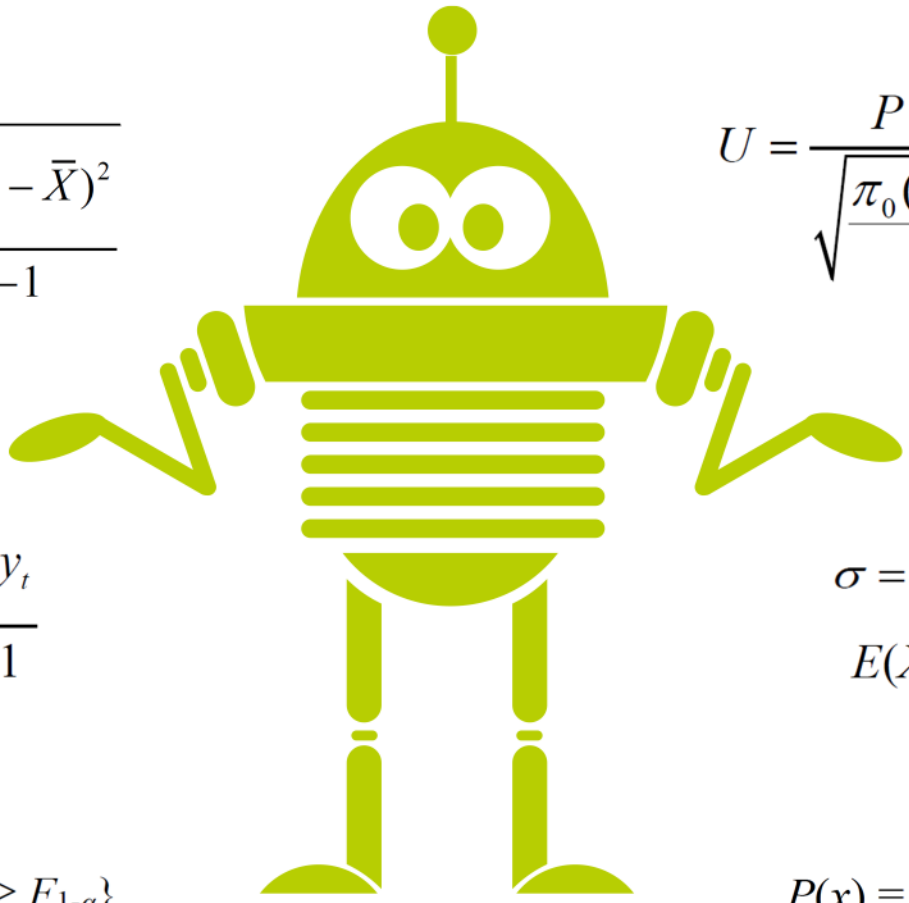
STATISTIKA

VŠTE V ČESKÝCH BUDĚJOVICÍCH

DRUHÁ ČÁST STUDUJNÍHO PLÁNU

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

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



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

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

$$E(X) = \pi$$

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

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

EDU FOR LIFE

VZDĚLÁNÍ, KTERÉ SE TI BUDE HODIT

Adriana Řeháčková

www.statistickyneklasicky.cz

STATISTIKA
VŠTE V ČESKÝCH BUDĚJOVICÍCH

Ukážeme si, že statistika není žádná nuda a že má opravdu smysl! Cílem kurzu je, abys pochopil základní až lehce pokročilé statistické metody. Nemine Tě ani osvojení Excelu a orientace ve vzorcích. Po kurzu bys měl vědět, co, kdy a jak použít a dokázat tyto znalosti využít na reálných datech.

OBSAH KURZU:
Dvouvýběrové testy
Analýza rozptylu - ANOVA
Testy o poměrech
Regresní přímka
Regresní analýza

O TOMTO MATERIÁLU:

