaví čtenáře.

$\alpha = 0,05$

$r = 2$

$s = 5$

$x (4)$

$(r - 1) * (s - 1) = (2 - 1) * (5 - 1) = 4$

$$N_{11} = n_{1.} \cdot n_{.1} / n = 510 \cdot 24 / 800 = 15,3 - \text{teoretická četnost}$$

Bez závislosti by si 15,3 žen vybranou populárně naučnou.

$$= 510 \cdot 152 / 800 = 96,9 - \text{teoretická četnost}$$

Bez závislosti by si 96,9 žen vybralo průvodce.

$$= 510 \cdot 126 / 800 = 193,8 \text{ (pro romány)}$$

$$= 510 \cdot 156 / 800 = 112,2 \text{ (pro časopisy)}$$

$$= 510 \cdot 126 / 800 = 91,8 \text{ (pro nic)}$$

Muži též

Test o nezávislosti kategoriálních znaků

- Jestliže byla prokázána mezi proměnnými závislost, můžeme zjišťovat sílu této závislosti.
- Použit pearsonův koeficient (namísto cramerova koeficientu)
- $C_c = \sqrt{G / n} = \sqrt{158,052 / 800} = 0,444$

Závislost odpovědi na otázku a pohlaví čtenáře je spíše slabší.

Asociační tabulka

$x_i \backslash y_j$	y_1	y_2	...	y_k	Součty četností $n_{i.}$
x_1	n_{11}	n_{12}	...	n_{1k}	$n_{1.}$
x_2	n_{21}	n_{22}	...	n_{2k}	$n_{2.}$
...	...	...	...	...	...
x_r	n_{r1}	n_{r2}	...	n_{rk}	$n_{r.}$
Součty četností $n_{.j}$	$n_{.1}$	$n_{.2}$	...	$n_{.k}$	n

1) H_0 :

H_1 :

2) Testové kritérium $G = n \cdot \text{vzorec}$

3) Kritický obor $W = (\text{vzorec})$

Míra těsnosti závislosti

K měření těsnosti závislosti jevů A a B se používá koeficient asociace r_{AB}

r_{AB} = vzorec

$r_{AB} = 1$ --> úplná kladná asociace (úplná přímá závislost)

$r_{AB} = (-1)$ --> úplná záporná asociace (úplná nepřímá závislost) - "pokud nastane A, nenastane B"

$r_{AB} = 0$ --> nezávislost znaků

Příklad

200 náhodných respondentů ... dotázání zda je výrobek kvalitní a líbí se jím reklama ... viz tabulka; na hladině významnosti 5 % zda posouzení kvality a líbí reklama na sobě závisí; změřte sílu a směr dané závislosti pomocí vhodné charakteristiky

Je výrobek kvalitní?	Líbí se reklama?		n _i *
	ANO	NE	
ANO	75	50	125
NE	20	55	75
n _j	95	105	200

Test o nezávislosti kategoriálních znaků

1) H_0 : hodnocení kvality výrobku a atraktivity reklamy závisí

H_1 : non H_0

2) $G = n \cdot \text{vzorec}$

3) $W = (G; G > 3,84)$

4) $G = 200 \cdot (75 \cdot 55 - 50 \cdot 20)^2 / 125 \cdot 75 \cdot 95 \cdot 105 = 20,886$

5) G náleží W --> zamítáme H_0 , přijímáme H_1

Na hladině významnosti 5 % jsme prokázali, že hodnocení kvality výrobku a atraktivity reklamy na sobě závisí.

$r_{AB} = 75 \cdot 55 - 50 \cdot 20 / \sqrt{125 \cdot 75 \cdot 95 \cdot 105} = 0,323$

Závislost mezi proměnnými je spíše slabší a přímá.

Znamená, že reklama částečně napomáhá, ale ne moc.

Analýza závislosti číselných proměnných (regresní a korelační analýza)

Rozlišujeme 2 základní typy závislosti mezi proměnnými:

- **Pevná (funkční)** – změna jednoho znaku vyvolá jednoznačnou změnu druhého znaku podle konkrétního funkčního znaku

- **Volná (statistická)** – změna jedné veličiny vyvolá změny druhé veličiny tak, že určité hodnotě jednoho znaku může odpovídat více různých hodnot druhého znaku

Korelační tabulka

- Dvourozměrná tabulka

$x_i \backslash y_j$	y_1	y_2	...	y_k	Součty četností $n_{i\bullet}$
x_1	n_{11}	n_{12}	...	n_{1k}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	...	n_{2k}	$n_{2\bullet}$
...	...	...	...	...	...
x_r	n_{r1}	n_{r2}	...	n_{rk}	$n_{r\bullet}$
Součty četností $n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet k}$	n

CVIČENÍ 10. 3.

Příklad – stanovení minimálního rozsahu výběru

Určete rozsah výběru osob, kterého by bylo zapotřebí ke zjištění – se spolehlivostí 0,95 a chybou 0,02 - podílu kuřáků, předpokládá-li se, že tento podíl bude okolo 30 %.

$$1 - \alpha = 0,95$$

$$\Delta = 0,02$$

$$p = 0,3 \text{ (ve vzorci, pokud nevím 0,5)}$$

$$n \geq 1,96^2 \cdot 0,3 \cdot 0,7 / 0,02^2$$

$$n \geq 2016,84$$

Potřebujeme data od 2017 respondentů.

- Klasický rozptyl
- Výběrový rozptyl - budeme používat

Odhad parametru

- Bodový odhad
- Intervalový odhad

Příklad - odhad parametru o 2 normálního rozdělení

Sestrojte oboustranný 95% interval spolehlivosti pro rozptyl normálně rozdělené náhodné veličiny na základě následujících údajů - jsou zadány hodnoty:

42, 48, 60, 43, 36, 50, 52, 38, 56, 45

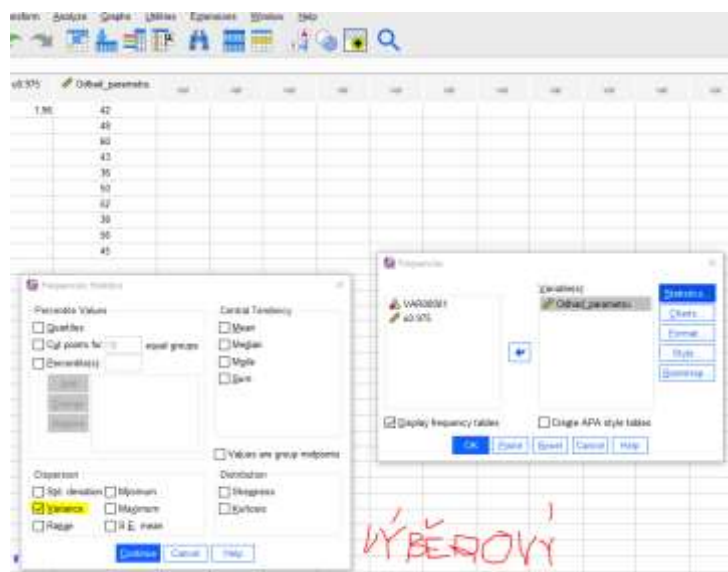
$n = 10$

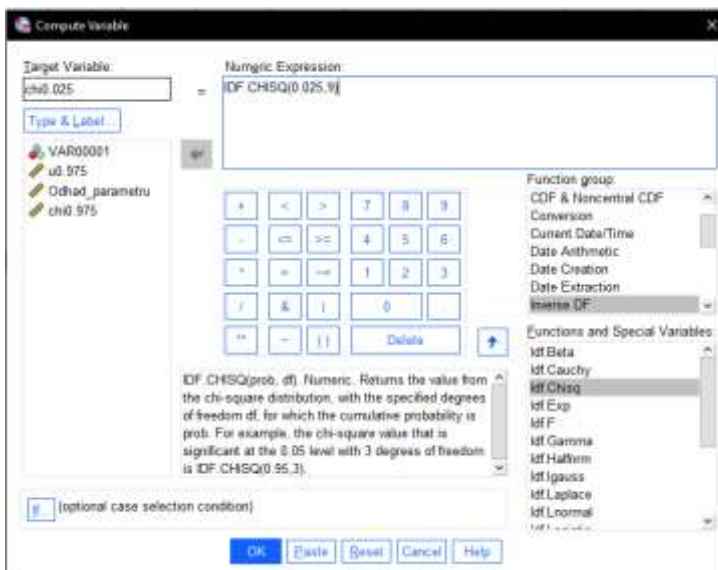
$1 - \alpha = 0,95$

$\alpha/2 = 0,025$

$1 - \alpha/2 = 0,975$

Hodnota výběrového rozptylu je potřeba dopočítat z výběrových dat.





chi0.975	chi0.025
19,023	2,70

$$P(9 * 59,11 / 19 < o2 < 9 * 59,11 / 2,7) = 0,95$$

$$P(27,99 < \sigma^2 < 197,03) = 0,95$$

Příklad – odhad parametru σ^2 normálního rozdělení

Zadání příkladu

Pevnost proužků textilie je náhodná veličina s normálním rozdělením. Tato pevnost bylo změřena u 25 náhodně výběrových vzorků. Z tohoto výběru byly vypočteny tyto výběrové charakteristiky: $\bar{x} = 85,1$; $s' = 3,25$. Určete se spolehlivostí 99 %, v jakých mezích se bude pohybovat rozptyl a směrodatná odchylka pevnosti.

$$1 - \alpha = 0,99$$

$$\alpha = 0,01$$

$$\alpha/2 = 0,005$$

$$1 - \alpha/2 = 0,995$$

$$\chi^2 = 9,985$$

TECHNICKÁ UNIVERZITA V LIBERCI

Odhad parametru σ^2 normálního rozd.

$$\chi^2_{1-\frac{\alpha}{2}}(n-1) = \chi^2_{0,995}(24) = 45,559$$

$$\chi^2_{\frac{\alpha}{2}}(n-1) = \chi^2_{0,005}(24) = 9,886$$

$$P\left(\frac{(n-1)s'^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} < \sigma^2 < \frac{(n-1)s'^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right) = 1 - \alpha$$

$$P\left(\frac{24 \cdot 3,25^2}{45,559} < \sigma^2 < \frac{24 \cdot 3,25^2}{9,886}\right) = 0,99$$

Chyba I. druhu = zamítáme H_0 i když ve skutečnosti platí, pravděpodobnost $P(I) = \alpha$

Chyba II. druhu = $P(II) = \beta$

1 - β = síla testu

Zadání příkladu

Pevnost proužků textilie je náhodná veličina s normálním rozdělením. Tato pevnost byla změřena u 20 náhodně vybraných vzorků. Z tohoto výběru byly vypočteny tyto výběrové charakteristiky

$n = 20$

$\bar{x} = 85,1$

$S' = 3,25$

$\alpha = 0,05$

Testujte na hladině významnosti 5 %, zda je průměrná pevnost proužků textilie 88 jednotek.

1) $H_0: \mu = 88$

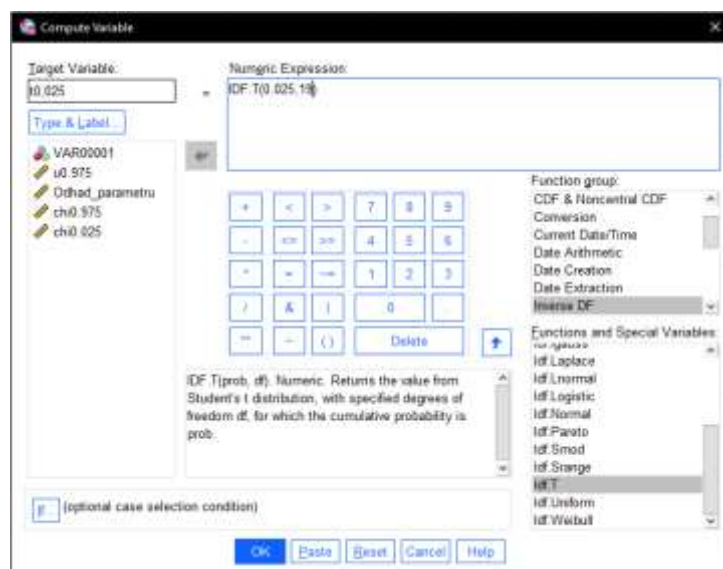
$H_1: \mu \neq 88$

2)
$$t = \frac{\bar{x} - \mu_0}{\frac{s'x}{\sqrt{n}}}$$

3) $W =$

$W =$

$t_{0,975}(19) = 2,093$



 t0.025	 t0.975
-2,093	2,093

$W = (t; t \leq -2,093 \cup t \geq 2,093)$

4) $t = -3,99$

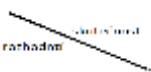
h) $t \in W \rightarrow H_0$ zamítáme, přijímáme H_1

Na hladině významnosti 5 % jsme prokázali, že pevnost proužků textilie není 88 jednotek.

SEMÍNÁŘ 14. 3.