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

STATISTICKY NEKLASICKY

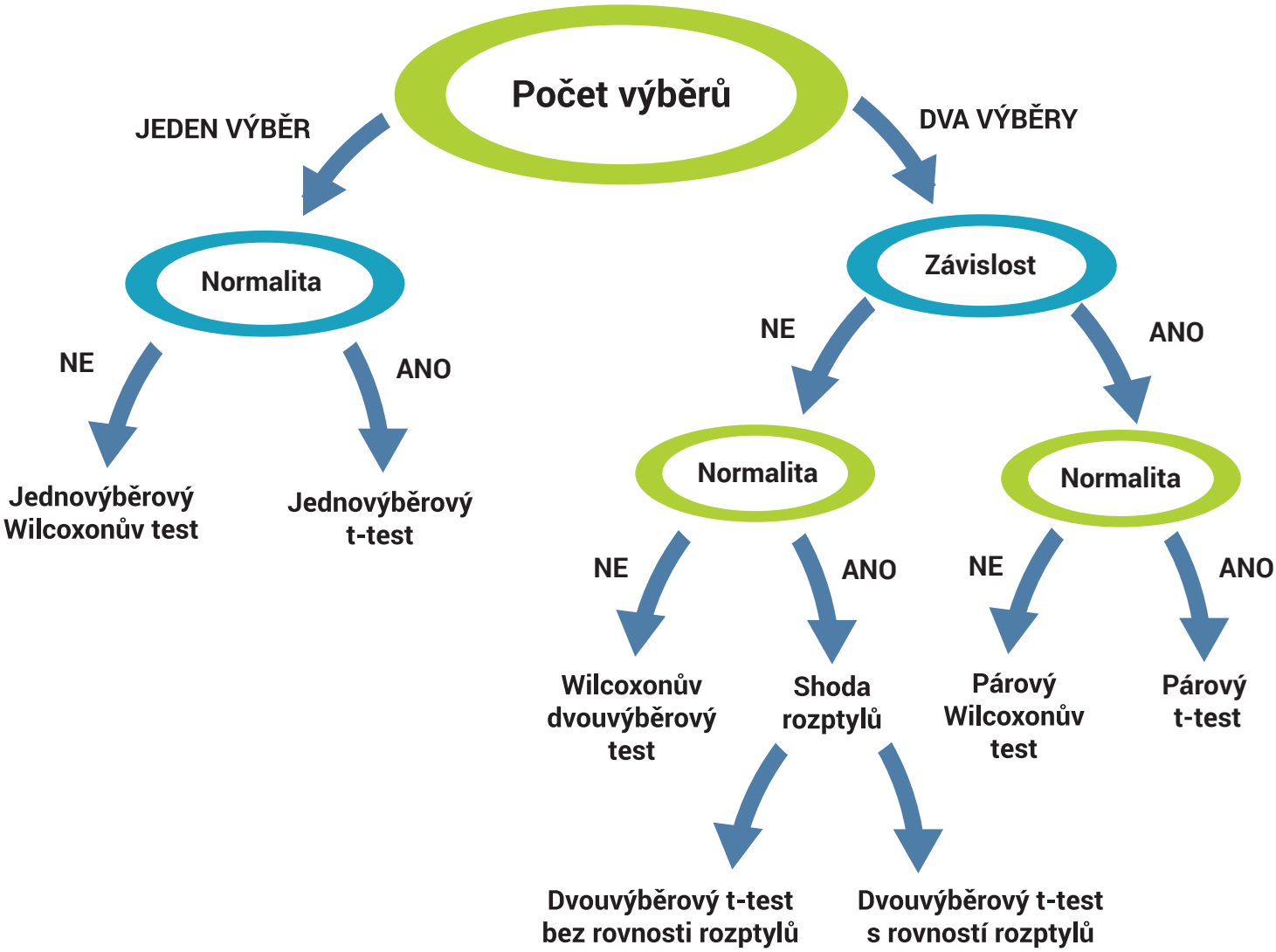
DVOUVÝBĚROVÉ TESTY

Máme testy pro nezávislé a závislé výběry. Je zde důležité uvědomit si základní rozdíl mezi těmito dvěma výběry.

testy pro nezávislé výběry - porovnáváme průměry dvou vzorků/skupin. Například: skóre mužů a žen, průměrný věk dožití v Japonsko a v ČR, hmotnost samců a samic, gympl vs učiliště,...

t-testy pro závislé výběry - porovnáváme dva průměry jednoho vzorku/skupiny. Většinou se používá pro porovnání dvou měření u stejných osob (máme jednu skupinu a na ní testujeme 2 odlišné věci). Například: známky z testů před a po přednášce, délka levé a pravé nohy, výška otce a syna (jsou provazané).

Následně rozhodneme podle normality, zda zvolit parametrický nebo neparametrický test.



1) Studenti psali v rámci předmětu statistika dva průběžné testy se stejný rozpětím možných získaných bodů. Údaje o počtech bodů získaných 10 studenty jsou v tabulce. Předpokládejme normální rozdělení počtu získaných bodů. Získali studenti v průměru stejné množství bodů v prvním a ve druhém testu? Testujte na 5% hladině významnosti.