- Bodově
- Intervalově
- Hypotézy

$H_1 \rightarrow$ jaké máme podezření

	H_0	H_1
Nezamítáme H_0	ok	Chyba 2. druhu
Zamítáme H_0	Chyba 1. druhu	ok

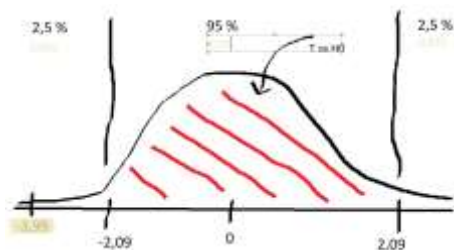
$$T = \frac{\bar{x} - \mu}{S_x} \sqrt{n}$$

Zamítáme H_0 proti $H_1: \mu \neq \mu_0$, pokud $|T| \geq t_{n-1}(1-\alpha)$

Příklad:

$$T = \frac{85,1 - 88}{3,25_x} \sqrt{20}$$

$3,99 (T) \geq 2,09 (\text{Idf.T})$



Příklad 2:

- 1) Určitá linka městské autobusové dopravy jezdí v době dopravní špičky v centru města průměrnou rychlostí 8 km/h. Vedení dopravního podniku uvažovalo o tom, zda by změna trasy dané linky vedla ke zvýšení průměrné rychlosti. Nová trasa proto byla projeta v deseti náhodně vybraných dnech v době dopravní špičky a byly zjištěny tyto průměrné rychlosti (v km/h):

8,5 9,5 7,8 8,2 9,0 7,5 8,2 7,8 9,0 8,5

Rozhodněte na 5 %-ní hladině významnosti, zda změna trasy vede ke zvýšení průměrné rychlosti, přičemž předpokládejte normální rozdělení rychlosti autobusu v centru v době dopravní špičky. 2 BODY

$$> \alpha = 0,05$$

$$H_0: \mu = 8 \text{ km/h}$$

$$H_1: \mu > 8 \text{ km/h}$$

T-Test						
One-Sample Statistics						
	N	Mean	Std. Deviation	Std. Error Mean		
deprava	10	8.4000	.82893	.19689		
One-Sample Test						
				Test Value = 8		
	t	df	Significance	Mean Difference	95% Confidence Interval of the Difference	
			One-Sided p	Two-Sided p	Lower	Upper
deprava	2.911	9	.038	.075	-.0088	.8498
One-Sample Effect Sizes						
	Standardized ^a	Point Estimate	95% Confidence Interval			
			Lower	Upper		
deprava	Cohen's d	.82893	.638	1.308		
	Hedges' correction	.68820	.591	1.193		

a. The denominator used in estimating the effect sizes.
Cohen's d uses the sample standard deviation.
Hedges' correction uses the sample standard deviation, plus a correction factor.

$$(T \geq t_{0,95} (9))$$

$$(2,011 \geq 1,893 (9))$$

P – hodnota je menší než 2,011 takže je to v těch 95 %

PŘEDNÁŠKA 16. 3.

Závisle proměnná – y

Nezávisle proměnná – x

Příklad:

Máme údaje o dvou číselných proměnných x a y. Uspořádejte tato data do vhodné tabulky a vypočítejte hodnoty podmíněných průměrů a podmíněných rozptylů proměnné y.

x	3	6	1	6	1	1	6	3	6	1
y	4	5	4	2	2	4	4	5	2	5

Korelační tabulka

xi	y			ni.	yi	S ² i
	2	4	5			
1	1	2	1	4	3,75	1,1875
3	0	1	1	2	4,50	0,2500
6	2	1	1	4	3,25	1,6875
n.j	3	4	3	10	x	X

$$y = 2 \cdot 0 + 4 \cdot 1 + 5 \cdot 1 / 2 = 4,5$$

Čára podmíněných průměrů a podmíněných rozptylů

Tabulka s daty domácnostní podle počtu členů a spotřeby vody. Byly vypočítány hodnoty podmíněných průměrů a podmíněných rozptylů spotřeby vody. Vhodným způsobem tyto charakteristiky graficky znázorněte.

Domácnost			y	S
1	80,40,60			266,67
2	70,125,51,106			846,5
3	101,79,136			550,89
4	125,180,91,130,152			874,
5 a více	230,168			
Součet				

Regresní analýza

Lineární závislost – její průběh je možné popsat pomocí přímky

Postup regresní analýzy

1) Volba typu regresní funkce

- při volbě typu regresní funkce lze uplatnit různá kritéria (grafické znázornění, matematicko-statistická kritéria, věcně-logický rozbor zkoumané závislosti)

- volba by se měla v první řadě opírat o určitou teorii, tzn. Vyplyvat z věcného rozboru vztahů proměnných

- vždy se snažíme o jednoduchost modelu (ne příliš mnoho parametrů)

- je třeba změřit vhodnou charakteristickou přilnavost regresní funkce k datům

2) Odhad parametrů zvolené regresní funkce

- matematické modely, které vyjadřují představu o průběhu závislosti proměnných;

- umožňují odhady neznámých hodnot, závisle proměnné y ze známých hodnot nezávisle proměnné x , resp. nezávisle proměnných x_1, x_2

- obecný tvar regresního modelu

Empirickou regresní funkci lze považovat za odhad teoretické regresní funkce.

Regresní funkce

- Teoretický (hypotetický)
- Výběrová (empirická)

Typy modelu

1) aditivní

2) multiplikativní

Teoretická regresní funkce:

- existují různé typy regresních funkcí

- nejčastější jsou lineární regresní funkce

- linearita se může hodnotit jak z hlediska proměnných, tak z hlediska parametrů

- každá regresní funkce má určitý počet parametrů (jejich počet je k)

Funkce lineární z hlediska parametrů

Funkce nelineární z hlediska parametrů

3) Testování hypotéz o parametrech regresní funkce

4)

CVIČENÍ 17. 3.

Příklad: Výběrový soubor o rozsahu 20 jednotek, jehož prvky jsou plochy o rozměru 1m², na nichž se zjišťuje hmotnost neskliženého zrní v gramech. Hmotnost neskliženého zrní je náhodná veličina, která má normální rozdělení. Posuďte na hladině významnosti 5 %, zda je střední hmotnost neskliženého zrní nižší než 7 g na m².

3,2	8,7	6,5	5,6
6,1	9,5	5,5	6,3
8	6,2	5,5	6,2
3,2	8,5	5,6	6,8
3,2	6,6	7,8	6,0

$n = 20$

$\alpha = 0,95$

$\mu = 7$

Standartní testovací postup:

1) $H_0: \mu = 7$

$H_1: \mu < 7$

Okomentoval(a): [JH1]: Dle vlastností testu hypotéz

Dále je třeba stanovit testové kritérium:

2)
$$t = \frac{|\bar{x} - \mu_0|}{\frac{s}{\sqrt{n}}}$$

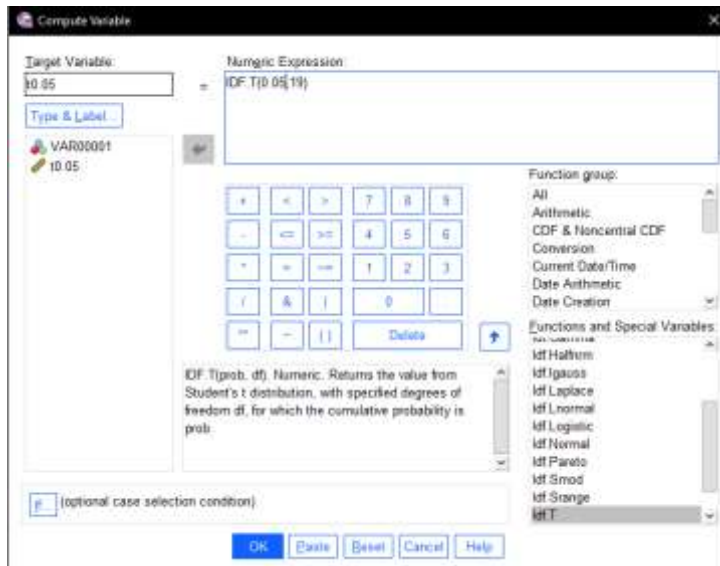
Okomentoval(a): [JH2]: malý rozsah

Stanovení kritického oboru:

3) $W = (t, t \leq t_{\alpha}(n-1))$

$W = (t, t \leq t_{0,05}(19))$

$W = (t, t \leq -1,729)$



$= -1,729$ (kritický obor)

Excel: `=T.INV(0,05;19)` = -1,729

4) $t = -1,692$

5) $t \notin W \rightarrow$ nezamítáme H_0

nepřijímáme H_1



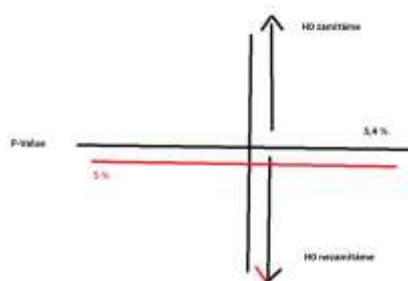
One-Sample Test						
	Test Value = 7					
	Significance		Mean		95% Confidence Interval of the Difference	
	t	df	One-Sided p	Two-Sided p	Lower	Upper
hmotnost	-1,692	19	,054	,107	-1,3423	,1423
	Mean Difference					
	-60000					

Minimální hladina významnosti pro zamítnutí nulové hypotézy.

$P-V = 0,054$

Porovnávám $P - V$ s α

$P - V > \alpha \rightarrow$



Na hladině významnosti 5 % jsme neprokázali, že by střední hmotnost nesklizeného zrní v gramech byla nižší než 7 g na m².

Test parametru π alternativního rozdělení

Je třeba mít k dispozici výběr dostatečně velkého rozsahu.

Příklad: Výrobce stanovil záruku 2 roky. Předpokládá, že bude opravovat 15 % výrobků. Na hladině významnosti 1 % ověřte, zda výrobce vycházel ze správného předpokladu, jestliže z 900 prodaných výrobků bylo reklamováno 150.

$\alpha = 0,99$

$n = 900$

$m = 150$

$n > 9 / \pi(1-\pi)$

$n > 9 / 0,167(1 - 0,167)$

$p = m/n = 150/900 = 0,167$

1) $H_0: \pi = 0,15$

H1: $\pi \neq 0,15$

2) U = vzorec

3) W = vzorec

$W = (U; U \leq -2,576 \cup U \geq 2,576)$

4) U = 1,429

Mean = střední hodnota normovaného normálního rozdělení (zde 0)ú

Kritické hodnoty = -2,576

$W = (U; U \leq -2,576 \cup U \geq 2,576)$

U = 1,429 --> není v kritickém oboru

H0 nezamítáme, H1 nepřijímáme

Na hladině významnosti 1 % nezamítáme předpoklad výrobce, že během záruční doby bude potřeba opravit 15 % výrobků.

IBM SPSS

1) H0: $\pi = 0,15$

H1: $\pi \neq 0,15$

2) Ověření podmínky výběru

x	n
0	750
1	150



3) U = 1,4

P-V = 0,161

4) $P-V > \alpha$ --> nezamítáme H0, nepřijímáme H1

SEMINÁŘ 21. 3.

O produkci podniku, který vyrábí dva druhy výrobků, máte v následující tabulce tyto údaje za rok 2000 a 2001:

Výrobek	Hodnota výroby v tis. Kč		Individuální jednoduchý index množství roku 2001/2000
	2000	2001	
postelky	60	88	1,6
řísy	72	57,6	0,9

Určete, jak se celkově změnila úroveň prodejních cen ve firmě, jak se změnilo množství vyrobeného a prodaného zboží ve sledovaném období a jak se změnila hodnota výroby. Dále zhodnoťte, jak se na změně hodnoty výroby podílel faktor cen a jak faktor prodaného množství výrobků.

3 BODY

$$P\left[\pi < p + u_{1-\alpha} \cdot \sqrt{\frac{p(1-p)}{n}}\right] = 1 - \alpha$$

$$n \geq \frac{u_{1-\frac{\alpha}{2}}^2 \cdot \pi(1-\pi)}{\Delta^2}.$$

$$u = 1,96^2$$

$$\pi = 0,2$$

$$\Delta = 0,02$$

PŘEDNÁŠKA 23. 3.

Jednoduchá regresní analýza

Odhad parametrů regresní funkce

2) Funkce nelineárních rozměrů

Metoda nejmenších čtverců

- Vychází z podmínky:

Regresní přímka

Popište pomocí regresní přímky závislost spotřeby vody v domácnosti na typu domácnosti dané počtem členů.

i	x	y	xy	x	i	x	y	xy	x
1	1	80	80	1	10	3	136	408	9
2	1	40	40	1	11	4	125	500	16
3	1	60	60	1	12	4	180	720	16
4	2	70	140	4	13	4	91	364	16
5	2	125	250	4	14	4	130	520	16
6	2	51	102	4	15	4	152	608	16
7	2	106	212	4	16	5	230	1150	25
8	3	101	303	9	17	5	168	840	25
9	3	79	237	9	suma	50	1924	6534	176

Regresní přímka

Koeficient $b_1 = 30,24$

Koeficient $b_0 = y - b_1x$

Koeficient $b_0 = 24,24$

$Y = 24,24 + 30,24x$

Interpretace:

Pokud vzroste počet členů domácnosti, vzroste průměrná spotřeba vody o ??? M3.

Postupným zlepšováním řešení pomocí iteračních postupů nalezneme odhad s požadovanou přesností.

Metody počátečního odhadu

- 1) Metoda vybraných bodů
- 2) - //- částečných součtů
- 3) Metoda apriorní informace
- 4) Metoda imeritizující transformace

Celkový F test je významný.

Možné výsledky testování

- 1) F-test významný a většina t-testů významný, pak můžeme danou regresní funkci považovat za vhodnou k popisu dané závislosti.
- 2) Pokud je F-test významný, ale většina t-testů není významná, pak je potřeba zvážit jestli daná regresní funkce nemá příliš velký počet regresorů a nebylo by vhodné některé z nich vypustit.
- 3) Pokud celkový F-test je nevýznamný, potom není vhodné danou regresní funkci k popisu dané závislosti použít a to ani tehdy, jsou-li některé z t-testů významné.

Regresní přímka (testy)

Zadání příkladu

Závislost spotřeby vody v domácnosti na typu domácnosti dané počtem členů byla popsána pomocí regresní přímky $Y = 24,24 + 30,24x$. Ověřte na hladině významnosti 5 % vhodnost této volby.

Vhodnost celého regresního modelu ověřujeme pomocí celkové F-testu:

- 1) $H_0: \beta_0 = c; \beta_1 = 0$ (tj. Zvolená přímka není vhodným modelem pro popis spotřeby vody na počtu členů domácnosti)

H_1 : non H_0

- 2) F = vzorec
- 3) W = vzorec

Na hladině významnosti 5 % jsme prokázali, že přímka je vhodná k popisu závislosti spotřeby vody na počtu členů domácnosti.

Regresní přímka (testy)

Vhodnost parametrů regresního modelu ověřujeme pomocí individuálních t-testů o nulové hodnotě regresních parametrů:

- 1) $H_0: \beta_0 = 0$

H_1 : non H_0

- 2) $t = b_0 / s_{b_0}$
- 3) $W =$

$W = (t; t \leq -2,131 \text{ } \geq 2,131)$

- 4) $t = 24,24 / 18,35 = 1,321$
- 5) t nenáleží W ; H_0 nezamítáme, H_1 nepřijímáme.

Na hladině významnosti 5 % nezamítáme předpoklad o tom, že parametr β_0 je statisticky nevýznamný.

Na hladině významnosti 5 % jsme prokázali, že parametr β_1 je statisticky významný.

Vzhledem k tomu, že celkový F-test je významný a jeden z t-testů také, je regresní přímku možné použít k popisu dané závislosti.

CVIČENÍ 24. 3.

Zácvik laboranta na optickém přístroji považujeme za ukončený, jestliže při opakovaném měření určitého objektu dosahuje směrodatné odchylky menší než 0,14 mm. Posuďte na hladině významnosti 5 %, zda můžeme zácvik laboranta považovat za ukončený, jestliže výsledky jeho měření jsou následující: 6,42; 6,44; 6,38; 6,6; 6,51. Předpokládáme normalitu rozdělení.

$$n = 7$$

$$\alpha = 0,05$$

$$S^{*2} =$$

$$1) \quad H_0: \sigma^2 = 0,14^2$$

$$H_1: \sigma^2 < 0,14^2$$

Dále je potřeba vybrat vhodné testové kritérium.

$$2) \quad \chi^2$$



$$= 1,635$$

$$3) \quad W = (\chi^2; \chi^2 \leq \chi^2_{\alpha} (n-1))$$

$$W = (\chi^2; \chi^2 \leq 1,64)$$

Variance	,01470
----------	--------

Test H o shodě 2 rozptylů

Pro porovnání rozptylů stipendií mezi studenty 2. a 4. ročníku VŠ byly provedeny náhodné výběry - ve 2. ročníku 10ti studentů a ve 4. ročníku 12ti studentů. Posuďte na hladině významnosti 5 %, zda jsou rozptyly jejich stipendií v obou ročnících stejné. Předpokládáme normalitu rozdělení stipendií.

2. ročník	2600	0	0	3200		0	1260	4400	4400	0	1520	
4. ročník	0	3600	2400	9800		0	2400	0	500	0	500	???

$V\alpha = 0,05$

$n_1 = 10$

$n_2 = 12$

1) H_0

H_1

2) F

3) $W = (F; F \leq 0,256 \cup F \geq 3,588)$

4) $F = 0,416$

5) F nenáleží $W \rightarrow H_0$ nezámítáme

$F_{0,025}(9,11)$

$F_{0,975}(9,11)$

Group Statistics

	dat	N	Mean	Std. Deviation	Std. Error Mean
data	2	10	1738,00	1807,748	571,660
	4	12	1841,67	2801,447	808,708

Independent Samples Test										
Levene's Test for Equality of Variances				t-Test for Equality of Means				95% Confidence Interval of the Difference		
	F	Sig.		t	df	Mean Difference	Std. Error Difference	Lower Bound	Upper Bound	
data	1,023	,320		-1,017	20	-60,667	59,5028	-171,168	50,833	
				-1,017	18,889	-60,667	59,5028	-171,168	50,833	

Independent Samples Effect Sizes

		Standardizer ^a	Point Estimate	95% Confidence Interval	
				Lower	Upper
data	Cohen's d	2405,626	-,043	-,882	,797
	Hedges' correction	2500,795	-,041	-,848	,766
	Glass's delta	2801,447	-,037	-,876	,803

a. The denominator used in estimating the effect sizes.

Cohen's d uses the pooled standard deviation.

Hedges' correction uses the pooled standard deviation, plus a correction factor.

Glass's delta uses the sample standard deviation of the control group.

SEMINÁŘ 28. 3. 2022

Příklad 1:

$n = 100$

$1 - \alpha = 0,95$

$X(\text{průměr}) = 150 \text{ g}$

$s^2 = 16 \text{ g}^2$

Bez sigmy

--> oboustranný interval

T-test = vždy normální rozdělení



One-Sample Statistics				
	N	Mean	Std. Deviation	Std. Error Mean
maien	7	150,00	2,160	,816

One-Sample Test						
Test Value = 0						
	t	df	Significance One-Sided p	Two-Sided p	Mean Difference	95% Confidence Interval of the Difference Lower Upper
maien	183,712	6	<.,001	<.,001	150,000	148,00 152,00

Příklad 2:

Třídění odpadů

Minimální podíl, který na pravděpodobnosti třídí --> levostranný interval spolehlivosti

$$p = 386 / 1000 = 0,386$$

$$\alpha = 0,05$$



$$u = 1,6449$$

Příklad 3:

Výrobce balí třtinový cukr

Průměr = 499 g

Směrodatná odchylka = 3,8 g

Deklarace, nepřekročí $\sigma < 5$ g

Stanovte mez, o které lze s 95ti % spolehlivostí prohlásit, že ji směrodatná odchylka hmotnosti cukru v balíčku nepřekročí. Náhodná veličina, normální rozdělení.

> intervalový odhad

H0: $\sigma \geq 5$ g

H1: $\sigma < 5$ g

5 g nenáleží (∞ , 4.864)

H0 zamítáme

Příklad 4:

Předpoklad 35 % kuřáků v populaci

n = 200

Kuřáci = 76

Hladina významnosti 5 %

PŘEDNÁŠKA 30. 3.

Postup regresní analýzy

1. Volba typu regresní funkce
2. Odhad parametrů zvolené regresní funkce
3. Testování hypotéz o parametrech regresní funkce
4. Ověření vhodnosti zvoleného regresního modelu

Kritéria pro posouzení kvality regresní funkce

1. Index determinace
 - Za nejvhodnější se považuje ta regresní funkce, u které je hodnota I^2 vyšší.
 - Udává, jaký podíl variability proměnné y lze vysvětlit zvolenou regresní funkcí (lze udávat i v %)
 - Je zároveň mírou těsnosti závislosti proměnné y na proměnné x.
 - Obecná míra, nezávislost na typu regresní funkce (lze použít i pro měření nelineární závislosti).
 - Tato míra není symetrická.
2. Testové kritérium F
 - Za vhodnější je považována ta funkce, u níž je hodnota F vyšší.
3. Residuální součet čtverců

- Za nejvhodnější považujeme
- 4. Reziduální rozptyl
- Za nejvhodnější považujeme funkci, která má reziduální rozptyl nižší. Reziduální rozptyl

Zadání příkladu:

Máme k dispozici data o proměnných x a y – viz tabulka. Vystihněte závislost proměnné y na proměnné x pomocí vhodné funkce (uvažujte přímku, parabolu, hyperbolu). Uvažujte $\alpha = 0,05$.

x	3	5	5	6	8	8	10	11	12	14
y	15	16	14	12	12	10	10	8	7	6

$$Y = b_0 + b_1x$$

$$Y = 18,568 - 0,923x$$

- 1) $H_0: \beta_0 = c; \beta_1 = 0$ (tj. Zvolená přímka není vhodným modelem pro popis závislosti proměnné y na proměnné x)

H_1 : non H_0

$$2) F = \frac{\frac{ST}{p-1}}{\frac{SR}{n-p}}$$

$$3) W = (F; F > F_{1-\alpha}(p-1, n-p))$$

$$W = (F; F > F_{0,95}(1,8))$$

$$W = (F; F > 5,318)$$

$$4) F = 95,0627/1 / 8,93728/8 = 85,09$$

$$5) F \text{ náleží } W \rightarrow \text{zamít. } H_0; \text{ přij. } H_1$$

Na hladině významnosti 5 % jsme prokázali, že přímka je vhodná k popisu závislosti proměnné y na proměnné x .

Vhodnost parametrů regresního modelu ověřujeme pomocí individuálních t-testů o nulové hodnotě regresních parametrů.

$$1) H_0: \beta = 0$$

H_1 : non H_0

$$2) t = b_0 / sb_0 \text{ (směrodatná chyba)}$$

$$3) W = \text{(vzorec)}$$

$$4) t = 18,568 / 0,8856 = 20,959$$

Na hladině významnosti 5 % jsme prokázali, že parametr β_0 je statisticky významný.

Dále budeme testovat významnost parametru β_1 :

$$1) H_0: \beta_1 = 0$$

H_1 : non H_0

2) $t = b_1 / s_{b1}$

3) $W = (\text{vzorec})$

$$W = (t; t \leq -2,306 \text{ u } t \geq 2,306)$$

Přijímáme hypotézu alternativní.

Celkový F test je významný.

Přímka je vhodným modelem.

Na hladině významnosti 5 % jsme prokázali, že parametr β_1 je statisticky významný.

Vzhledem k tomu, že celkový F-test je významný a oba t-testy také, je regresní přímku možné použít

Regresní hyperbola

$$Y = b_0 + b_1/x$$

$$Y = 5,632 + 35,893/x$$

Vhodnost celého regresního modelu (zde hyperboly) ověřujeme pomocí proměnné.

1) $H_0: \beta_0 = c; \beta_1 = 0$

$$H_1: \text{non } H_0$$

2) $F = ST/p-1 / SR/n-p$

3) $W = (F; F > 5,318)$

4) $F = 73,1063/1 / 30,8937/8 = 18,93$

5) F náleží $W \rightarrow$ zamítáme H_0 , přijímáme H_1

Na hladině významnosti 5 % jsme prokázali, že hyperbola je vhodná k popisu závislosti proměnné Y na proměnné X .

Vhodnost parametrů regresního modelu ověřujeme pomocí individuálních t-testů o nulové hodnotě regresních parametrů.

1) $H_0: \beta_0 = 0$ (tj. Regresní parametr β_0 není statisticky významný)

$$H_1: \text{non } H_0$$

2) $t = b_0/s_{b0}$

3) $W = (t; t \leq -2,306 \text{ u } t \geq 2,306)$

4) $t = 35,893 / 8,2494 = 4,351$

T je prvkem kritického oboru.

H_0 zamítáme, H_1 přijímáme

Regresní parabola

$$Y = b_0 + b_1x + b_2x^2$$

$$Y = 18,483 - 0,899x - 0,001x^2$$

Zvolená parabola není vhodná k vystižení závislosti proměnné y na proměnné x.

5) t = zamítáme H_0 , přijímáme H_1

Regresní parabola

1) H_0 :

2) $t = b_1 / sb_1$

3) $W = (t; t \leq -2,365 \text{ u } t \geq 0,365)$

4) $t = -1,530$

5) H_0 nezamítáme

Na hladině významnosti 5 % jsme nezamítli předpoklad o tom, že parametr β_1 je statisticky nevýznamný.

1) $H_0: \beta_2 = 0$

2) $T t = b_2 / sb_2$

3) W

4) $T t = -0,001/0,0341 = 0,029$

5) H_0 nezamítáme

Pokud je celkový F test významný, ale většina t-testů ne, není vhodné. Tuto parabolu by nebylo moc vhodné používat.

Regresní funkce (srovnání)

Nyní máme dvě funkce, které jsou vhodné k popisu

Kritérium	přímka	Hyperbola
Index determinace - I^2	0,914	0,703
Hodnota reziduálního součtu čtverců - S_R	8,937	30,894
S^2_R	1,117	3,862
F	85,090	18,930

Vhodnější z obou funkcí k popisu závislosti proměnné y na proměnné x je **přímka**.

CVIČENÍ 31. 3.