Student	1	2	3	4	5	6	7	8	9	10
1. test	20	16	9	15	15	16	7	14	15	14
2. test	15	13	14	13	11	12	12	12	11	7

2) U 9 dvojčat narozených v místní nemocnici byla zjištěna následující porodní váha (v gramech):

starší	1340	2500	2600	1600	2250	1660	1750	2300	1900
mladší	1600	2003	2050	1 700	2450	2050	1500	2000	2120

Pomocí vhodného testu zjistěte, zda porodní váha staršího z dvojčat je vyšší než u mladšího z dvojčat.

3) Auta měly následující spotřebu PHM: 64, 67, 72, 60, 70. Jsou zkoušena nová vylepšení a zjišťováno, zda se spotřeba změnila. Nová spotřeba, po vylepšení byla 60, 76, 68, 70, 72. Testujte na 1% hladině významnosti, zda se spotřeba změnila či nikoliv.

4) 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).

skupina s placebem	skupina s test. lékem
95	75
85	60
100	30
120	65
80	100
90	70
85	40
80	55
75	65
120	110

5) Pojišťovna testuje nový typ pojištění a chce zjistit, zda se průměrná výše pojistného plnění liší mezi muži a ženami. Po prvních několika týdnech máme následující údaje o pojistném plnění. Otestujte na 5% hladině významnosti, zda se významně odlišuje průměrná výše pojistného plnění u mužů a žen.

Muži: počet: 21; směrodatná odchylka: 1 542; průměrná výše pojistného: 32 256 Kč

Ženy: počet: 27; směrodatná odchylka: 1 769; průměrná výše pojistného: 35 789 Kč

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-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%)?
Je 10% studentů s hodnocením výborně, 20% s chvalitebně, 50% s dobře a 20% s dostatečně?
Je hod kostkou spravedlivý?

1) Nechť v průzkumu bylo 40 žen a 30 mužů , je rozdělení žen a mužů rovnoměrné?

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

3) Hodíme 60krát šestistěnnou hrací kostkou. Jednotlivé stěny padly v následujícím poměru: 7:9:10:6:15:13. Proveďte test na 5% hladině významnosti, zda je kostka v pořádku.

DOMÁCÍ ÚKOL: Marketingový plán tvrdil, že záznam koncertu skupiny Kapička se prodá v poměru 70% CD, 20% DVD a 10% kazety. Za měsíc se skutečně prodá 2400 kusů CD, 1075 DVD a 384 kusů kazet tohoto koncertu. Ověřte, zda byl předpoklad marketingového plánu správný.

Řešení: $\{152,73 \geq 5,99\}$ Na 5% hladině významnosti zamítáme H0.
Předpoklad nebyl správný.

ANOVA -jednofaktorová analýza rozptylu

Analýzu rozptylu využíváme v případě, že máme víc druhů testovaného předmětu, a zároveň každý druh testujeme vícekrát. Sledujeme například plat ve skupině absolventů ZŠ, SŠ, VŠ a chceme prokázat, zda plat závisí na dosaženém vzdělání.

Zkrátka nás zajímá zda Y závisí na tzv. kategoriálním faktoru (na vzdělání, na typu stroje,…), odtud označení „jednofaktorová ANOVA“.

H0: Nezávislost na faktoru
H1: Závislost na faktoru

Vnitroskupinový rozptyl - “Jaká je přirozená variabilita uvnitř jednotlivých výběrů v našem měření” Rozptyl jedné skupiny.

Meziskupinový rozptyl - “Jak daleko jsou průměry od sebe”. Rozptyl mezi skupinami.

1) Zkoumali jsme tři druhy benzínu a u každého jsme udělali 5 měření spotřeby. Meziskupinový rozptyl jsme vyčíslili na 0,28 a vnitroskupinový se rovná 0,09. Ověřte hypotézu, že se spotřeby u těchto třech druhů rovnají a určete intenzitu této závislosti.

2) Ve 4 lokalitách bylo náhodně osloveno celkem 28 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.