Test hypotézy o shodě 2 průměrů

- 2 nezávislé náhodné výběry z normálního rozdělení

- 1) a) $H_0: \mu_1 = \mu_1$
 b) $H_1: \mu_1 \neq \mu_1$
 c) $H_1: \mu_1 > \mu_1$
 d) $H_1: \mu_1 < \mu_1$

- 2) Test. Kritérium - 3 možnosti

známe

ano --> u

ne --> t

Příklad 1)

Určení rozdílů ve spotřebě PHM 2 typů vozů bylo náhodně vybráno 10 vozů s naměřenou spotřebou:

Insignia	6,2	7,3	6,3	5,5	6,8	6,5	6,3	6,6	7,1	6,5
Mondeo	5,7	5,0	5,3	5,6	6,1	5,3	5,8	5,7	5,4	5,5

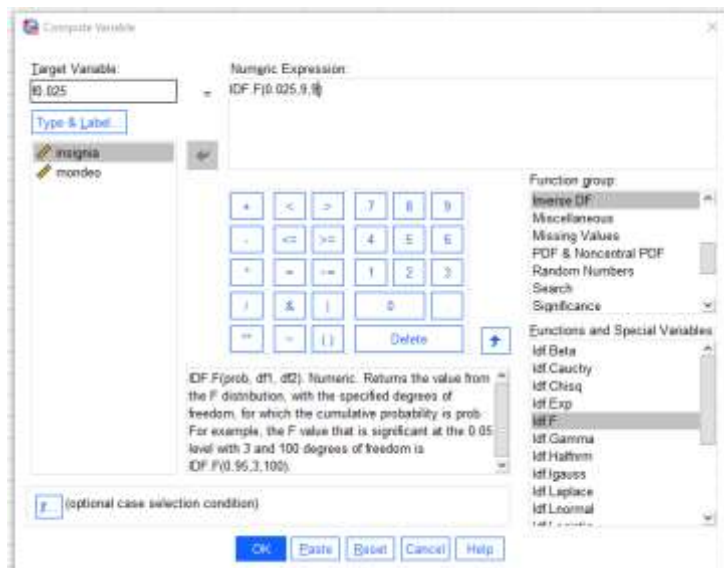
Řešení

- 1) $H_0: \mu_1 = \mu_2$ (střední hodnoty jsou si rovny)

$H_1: \mu_1 \neq \mu_2$ (statisticky významný - nerovnají se sobě)

Dále je potřeba vybrat vhodné testové kritérium.

- 2) $F = s'^2_1 / s'^2_2$



0.025 --> 0,248

0.075 --> 4,026

Statistics

		insignia	mondeo
N	Valid	10	10
	Missing	0	0
Std. Error of Mean		,15878	,09798
Std. Deviation		,50211	,30984
Variance		,252	,096
Range		1,80	1,10
Minimum		5,50	5,00
Maximum		7,30	6,10

Na hladině významnosti 5 % jsme nezamítli předpoklad o tom, že rozptyly v obou základních souborech jsou shodné.

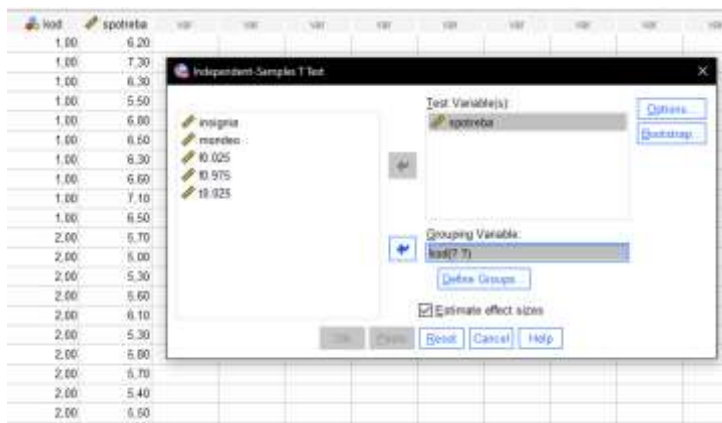
Test o shodě 2 průměrů:

2) $t =$

Následuje stanovení kritického oboru

3) $- 2,101 / 2,101 +$

- 4) $t = 5,199$
- 5) t náleží $W \rightarrow H_0$ zaítáme, H_1 nepřijímáme



Grouping Variable:

kod(1 2)

Define Groups...

Independent Samples Test									
Levene's Test for Equality of Variances				t-test for Equality of Means					
	F	Sig.	t	df	Significance: One-tailed p	Two-tailed p	Mean Difference	90% Confidence Interval of the Difference	
Equal variances assumed	.267	.607	5.199	18	<.001	<.001	.67900	.18698	.51692 1.36108
Equal variances not assumed			5.199	14.836	<.001	<.001	.67900	.18698	.57229 1.36771

Zadání:

Výběrové průměry $x_1 = 6,51$ $x_2 = 5,54$

Výběrové rozptyly $s_{12} = 0,2521$ $s_{22} = 0,096$

Dále $n = 10$ $\alpha = 0,05$

Na hladině významnosti 5 % rozhodněte, zda je průměrná spotřeba druhého typu automobilu významně nižší než u prvního typu. Předpokládáme normalitu rozdělení spotřeby benzínu.

1) $H_0: \mu_1 = \mu_2$

$H_1: \mu_1 > \mu_2$

T Test Computed from Summary Data

Sample 1

Number of cases: 10

Mean: 6.51

Standard Deviation: 0.502

Label: Sample 1

Sample 2

Number of cases: 10

Mean: 5.54

Standard Deviation: 0.310

Label: Sample 2

Confidence Level (%): 95

Marta Garcia-Granero provided valuable assistance with this procedure

This dialog requires the Python plugin

Note: using syntax you can do many sets of tests in one command

OK Paste Reset Cancel

Pomocný test Hartliho

Summary T-Test

Summary Data

	N	Mean	Std. Deviation	Std. Error Mean
Sample 1	10,000	6,510	,502	,159
Sample 2	10,000	5,540	,310	,098

Independent Samples Test

	Mean Difference	Std. Error Difference	t	df	Sig. (2-tailed)
Equal variances assumed	,970	,187	5,199	18,000	<,001
Equal variances not assumed	,970	,187	5,199	14,993	<,001

Hartley test for equal variance: F = 2.622, Sig. = 0.0721

95.0% Confidence Intervals for Difference

	Lower Limit	Upper Limit
Asymptotic (equal variance)	,604	1,336
Asymptotic (unequal variance)	,604	1,336
Exact (equal variance)	,578	1,362
Exact (unequal variance)	,572	1,368

SEMINÁŘ 4. 4. 2022

Příklad 1

Průměrná pevnost 88 jednotek, u 20 náhodně vybraných ...

Náhodný výběr	20
Výběrový průměr	85,1
Směrodatná odchylka	3,25
Na hladině významnosti	5 %

Testujte hypotézu, že průměrná pevnost proužků je 88 jednotek. Normalita rozdělení náhodné veličiny, kterou je pevnost proužků netkané textilie.

1) $H_0: \mu = 88$

$H_1: \mu > 88$

2) $t = -3,99$

3) $t_{0,025}(19) = -2,09$

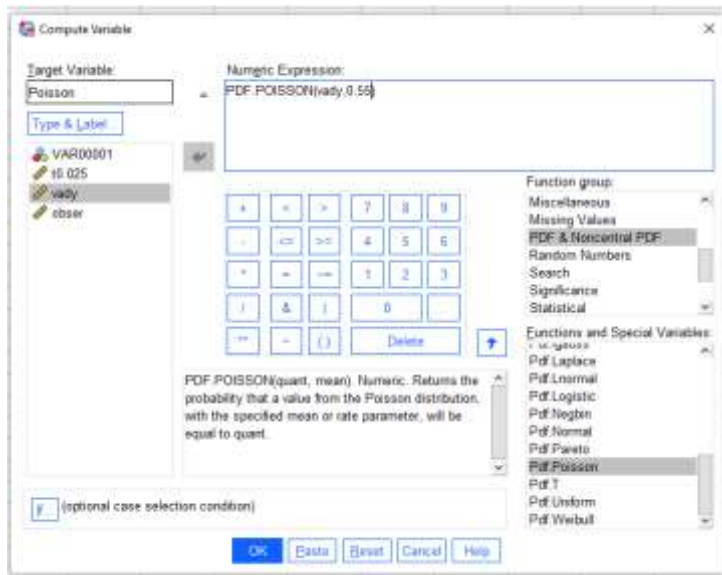
$t_{0,975}(19) = 2,09$

3) $w = t \leq -2,093$ nebo $t \geq 2,093$

Příklad 2

Počet vad	n
0	65
1	20
2	10
3+	5
Celkem	100

 vady	 obser
,00	65,00
1,00	20,00
2,00	10,00
3,00	5,00



vady	obser	Poisson
,00	65,00	,5769
1,00	20,00	,3173
2,00	10,00	,0873
3,00	5,00	,0160

G =

p2 = 0,01846

chi = 7,81

Příklad 3

count * y Crosstabulation

Count

		y					Total
		1,00	2,00	3,00	4,00	5,00	
count	,00	0	1	0	0	1	2
	1,00	2	1	2	2	2	9
	2,00	0	2	2	1	1	6
	3,00	2	0	0	1	0	3
Total		4	4	4	4	4	20

$$2 * 7/8 = 1,75$$

PŘEDNÁŠKA 6. 4.

Korelační analýza

- Zabývá se především intenzitou vzájemného vztahu proměnných
- Je základné metodou měření síly lineární závislosti číselných proměnných
- „correlatio“ – vzájemná souvislost
- Z výpočetních a interpretačních hledisek se regresní a korelační analýza prolínají, nelze mezi nimi stanovit ostrou hranici

Párově sdružené regresní koeficienty

1. $b_{yx} = b_{xy} = 0 \Rightarrow x$ a y jsou korelačně nezávislé; sdružené regresní přímky svírají pravý úhel.
2. $b_{yx} = \frac{1}{b_{xy}} \Rightarrow x$ a y jsou úplně závislé; sdružené regresní přímky svírají nulový úhel (splývají).

Míry těsnosti lineární závislosti

Koeficient determinace: $r_{yx}^2 = r_{xy}^2 = b_{yx} \cdot b_{xy} = \frac{s_{xy}}{s_x^2} \cdot \frac{s_{xy}}{s_y^2} = \frac{s_{xy}^2}{s_x^2 \cdot s_y^2}$; $r_{xy}^2 \in \langle 0; 1 \rangle$

Koeficient korelace: $r_{yx} = r_{xy} = \sqrt{r_{yx}^2} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2) \cdot (\overline{y^2} - \bar{y}^2)}} \cdot \sqrt{\overline{x^2} - \bar{x}^2} \cdot \sqrt{\overline{y^2} - \bar{y}^2}$; $r_{xy} \in \langle -1; 1 \rangle$

- měří sílu lineární závislosti, nikoli závislosti obecně (lineární závislost = korelovanost).
- koeficient je symetrický.

Jaký podíl závisle proměnné je možné objasnit pomocí dané regresní přímky.

b_{xy}	b_{yx}	r_{yx}
-	-	-
+	+	+
$\swarrow$ +	$\swarrow$ -	
$\searrow$ -	$\searrow$ +	

Test hypotézy o nulové hodnotě korelačního koeficientu

1) $H_0: \rho_{xy} = 0$ (lineární nezávislost x a y)

$H_1: \text{non } H_0$

Obvykle oboustranná

2) **Testové kritérium:**

$$t = \frac{r_{yx} \cdot \sqrt{n-2}}{\sqrt{1-r_{yx}^2}}; \quad \text{Statistika } t \text{ má při platnosti } H_0 \text{ rozdělení } t \text{ s } (n-2) \text{ stupni volnosti}$$

3) **Kritický obor:**

$$W = \left\{ t; t \leq t_{\frac{\alpha}{2}}(n-2) \cup t \geq t_{1-\frac{\alpha}{2}}(n-2) \right\}$$

4) **Závěr testu:**

Pokud leží hodnota testového kritéria v kritickém oboru, zamítáme H_0 a přijímáme H_1 , tzn. prokázali jsme hypotézu o lineární závislosti proměnných x a y .

Příklad 1:

Z výběrového souboru o rozsahu 11 jednotek byl vypočítán korelační koeficient $r_{yx} = 0,3$. Posuďte na hladině významnosti 5 %, zda je korelační koeficient v základním souboru statisticky významný.

Statistický nevýznamný $\rightarrow = 0$

$$t = \frac{r_{yx} \cdot \sqrt{n-2}}{\sqrt{1-r_{yx}^2}};$$

$$t = (0,3 * \sqrt{11-2}) / (\sqrt{1-0,3^2})$$

1) $H_0: \rho_{yx} = 0$

$H_1: \rho_{yx} \neq 0$

2) $t =$

3) $w = t \leq -2,262$ nebo $t \geq 2,262$

4) $t = 0,943$

Hodnota se nenachází na kritickém oboru $\rightarrow$ nezamítáme H_0

Na hladině významnosti 5 % nezamítáme předpoklad o tom, že korelační koeficient v základním souboru je statisticky nevýznamný.

Pokračování příkladu:

Co kdyby byl výběrový soubor 51 jednotek?

$n = 51$

$r_{yx} = 0,3$

$\alpha = 0,95$

$w = \geq 2,01$ nebo $\leq -2,01$

zamítáme H_0 , přijímáme H_1

Na hladině významnosti 5 % jsme prokázali, že korelační koeficient v základním souboru je statisticky významný.

Lineární závislost mezi proměnnými x a y je přímá a spíše slabší.

Příklad:

Dva degustátoři vedli ochutnávku českých vín. Po ochutnání všech 10 vzorků každý z nich sestavil žebříček vín podle toho, jak celkově vína zhodnotil (od nejlepšího po nejhorší). Posuďte na hladině významnosti 5 %, zda je mezi oběma hodnoceními závislost, příp. změřte sílu a směr.

Víno	A	B	C	D	E	F	G	H	I	J	Celkem
Degustátor X	3	6	8	5	1	4	10	9	2	7	X

Degustátor Y	2	5	9	6	1	4	8	10	3	7	X
$(i_k - i_y)^2$	1	1	1	1	0	0	4	1	1	0	10

$$n = 10$$

$$\alpha = 0,05$$

Vypočítáme Spearmanův koeficient:

$$R_s = 1 - 60 / 990 = 0,939$$

$$1) \quad H_0: \rho_s = 0$$

$$H_1: \rho_s \neq 0$$

$$2) \quad t =$$

$$3) \quad \text{testové kritérium} \quad = < -2,306 / >= 2,306$$

$$4) \quad t = 7,722$$

Časové řady

Chronologický průměr

a) prostý

b) vážený

CVIČENÍ 7. 4.

Parametrické testy

Neparametrické testy

- Chi-kvadrát test dobré shody
 - Pokud nulová hypotéza uvádí

Příklad – chí test dobré shody:

Výrobek je dodáván ve 3 provedeních – A, B, C. Výrobce očekává zájem zákazníků ve struktuře:

A – 46 %

B – 32 %

C – 22 %

Ověřte na významnosti 5 %, zda je předpoklad výrobce správný, a to na základě údajů o prodejkách za první měsíc, kdy bylo prodáno 600 ks výrobků ve struktuře, kterou uvádí následující tabulka.

Typ	Prodej v ks (n_i)	$\pi_{0,i}$	$n\pi_{0,i}$	$\frac{(n_i - n\pi_{0,i})^2}{n\pi_{0,i}}$	
A	280	0,46	276	0,06	
B	190	0,32	192	0,02	
C	130	0,22	132	0,03	
Celkem	600	1,00	600	0,11	

1) $H_0: \pi_1 = 0,46; \pi_2 = 0,32; \pi_3 = 0,22$

(nebo H_0 : předpoklad výrobce je správný)

H_1 : non H_0

Dále je potřeba vybrat vhodné testové kritérium.

2) G = vzorec

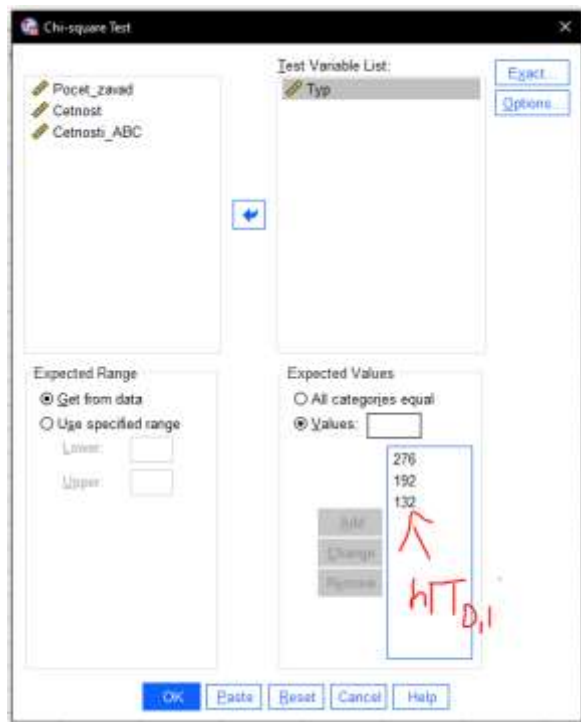
3) $W = G > 5,991$

4) $G = 0,11$

5) $G \neq W > H_0$ nezamítáme, H_1 nepřijímáme

Na hladině významnosti 5 % nezamítáme předpoklad o tom, že výrobce odhadl strukturu zájmu zákazníků o různé typy výrobků správně.

Weight cases – Cetnost_ABC



Typ

	Observed N	Expected N	Residual
1	280	276,0	4,0
2	190	192,0	-2,0
3	130	132,0	-2,0
Total	600		

Test Statistics

	Typ
Chi-Square	,109 ^a
df	2
Asymp. Sig.	,947

a. 0 cells (0,0%) have expected frequencies less than 5. The minimum expected cell frequency is 132,0.

1) $H_0: \pi_1 = 0,46; \pi_2 = 0,32; \pi_3 = 0,22$

H_1 : non H_0

2) Výpočet $n\pi_{0,i}$

3) $G = 0,109$ (0,11)

4) P-Value (Sig) = 0,947 > $\alpha \Rightarrow$ nezamítáme H_0 , nepřijímáme H_1

Pracovníci zpracovávají přehledy. Každý list nese záznam o 200 poukázkách. Výkon pracovníků se poměřuje počtem závad na pracovních listech. Pro normování práce je nutné získat představu o rozdělení listů podle počtu závad. Máme k dispozici údaje o 500 listech. Rozhodněte na hladině významnosti 1 %, zda by se v této situaci vyhovovalo Poissonovo rozdělení se střední hodnotou $\lambda = 4$.

Počet závad	Počet listů	Poisson - pravdě funkce	$n * \pi_{0,i}$	$\frac{(n_i - n\pi_{0,i})^2}{n\pi_{0,i}}$
0	11	0,0183	9,15	0,3740
1	37	0,0733	36,65	0,0000
2	65	0,1465	73,25	0,93
3	100	0,1954	97,7	
4	103	0,1954	97,7	
5	80	0,1563	75,15	
6	49	0,1042	52,1	
7	33	0,0595	29,75	
8	10	0,0298	14,9	
9	8	0,0132	6,6	
10	3	0,0053	2,65	
11 a více	1	0,0028	1,4	
Celkem	500	1,0000	500,00	

1) H_0 : počet závad se řídí Po(4)

H_1 : non H_0

2) Výpočet $\pi_{0,i}$ a $n\pi_{0,i}$

G = vzorec

Následuje stanovení kritického oboru.

3) $W = G > 24,725$

Okomentoval(a): [JH3]: s parametrem 4

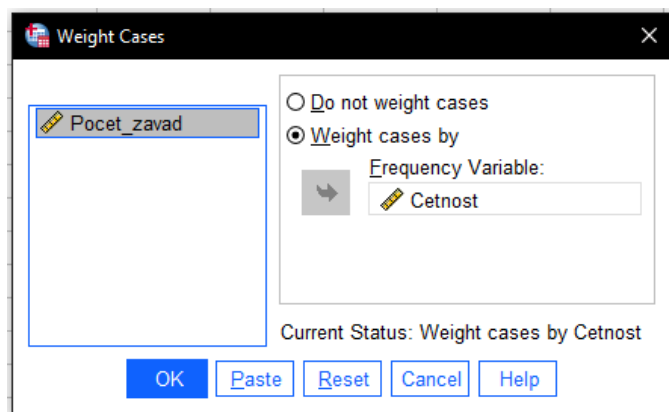


4) $G = 4,29$

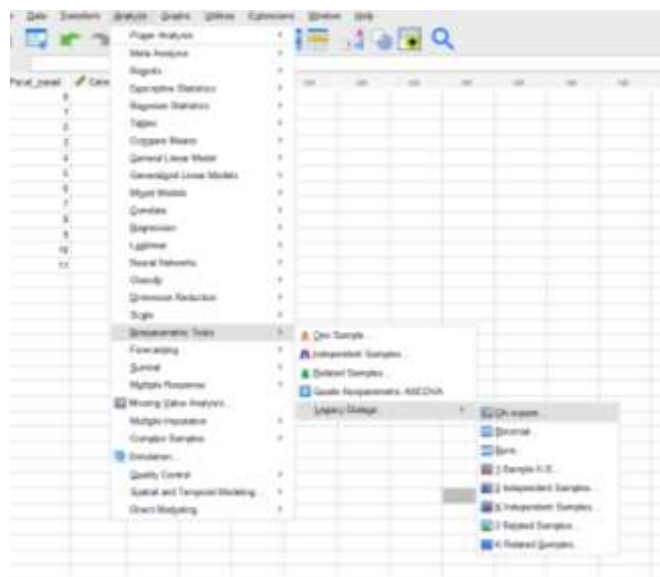
5) $G \neq W > H_0$ nezamítáme. H_1 nepřijímáme

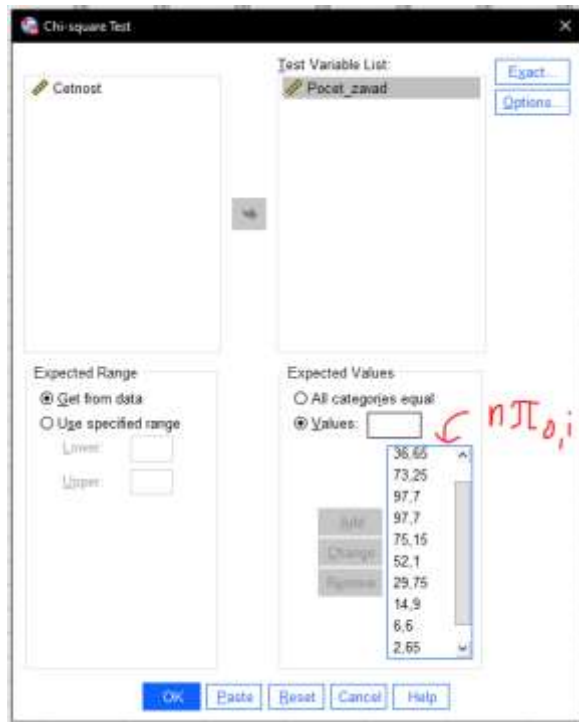
Na hladině významnosti 1 % nezamítáme předpoklad o tom, že se počet závad na zpracovaných listech řídí Poissonovým rozdělením se střední hodnotou $\lambda = 4$.

Postup



--> **Weight On**





Test Statistics

	Pocet_zavad
Chi-Square	4,524 ^a
df	11
Asymp. Sig.	,952

a. 2 cells (16,7%) have expected frequencies less than 5. The minimum expected cell frequency is 1,4.

- 5) P-Value (Sig) = 0,952 > α (= 0,01) => nezamítáme H0, nepřijímáme H1

SEMINÁŘ 11. 4.

Zadání příkladu

Sledujeme výzkum, který se zabývá vlivem 4 typů učebních metod angličtiny na výsledky studentů, posuzované podle výsledků dosažených v závěrečném testu (v bodech). Údaje jsou v následující tabulce. Data byla získána náhodným výběrem, rozsah souboru je 20 studentů. Předpokládáme, že výkon studentů má normální rozdělení. Pomocí analýzy rozptylu posuďte závislost proměnné y (výsledky v bodech) na metodě výuky (x). Posuďte sílu případné závislosti.

Typ učební metody (x_i) $i = 1,2,3,4$	Počet bodů v závěrečném testu (y_{ij}) $j = 1,2,3,4,5$	Četnost (n_i)	Podmíněné průměry $\bar{y}_i = \frac{\sum y_{ij}}{n_i}$	Podmíněné rozptyly $s_i^2 = \frac{\sum (y_{ij} - \bar{y}_i)^2}{n_i}$
1. typ	71 83 65 77 84	5	76	52
2. typ	60 51 54 80 55	5	60	108,4
3. typ	55 55 62 65 63	5	60	17,6
4. typ	68 73 67 59 53	5	64	50,4
Suma		20	X	X

- 1) $H_0: \mu = \mu = \mu = \mu$
- 2) Testové kritérium

1) Analyze>Compare Means>One-Way ANOVA

Dependent List: zisk
Factor: metoda

ANOVA

zisk

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	862,000	3	286,667	4,313	,026
Within Groups	3142,000	16	196,375		
Total	4004,000	19			

- 3) Kritický obor

$$W = (F, F \geq F_{1-\alpha}(k-1; n-k))$$

$$W = (F, F \geq F_{0,95}(3; 16))$$

$$W = (F, F \geq 3,2391)$$



 f0.95
3,2389

4) 4,016 >= 3,2391

H0 nezamítáme, H1 nepřijímáme

Sílu závislosti

$$P^2 = \frac{S_{ym}}{S_y} = \frac{860}{2002} = 42,96\%$$

ANOVA

zisk

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	860,000	3	286,667	4,016	,026
Within Groups	1142,000	16	71,375		
Total	2002,000	19			

Zadání příkladu

Na základě průzkumu provedeného u čtenářů časopisů A, B, C byla sestavena kontingenční tabulka. Rozhodněte, zda výběr časopisu a bydliště čtenáře jsou znaky závislé, neboli zda se struktura čtenářů z hlediska bydliště u časopisů A, B, C liší.