Lokalita	Počet návštěv
A	2,3,3,2,4,3,3
B	4,5,4,4,5,4,5
C	2,3,3,2,1,2,2
D	3,3,2,3,3,4,3

Lokalita	Počet návštěv						
A	2	3	3	2	4	3	3
B	4	5	4	4	5	4	5
C	2	3	3	2	1	2	2
D	3	3	2	3	3	4	3

Excel: Data - Analýza Dat - Anova: jeden faktor

Anova: jeden faktor

Faktor

Výběr	*Počet*	*Součet*	*Průměr*	*Rozptyl*
Řádek 1	7	20	2,8571	0,4762
Řádek 2	7	31	4,4286	0,2857
Řádek 3	7	15	2,1429	0,4762
Řádek 4	7	21	3	0,3333

ANOVA

Zdroj variability	*SS*	*Rozdíl*	*MS*	*F*	*Hodnota P*
Mezi výběry	19,25	3	6,4167	16,333	5,3619E-06
Všechny výběry	9,429	24	0,3929		
Celkem	28,68	27			

3) Doplňte chybějící tabulku, udělejte test včetně zapsání hypotéz a vyvoďte závěry, vhodným ukazatelem změřte sílu závislosti a závěr též okomentujte.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	456, 1333333	228, 0666	?	<.0001
Error	12	23, 33333333	?		
Corrected Total	14	?			

Zabýváme se dvojicí veličin:
Y - vysvětlovaná, závisle proměnná
X - vysvětlující, nezávisle proměnná

Hledáme vysvětlení chování Y při dané hodnotě X.

- $x_1, ..., x_n$ - dané hodnoty proměnné X
- $y_1, ..., y_n$ - naměřené (náhodné) hodnoty proměnné Y
- ε_i - náhodná chyba $N(0, \sigma^2)$ (normální rozdělení)
- β_1 - průměrná změna Y při jednotkové změně X
- β_0 - průměrná hodnota Y při X=0 (konstanta)

součet čtverců	
Celkový	$S_y = \sum (y_i - \bar{y})^2$
Teoretický	$S_T = \sum (\hat{y}_i - \bar{y})^2$
Reziduální	$S_R = \sum (y_i - \hat{y}_i)^2 = \sum \hat{\varepsilon}_i^2$

Regresní přímka: $y = \beta_0 + \beta_1 x + \varepsilon,$

Regresní parabola: $Y = b_0 + b_1 x + b_2 x^2, \quad \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$

Vícenásobná lineární regrese: $Y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k, \quad \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$

Koeficient determinace (Index determinace):

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

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

Upravený index determinace:

$$I^2_{ADJ} = R^2_{ADJ} = 1 - (1 - I^2)\frac{n-1}{n-p}$$

Je možné jej použít např. proto, abychom rozhodli, zda je lepším modelem regresní přímka nebo regresní parabola. Což Indexem determinace nemůžeme.

Korelační koeficient:

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

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

a) Modelujte závislost nákladů na údržbu na ceně regresní přímkou.
b) Zhodnoťte kvalitu modelu pomocí koeficientu determinace.
c) Ověřte pomocí testu, zda se jedná o významnou závislost.
d) Interpretujte věcně hodnotu regresního koeficientu b1.
e) Odhadněte střední hodnotu nákladů u domů za 80 tis. dolarů

VÝSLEDEK

Regresní statistika	
Násobné R	0,984743
Hodnota spolehlivosti R	0,969718
Nastavená hodnota spoleh	0,965933
Chyba stř. hodnoty	58,59154
Pozorování	10

ANOVA

	Rozdíl	*SS*	*MS*	*F*	*ýznamnost F*
Regrese	1	879463,2	879463,2	256,1815	2,33E-07
Rezidua	8	27463,75	3432,969		
Celkem	9	906926,9			

	Koeficienty	*ba stř. hodn*	*t Stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*	*Dolní 99,0%*
Hranice	-160,347	41,00253	-3,91066	0,004477	-254,899	-65,7949	-297,926
Soubor X 1	7,573698	0,473188	16,00567	2,33E-07	6,482524	8,664873	5,985968

Excel: Data - Analýza Dat - Regrese (x - nezávislá, y-závislá)

2) Tabulka obsahuje údaje o stáří, počtu najetých km a ceně 20 ojetých aut značky Octavia Combi.
a) sestavte dvojnásobný reg. model pro závislost ceny na stáří a počtu najetých km.
b) Odhadněte bodově průměrnou cenu nového auta.
c) Z kolika % zle změny v ceně vysvětlit nalezeným reg. modelem.
d) Použijte reg. model k odhadu ceny auta starého 6 let, které má najeto 60 tis.km a to jak bodově, tak intervalově.

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