	Časopis			
Bydliště	A	B	C	Celkem
Liberec	75	75	50	200
Jablonec	40	70	40	150
Česká Lípa	35	5	10	50
Celkem	150	150	100	400

bydliste * casopis Crosstabulation

			casopis			Total
			a	b	c	
bydliste	CSL	Count	35	5	10	50
		Expected Count	18,8	18,8	12,5	50,0
	JBC	Count	40	70	40	150
		Expected Count	56,3	56,3	37,5	150,0
	LBC	Count	75	75	50	200
		Expected Count	75,0	75,0	50,0	200,0
Total	Count		150	150	100	400
	Expected Count		150,0	150,0	100,0	400,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	32,889 ^a	4	<,001
Likelihood Ratio	34,516	4	<,001
N of Valid Cases	400		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 12,50.

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	,287	<,001
	Cramer's V	,203	<,001
	Contingency Coefficient	,276	<,001
N of Valid Cases		400	

Testové kritérium G

$$G = \sum_{i=1}^r \sum_{j=1}^s \frac{(n_{ij} - n_{ij}')^2}{n_{ij}'}$$

$$3. \quad W \equiv \left\{ G, G > \chi^2_{1-\alpha} [(r-1)(s-1)] \right\}$$

$$W \equiv (G, G \geq \chi^2(4))$$

$$W \equiv (G, G \geq 9,488)$$

$$4. \quad 32,889 \geq 9,488$$

Testové kritérium je prvkem kritického oboru. Nulovou hypotézu zamítáme, přijímáme hypotézu alternativní.

PŘEDNÁŠKA 13. 4. 2022

A. Rozdělení ČR podle časového hlediska rozhodného pro zjišťování údajů

Zadání příkladu:

V následující tabulce jsou o tržbách v maloobchodě u fyzických a právnických osob v ČR v letech 2002-2006 v mil. Kč. Charakterizujte úroveň hodnot této časové řady pomocí vhodné charakteristiky.

Rok	2002	2003	2004	2005	2006
Tržby v mil. Kč	621 870	667 196	683 815	714 914	750 713

Aritmetický průměr

$$y = 3\,438\,508 / 5 = 687\,701,6$$

U let – zvážit zda 1. a poslední rok nepůlit → Průměrný počet obyvatel ve sledovaném období.

Zadání příkladu:

Datum	Počet zaměstnanců	$\bar{y}_i$	d_i	$y_i d_i$
1. 1. 2019	190	$190 + 230 / 2 = 210$	151	31 710
1. 6. 2019	230	$230 + 280 / 2 = 255$	61	15 555
1. 8. 2019	280	$280 + 250 / 2 = 265$	122	32 330
1. 12. 2019	250			
Celkem	x	x	334	79 595

$$\frac{\sum_{i=1}^{n-1} \bar{y}_i d_i}{\sum_{i=1}^{n-1} d_i} = 79\,595 / 334 = 238,3$$

Průměrný počet zaměstnanců byl od 1.1. do 1.12. 238,2 osob.

→ okamžikový interval

→ vážený průměr

$$\frac{\sum_{i=1}^{n-1} \bar{y}_i d_i}{\sum_{i=1}^{n-1} d_i}$$

Absolutní difference (přírůstky) – charakterizují absolutní změnu (přírůstek nebo úbytek) hodnoty ukazatele v časovém okamžiku t oproti období předcházejícímu (t-1)

Zadání příkladu:

Máme údaje o produkci proutěných košů ve firmě v letech 2014-2018. Charakterizujte vývoj produkce pomocí jednoduchých charakteristik časových řad – vypočítejte absolutní přírůstky, koeficienty růstu, průměrný absolutní přírůstek a průměrný koeficient růstu. Vypočítané hodnoty interpretnete.

Rok	Počet vyrobených košů	$\Delta_t^{(1)}$	$\Delta_t^{(2)}$	k_t
2014	10 120	-	-	-
2015	12 460	2 340	-	1,231
2016	9 820	- 2 340	- 4 980	0,788
2017	9 970	150	2 790	1,015
2018	11 850	1 880	1 730	1,189

V roce 2015 vzrostla produkce proutěných košů oproti roku o 2 340 Ks.

Koeficient růstu:

$$k_t = y_2 / y_1 = 12\,460 / 10\,120 = 1,231$$

V roce 2015 vzrostla produkce proutěných košů oproti roku 2014 o 23,1 % (NEBO 1,231 krát).

$$k_3 = y_3 / y_2 = 9\,820 / 12\,460 = 0,788$$

V roce 206 poklesla produkce proutěných košů oproti roku 2015 o 21,2 %.

$$\Delta = 2\,340 + (-2\,640) + 150 + 1\,880 / 4 = 432,5$$

$$\Delta = \frac{Y_n - Y_1}{n - 1} = 11\,850 - 10\,120 / 4 = 432,5$$

V letech 2014-2018 se počet vyrobených košů v průměru zvýšilo 432,5 ks.

Průměrný koeficient růstu vypočítáme podle:

$$k = \sqrt[4]{1,040}$$

Počet vyrobených košů se průměrně zvýšil o 4 %.

CVIČENÍ 14. 4.

Gambleři mají tušení, že jedna z hracích kostek je cinklá.

Počet pokusů	n_i	$n\pi_{0,i}$	
1	13	10	$(13-10)^2 / 10 = 0,9$
2	10	10	$(10-10)^2 / 10 = 0,0$
3	9	10	0,1
4	11	10	0,1
5	11	10	0,1
6	6	10	1,6
Σ	60	60	2,8

--> proporce (jaká je pravděpodobnost ?

Na hladině významnosti 1 % je cinklá?

$$n = 60 \quad k = 6 \quad \alpha = 0,01$$

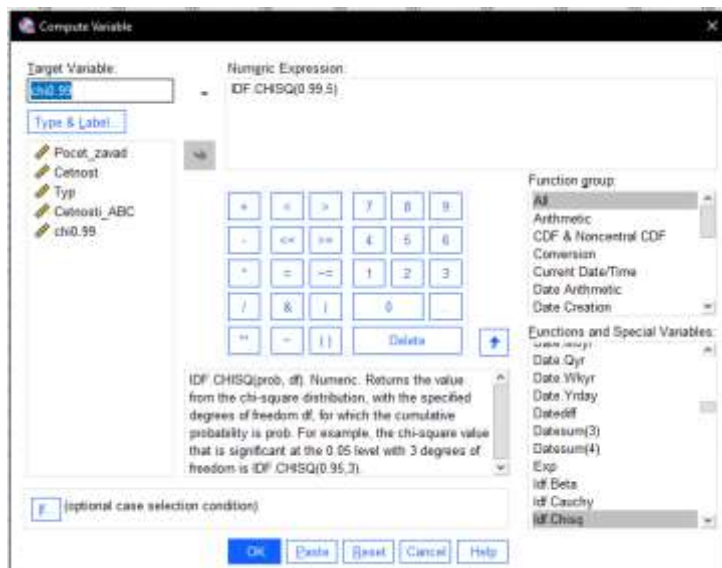
1) H_0 : kostka není cinklá ($\pi_n = 1/6$)

H_1 : kostka je cinklá

$$2) \quad G = \sum \frac{(n_i - n\pi_{0,i})^2}{n\pi_{0,i}}$$

$$3) \quad W \equiv \{G; G > X_{1-\alpha}^2(k-1)\}$$

$$W = (G; G > 15,086)$$

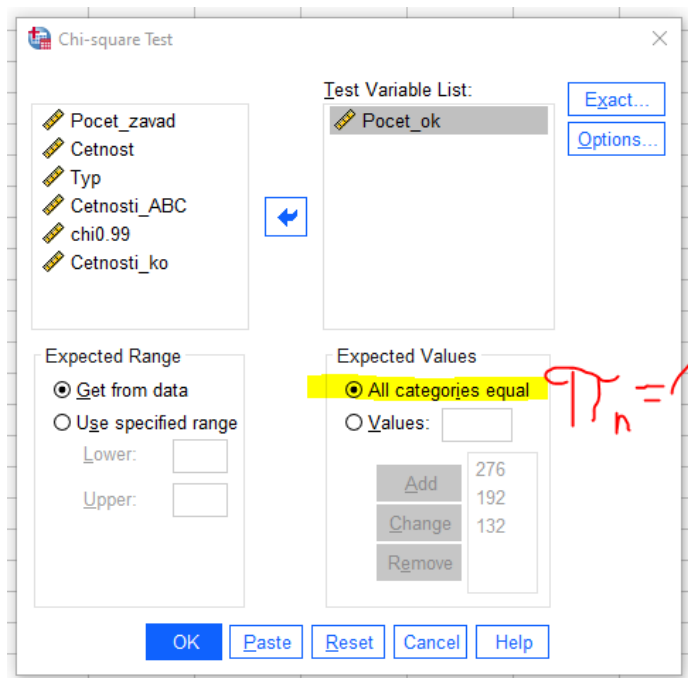


4) $G = 2,8$

5) $G \in V, G \neq W \Rightarrow \text{nezamítáme } H_0$

Na HV 1 % nezamítáme, že kostka není cinklá

-- Weight cases



Pocet_ok

	Observed N	Expected N	Residual
1	13	10,0	3,0
2	10	10,0	,0
3	9	10,0	-1,0
4	11	10,0	1,0
5	11	10,0	1,0
6	6	10,0	-4,0
Total	60		

Test Statistics

Pocet_ok	
Chi-Square	2,800 ^a
df	5
Asymp. Sig.	,731

a. 0 cells (0,0%) have expected frequencies less than 5. The minimum expected cell frequency is 10,0.

G = 2,8 - Chí

P-Value (Sig) = 0,731 > $\alpha 0,01$ --> nezamítáme H0, nepřijímáme H1

PŘEDNÁŠKA 20. 4. 2022

Základní koncepty modelování časových řad

2 přístupy k analýze

- Klasický
 - Model jednorozměrný
 - Nejjednodušší koncepty modelování ČŘ
 - Reálná hodnota ČŘ je funkce času (t)
 - Přístupy
 - Dekompozice
 - Boxova-Jenkinsonova metodologie
 - Spektrální analýza ČŘ
 - Model vícerozměrný
- Odeptovní

Složky časové řady

- Trendová složka (trend) = dlouhodobá tendence ve vývoji hodnot analyzovaného ukazatele v čase
- Sezónní složka = pravidelně se opakující odchylka od trendu, periodičita maximálně 1 rok
- Cyklická složka =
- Náhodná složka =

Typy modelů časových řad

Podle způsobu rozkladu rozlišujeme dva základní typy modelu:

Aditivní model: $y_t = T_t + S_t + C_t + \varepsilon_t$

Multiplikativní model: $y_t = T_t * S_t * C_t * \varepsilon_t$

Příklad:

Máme k dispozici časovou řadu domkončených bytů v ČR v letech 1995-2002. Vyrovnejte tyto data přímkou

Rok	Počet bytů (y_t)	t	t^2	$t*y_t$
1995	12 998	1	1	12 998
1996	14 037	2	4	28 074
1997	16 757	3	9	50 271
1998	22 163	4	16	88 652
1999	23 734	5	25	118 670
2000	25 207	6	36	
2001	24 758	7	49	
2002	27 700	8	64	
	167 354	36	204	

$$a = 8 * 844813 - 36 * 167354 / 8 * 204 - 36^2 = 733760 / 336 = 2183,81$$

$$a_0 = 167354 / 8 - 2183,81 * 36 / 8 = 11 092,11$$

Trendová přímka

Extrapolace pro rok 2003:

$$T = 30746,4$$

Extrapolace pro rok 2004:

$$T = 32930,2$$

Míry úspěšnosti zvolené trendové funkce

Trendová přímka (vhodnost)

Časová řada počtu dokončených bytů v ČR v letech 1995 – 2002 byla vyrovnána pomocí trendové přímky $T = 11092,11 + 2183,81t$. Ověřte na hladině významnosti 5 %, zda je přímka vhodným modelem pro vyrovnání těchto hodnot a ověřte, zda by třeba nebylo vhodnější použít namísto přímky trendovou parabolu.

Vhodnost celého modelu trendové přímky ověřujeme pomocí celkového F-testu.

- 1) $H_0: \alpha_0 = c; \alpha_1 = 0$ (přímka není vhodná k vyrovnání hodnot součtu dokončených bytů v ČR)

$H_1: \text{non } H_0$

2)
$$F = \frac{\frac{S_T}{p-1}}{\frac{S_R}{n-p}}$$

- 3) $W = (F; F > F_{0,95}(1,6))$

$W = (F; F > 5,987)$

Na hladině významnosti 5 % jsme prokázali, že trendová přímka je vhodná k vyrovnání dat o počtu dokončených bytů v ČR v letech 1995 – 2002.

Vhodnost **parametrů** trendové přímky ověřujeme pomocí individuálních t-testů.

- 1) $H_0: \alpha_0 = 0$ (parametr α_0 není statisticky významný)

$H_1: \text{non } H_0$

- 2) $t = a_0 / s_{a0}$

- 3) $W = (t; t \leq t_{0,025}(6) \text{ nebo } t \geq t_{0,975}(6))$

$W = (t; t \leq -2,447 \text{ nebo } t > 2,447)$

H_0 zamítáme, H_1 přijímáme

Na hladině významnosti 5 % jsme prokázali, že parametr α_0 je statisticky významný.

Dále budeme testovat významnost parametru α_1 :

- 1) $H_0: \alpha_1 = 0$

$H_1: \text{non } H_0$

- 2) $t = a_1 / s_{a1}$

- 3) $W = (t; t \leq t_{0,025}(6) \text{ nebo } t \geq t_{0,975}(6))$

$W = (t; t \leq -2,447 \text{ nebo } t \geq 2,447)$

- 2) $t = a_1 / s_{a1} = 8,747$

- 5) H_0 zamítáme, H_1 přijímáme

Vzhledem k tomu, že celkový F-test je významný a oba t-testy jsou ...

Trendová parabola

Nyní data o počtu dokončených bytů v ČR v letech 1995-2002 vyrovnáváme pomocí trendové paraboly.

Rovnice výběrové trendové paraboly má tvar:

$$T = a_0 + a_1t + a_2t^2$$

$$T = 8262,11 + 3881,81t - 188,667t^2$$

1) H_0

????

Parabola je vhodná k vyrovnaní dat.

Vhodnost parametrů trendové paraboly ověřujeme pomocí individuálních t-testů.

- 1) $H_0: \alpha_0 = 0$ (parametr α_0 není statisticky významný)
 $H_1: \text{non } H_0$
- 2) $t = a_0 / sa_0$
- 3) $W = (t; t \leq -2,571 \text{ nebo } t \geq 2,571)$ kvantily t-testu
- 4) $t = 8262,11 / 1945,81 = 4,246$
- 5) $t \Rightarrow H_0$ zamítáme, H_1 přijímáme

Trendová parabola

Dále budeme testovat významnost parametru.

- 1) $H_0: \alpha_0 = 0$ (parametr α_1 není statisticky významný)
 $H_1: \text{non } H_0$
- 2) $t = a_1 / sa_1$
- 3) $W = (t; t \leq -2,571 \text{ nebo } t \geq 2,571)$ kvantily t-testu $t_{0,975}(5)$
- 4) $t = 3881,81 / 992,058 = 3,913$
- 5) $t \Rightarrow H_0$ zamítáme, H_1 přijímáme

- 1) $H_0: \alpha_0 = 0$ (parametr α_1 není statisticky významný)
 $H_1: \text{non } H_0$
- 2) $t = a_2 / sa_2$
- 3) $W = (t; t \leq -2,571 \text{ nebo } t \geq 2,571)$ kvantily t-testu $t_{0,975}(5)$
- 4) $t = -188 = ????$
- 5) $t \Rightarrow H_0$ nezamítáme, H_1 nepřijímáme

Vzhledem k tomu, že celkový F-test je významný a většina t-testů také, lze trendovou parabolu považovat za vhodný model pro vyrovnání sledovaných dat.

Trendové funkce - srovnání

Kritérium	Přímka	Parabola
I^2	-	-
I^2_{adj}	0,915	0,937
S_R	-	-
S^2_R	2617660	1945200
F	76,52	53,02
M.S.E.	-	-

Vybrala bych si přímku – dobře konstruované, dobré výsledky, mnohem jednodušší funkce, nevýznamnost 3. parametru paraboly ...

CVIČENÍ 21. 4.

Analýza rozptylu

Metody zkoumání závislosti

- Analýza rozptylu
- Test nezávislosti kategoriálních znaků – slovní, nominální
- Regresní analýza – číselné
- Korelační – číselné

Hlavní myšlenka analýzy rozptylu

- 2. vnitroskupinový (reziduální) rozptyl s^2_{yv}

Test hypotézy o nezávislosti

Poměr determinace P^2 :

Poměr korelace P:

Zadání příkladu:

Proběhl výzkum zaměřený na zkoumání vlivu učební metody angličtiny na výsledky studentů. Výsledky se bodově hodnotily v závěrečných testech. Posuďte pomocí analýzy rozptylu, na hladině významnosti 5 %, zda výsledky studentů závisejí na učební metodě angličtiny. Případně změřte sílu dané závislosti. Předpokládáme normalitu rozdělení bodových výsledků.

Typ učební metody (x_i)	Počet bodů v testu (y_{ij})	Počet studentů (n_i)	y_i	s^2_i
Typ A	71, 83, 65, 77, 84	5	76	52
Typ B	60, 51, 54, 80, 55	5	60	108,4
Typ C	55, 55, 62, 65, 63	5	60	17,6
Typ D	68, 73, 67, 59, 53	5	64	50,4

Součet		20	x	x
--------	--	----	---	---

Analýza rozptylu

- 1) Typ proměnných je vyhovující. Zejména u závisle proměnné y to hraje podstatnou roli.
- 2) Výběr pochází z normálního rozdělení.
- 3) Výběry jsou nezávislé.
- 4) Počet pozorování je větší než počet skupin, tj $n > k$; zde $20 > 4$
- 5) Homoskedasticita - ověření pomocí Bartlettova testu

Bartlettův test

- 1) $H_0: \sigma^2_1 = \sigma^2_2 = \sigma^2_3$

H_1 : non H_0

- 2) B (Bartlettovo testové kritérium)

- 3) 7,815



- 4)

Analýza rozptylu

Test o nezávislosti sledované proměnné y na faktoru x:

- 1) $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_1 : non H_0

- 2)
$$\frac{S_{ym}}{k-1} / \frac{S_{yp}}{n-k}$$

3) $w = (F ; F \geq 3,239)$

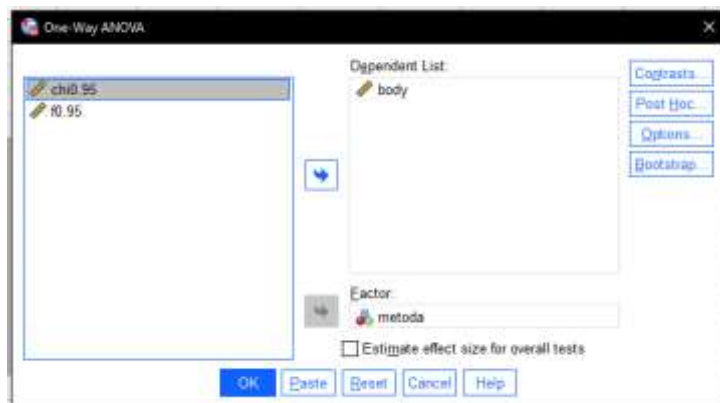
4) $F = \frac{\frac{860}{3}}{\frac{1142}{16}} = 4,02$

5) H_0 zamítáme, H_1 přijímáme

$P_2 = \frac{S_{ym}}{S_y} = \frac{860}{2002} = 0,430$

Závislost počtu bodů v testu je na učební metodě spíše nižší NEBO 43 % z celkové variability počtu bodů

IBM SPSS



One-Way ANOVA: Options

Statistics

☐ Descriptive

☐ Fixed and random effects

☒ Homogeneity of variance test

☐ Brown-Forsythe test

☐ Welch test

☒ Means plot

Missing Values

☒ Exclude cases analysis by analysis

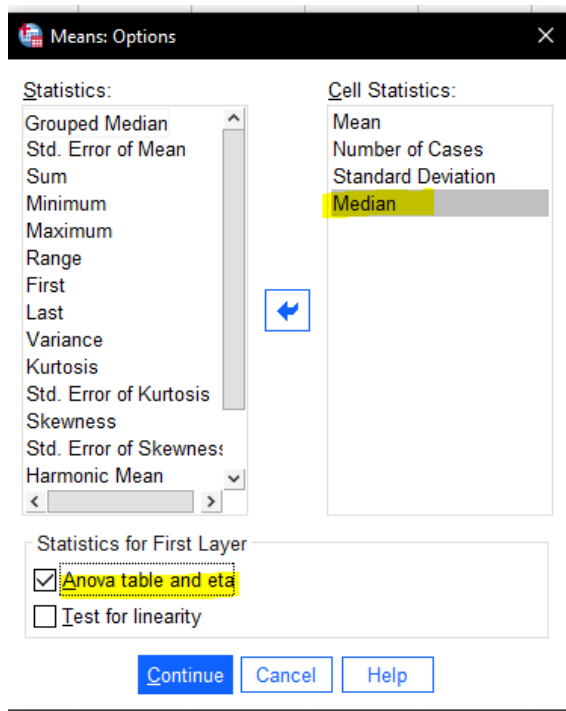
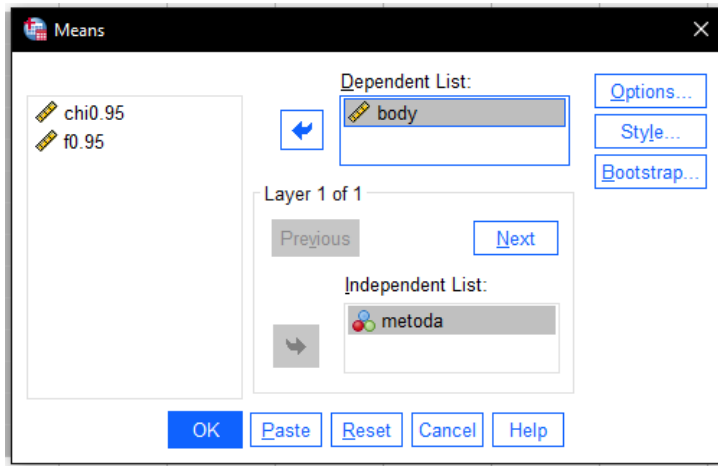
☐ Exclude cases listwise

Confidence Intervals

Level(%):

Tests of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
body	Based on Mean	,655	3	16	,592
	Based on Median	,254	3	16	,857
	Based on Median and with adjusted df	,254	3	8,811	,857
	Based on trimmed mean	,591	3	16	,630



Report

body

metoda	Mean	N	Std. Deviation	Median
1	76,00	5	8,062	77,00
2	60,00	5	11,640	55,00
3	60,00	5	4,690	62,00
4	64,00	5	7,937	67,00
Total	65,00	20	10,265	64,00

ANOVA Table

		Sum of Squares	df	Mean Square	F	Sig.
body * metoda	Between Groups (Combined)	860,000	3	286,667	4,016	,026
	Within Groups	1142,000	16	71,375		
	Total	2002,000	19			

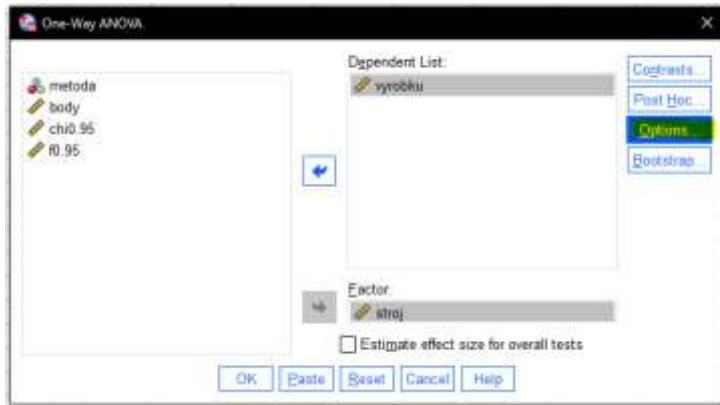
Zadání příkladu:

Výrobní podnik zamýšlí srovnat účinnost (efektivnost) 3 typů výrobních strojů. Každý stroj byl testován po 8 hodin a byly zaznamenány počty výrobků vyrobených během každé hodiny. Na hladině významnosti 5 % posuďte, zda počet vyrobených výrobků za 1 hodinu závisí na typu stroje. ...

Typ stroje	Počet vyrobených výrobků	Počet h
Stroj 1	65, 60, 63, 68, 70, 69, 65, 65	8
Stroj 2	59, 58, 63, 62, 57, 58, 60, 67	8
Stroj 3	57, 64, 62, 62, 60, 61, 61, 64	8
Součet		24

Předpoklady

- I. Typ proměnných - y je metrická, x je nominální => OK
- II. Výběry jsou nezávislé
- III. Výběry pocházejí z normálního rozdělení
- IV. $n > k$; $24 > 3$ => OK
- V. Homoskedasticita



Tests of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
vyrobku	Based on Mean	,823	2	21	,453
	Based on Median	,444	2	21	,647
	Based on Median and with adjusted df	,444	2	18,868	,648
	Based on trimmed mean	,780	2	21	,471

ANOVA

vyrobku

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	120,250	2	60,125	6,654	,006
Within Groups	189,750	21	9,036		
Total	310,000	23			

1) $H_0: \sigma^2_1 = \sigma^2_2 = \sigma^2_3$

H_1 : non H_0

Nezamítáme předpoklad, že

2) $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

3) $F = 6,654$

4) $\text{Sig.} = 0,006 < \alpha \Rightarrow$ zamítáme H_0 , přijímáme H_1

SEMINÁŘ 25. 4. 2022

	vykon	procento
1	73	4,80
2	55	3,80
3	53	3,90
4	60	4,00
5	70	4,60
6	69	4,40
7	68	4,40
8	63	3,80
9	63	4,00
10	58	3,90

Regression >

Curve Estimation

Dependent(s):
procento

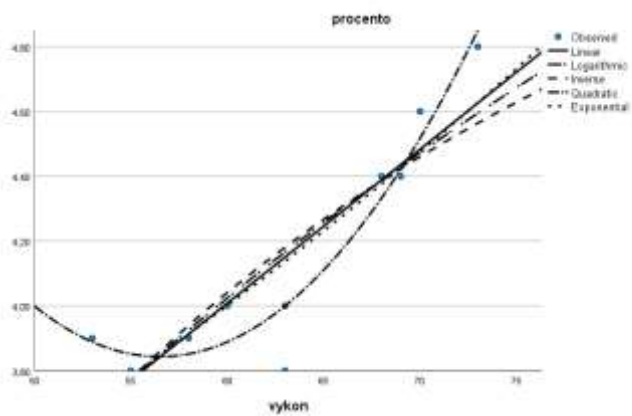
Independent:
☒ Variable: vykon
☐ Time

Case Labels:
☒ Include constant in equation
☒ Plot models

Models:
☒ Linear ☒ Quadratic ☐ Compound ☐ Growth
☒ Logarithmic ☐ Cubic ☐ S ☒ Exponential
☒ Inverse ☐ Power ☐ Logistic
 Upper bound:

☒ Display ANOVA table

OK Paste Reset Cancel Help



PŘEDNÁŠKA 27. 4.

Prosté klouzavé průměry

Předpoklad: na klouzavých částech o rozsahu $m = 2p + 1$ je definován lineární trend.

Vážené klouzavé průměry

Předpoklad: na klouzavých částech o rozsahu $m = 2p + 1$ je definován parabolický trend.

Zadání

t	yt	Tříčlenné klouzavé úhrny				
1	32					
2	38	110				
3	40	118				
4	40	121				
5	41	129				
6	48	145				
7	56	157				
8	53	165				
9	56	170				

Chyba! Chybný vložený objekt.

Máme k dispozici časovou řadu denních údajů za 3 týdny měsíce dubna 2019. V této časové řadě byla prokázána existence sezónnosti. Kvantifikujte velikost sezónních výkyvů a očistěte časovou řadu od sezónních vlivů.

Týden	Den	y _{it}	T _y	Sezonní rozdíly	Průměrné	Rozdílové	Očištěné údaje
15	PO	200	•	•			131,0714
	ÚT	140	•	•			134,6429
	ST	60	•	•			124,2857
	ČT	100	131,4286	-31,42857143	-34,2857	-35,3571	135,3571
	PÁ	230	130	100	95	93,92857	136,0714
	SO	110	127,1429	-17,14285714	-10,7143	-11,7857	121,7857
	NE	80	128,5714	-48,57142857	-55,7143	-56,7857	136,7857
16	PO	190	127,1429	62,85714286	70	68,92857	121,0714
	ÚT	120	125,7143	-5,714285714	6,428571	5,357143	114,6429
	ST	70	128,5714	-58,57142857	-63,2143	-64,2857	134,2857
	ČT	90	127,1429	-37,14285714			125,3571
	PÁ	220	130	90			126,0714
	SO	130	134,2857	-4,285714286			141,7857
	NE	70	132,8571	-62,85714286			126,7857
17	PO	210	132,8571	77,14285714			141,0714
	ÚT	150	131,4286	18,57142857			144,6429
	ST	60	127,8571	-67,85714286			124,2857
	ČT	90	130	-40			125,3571
	PÁ	210	•				116,0714
	SO	105	•				116,7857
	NE	85	•				141,7857
					7,5		
					1,071429		

Vhodné pro aditivní model

CVIČENÍ 28. 4.

Test nezávislosti

χ^2 – test nezávislosti lze použít tehdy, když....

Soubor 500 náhodných osob co vyletěli. Sledujeme 2 znaky – nejvyšší dokončené vzdělání a důvod výpovědi. Posuďte na hladině významnosti 5 %, zda je mezi nejvyšším dokončeným vzděláním a důvodem výpovědi nějaká závislost. Případně změřte její sílu pomocí vhodné charakteristiky.

Vzdělání	Důvod výpovědi					ní
	Druh práce	Finanční ohodnocení	Pracovní prostředí	Nepříjemný kolektiv	Ostatní	
ZŠ	31	55	30	8	6	130
U	27	34	22	5	2	90
SŠ	44	47	36	23	20	170
VŠ	50	25	19	10	6	110
n_j	152	161	107	46	34	500

Test o nezávislosti proměnných a a b.

1) H_0 : důvod výpovědi a nejvyšší dokončené vzdělání na sobě nezávisí

H_1 : non H_0

2) $G =$

3) $W = (G; G > \text{něco})$



	A	B	C	D	E	F	G
1	Vzdělání	Důvod výpovědi					ni
2		Druh práce	Finanční ohodnocení	Pracovní prostředí	Nepříjemný kolektiv	Ostatní	
3	ZŠ	31	55	30	8	6	130
4	U	27	34	22	5	2	90
5	SŠ	44	47	36	23	20	170
6	VŠ	50	25	19	10	6	110
7	nj	152	161	107	46	34	500
8							
9	Vzdělání	Důvod výpovědi					ni
10		Druh práce	Finanční ohodnocení	Pracovní prostředí	Nepříjemný kolektiv	Ostatní	
11	ZŠ	39,52	41,86	27,82	11,96	8,84	130
12	U	27,36	28,98	19,26	8,28	6,12	90
13	SŠ	51,68	54,74	36,38	15,64	11,56	170
14	VŠ	33,44	35,42	23,54	10,12	7,48	110
15	nj	152	161	107	46	34	500

$$C_c = \sqrt{g / N} \cdot H = \sqrt{37,994 / 500} \cdot 3 = 0,159$$

Třídění dat:

Výzkum 100 náhodných domácností:

28 má auto

20 má chatu

65 má hovno

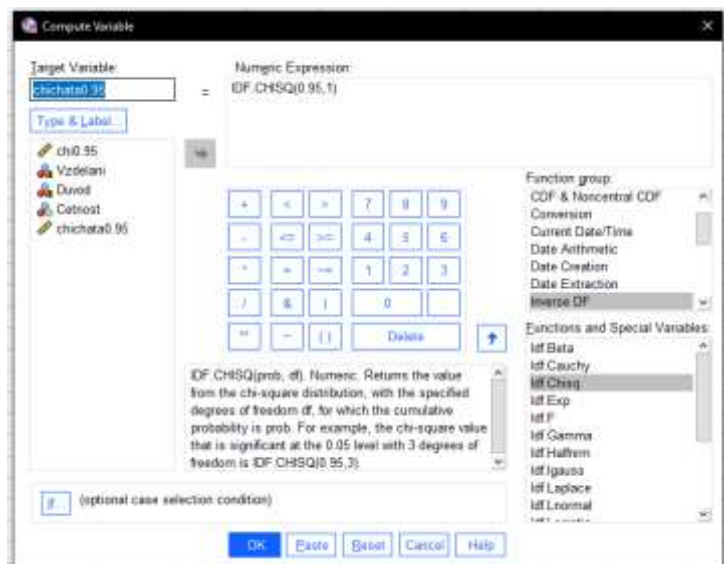
Vlastnictví chaty	Vlastnictví automobilu		ni
	ANO	NE	
ANO	13	7	20
NE	15	65	80
nj	28	72	100

1) H_0 : vlastnictví chaty a automobilu na sobě nezávisí

H_1 : non H_0

2) $G = n \cdot$

3) $W = (G; G >)$



$W = (G; G > 3,84)$

4) $G = 100 * (13*65 - 7*15)^2 / 20*80*28*72 = 16,977$

5) G náleží W --> zamítáme H_0 , přijímáme H_1

Test nezávislosti kategoriálních dat

Vhodnou charakteristikou pro změření síly

$RAB = 0,412$

Závislost mezi proměnnými je spíše slabší a přímá.

To znamená, že častěji pražské domácnosti vlastní automobil, tak i chatu, nebo nevlastní ani automobil, ani chatu.

Chata	Auto	Cestnost
4 ANO	ANO	13,00
4 ANO	NE	7,00
4 NE	ANO	15,00
4 NE	NE	65,00
4		

> WEIGHT CASES

Chata * Auto Crosstabulation

Count

		Auto		
		ANO	NE	Total
Chata	ANO	13	7	20
	NE	15	65	80
Total		28	72	100

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	16,977 ^a	1	<,001		
Continuity Correction ^b	14,760	1	<,001		
Likelihood Ratio	15,480	1	<,001		
Fisher's Exact Test				<,001	<,001
N of Valid Cases	100				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 5,60.

b. Computed only for a 2x2 table

G = 14,760 -->(Sig) = 0,000122 < α → kramerovo V

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	,412	<,001
	Cramer's V	,412	<,001
	Contingency Coefficient	,381	<,001
N of Valid Cases		100	

ABSENCE JAKO KRÁVA



Cvičení 12. 5.

Kritérium	Přímka	Parabola	Exponenciála
I^2	-	-	-
I^2_{adj}	0,599	0,956	0,599
F	11,449	76,41	11,442
S_R	-	-	-
$S^2_{R \text{ reziduální rozptyl}}$	0,477	0,053	0,011
Sigma	0,015	0,000179	0,015

Příklad: korelační analýza

Hladina významnosti - 5 %

x	1	2	2	3	3	4	5	5	7	8
y	6	5	7	4	5	3	3	4	2	1

Correlations

		x	y
x	Pearson Correlation	1	-,913**
	Sig. (2-tailed)		<,001
	N	10	9
y	Pearson Correlation	-,913**	1
	Sig. (2-tailed)	<,001	
	N	9	9

** . Correlation is significant at the 0.01 level (2-tailed).

(Bivariate)

Výběrový koeficient korelace: - 0,915

1) $H_0: \rho_{yx} = 0$

$H_1: \rho_{yx} \neq 0$

2) $\text{Sig} = 0,000204 < \alpha \rightarrow$ zamítáme H_0

$Y = 6,957 - 0,739 x$ přímka popisných závislostí y na x

$X = 8,533 - 1,133 y$ přímka popisných závislostí x na y

$b_{yx} = - 0,739$

$b_{xy} = -1,133$ párově regresní koeficienty

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	37,443	1	37,443	35,199	<,001
Residual	7,446	7	1,064		
Total	44,889	8			

The independent variable is y.

(Curve Estimation)

Vypočítejte hodnotu koeficientu determinace a interpretujte ji.

Když se hodnota nezávislé proměnné y zvýší o jednotku, potom se hodnota závislé proměnné x v průměru sníží o 1,133.

+ obráceně na 0,739

$$r_y^2 = 0,838$$

$$r_{yx}^2 = b_{yx} \cdot b_{xy}$$

$$r_{xy}^2 = -0,739 \cdot (-1,133) = 0,837$$

$$r_{yx} = \pm \sqrt{b_{yx} \cdot b_{xy}}$$

$$R_{yx} = -0,915$$

Pořadové korelace

Na výstavě potravinářského zboží soutěžilo 12 výrobků. Bylo stanoveno dvojí pořadí těchto výrobků – na základě hodnocení návštěvníků a poroty odborníků. Posuďte na hladině významnosti 5 %, zda je mezi oběma hodnoceními závislost, příp. Změřte její sílu a směr.

Výrobek	A	B	C	D	E	F	G	H	I	J	K	L
Návštěvníci	2	6	7	9	4	12	1	11	10	3	5	8
Odbor. porota	4	8	10	11	7	9	2	12	6	1	3	5

$$\alpha = 0,05$$

Correlations

		navsteva	porota
navsteva	Pearson Correlation	1	,741**
	Sig. (2-tailed)		,006
	N	12	12
porota	Pearson Correlation	,741**	1
	Sig. (2-tailed)	,006	
	N	12	12

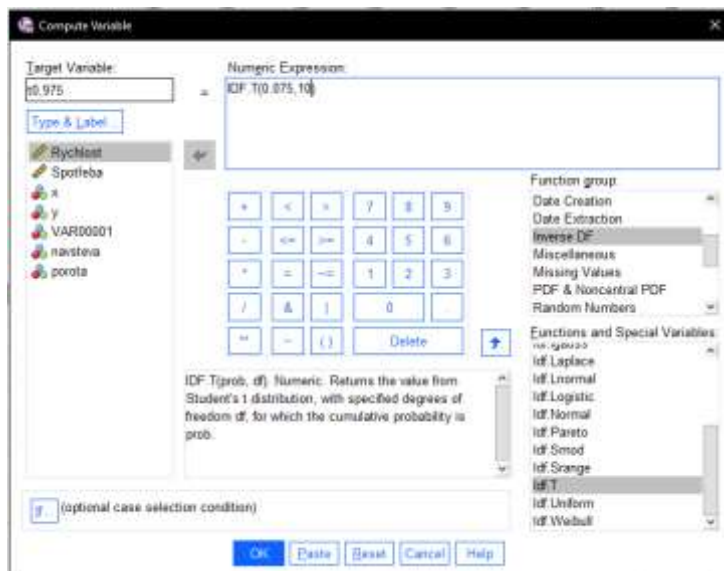
** Correlation is significant at the 0.01 level (2-tailed).

Vyjde tam sigma $0,006 < \alpha \rightarrow$ zamítáme.

1) $H_0: \rho_s = 0$

$H_0: \rho_s \neq 0$

2)



$t = -2,228$ a $2,228$

T je prvkem krizového oboru

H_0 zamítáme, H_1 přijímáme