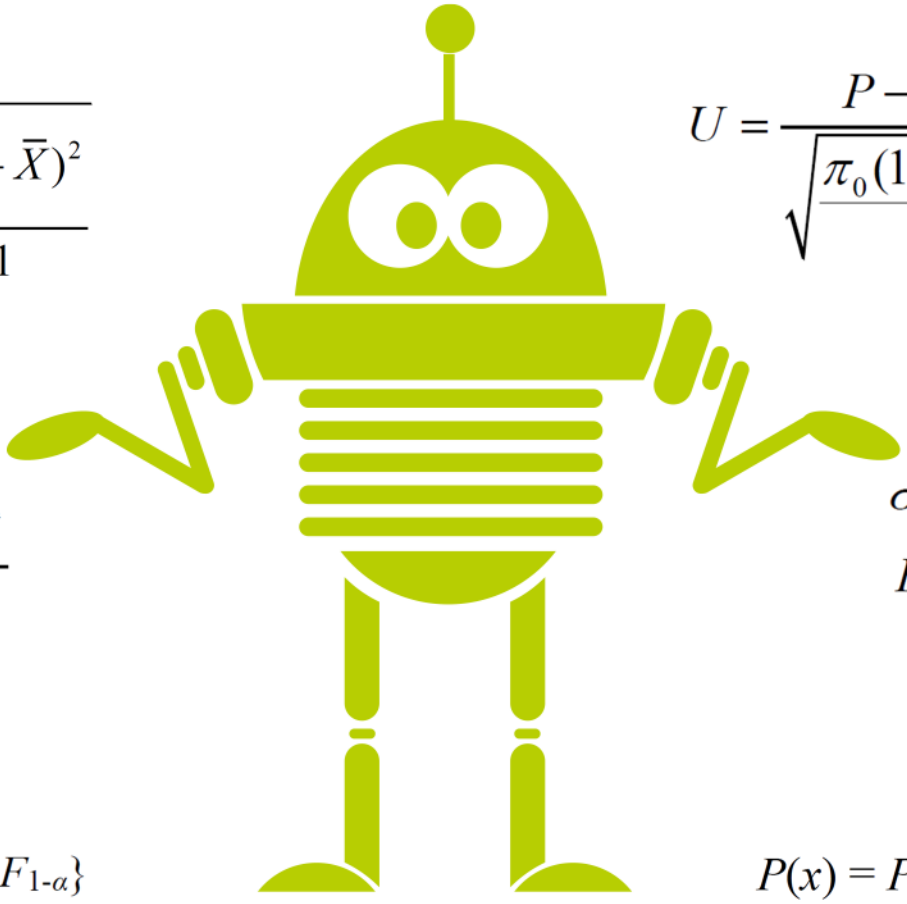
STATISTIKA II

ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA

ONLINE KURZ PRO PAE

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

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



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

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

$$E(X) = \pi$$

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

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

EDU FOR LIFE

VZDĚLÁNÍ, KTERÉ SE TI BUDE HODIT

Adriana Řeháčková

www.statistickyneklasicky.cz

STATISTIKA
ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA

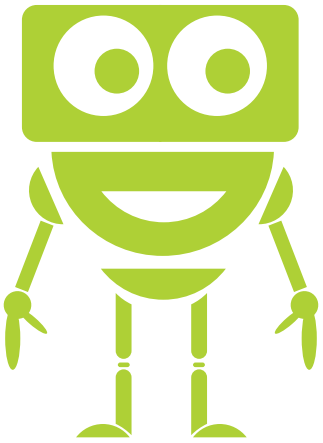
Ukážeme si, že statistika není žádná nuda a že má opravdu smysl! Cílem kurzu je, abys dokázal základní až lehce pokročilé statistické metody správně použít při řešení příkladů. Ukážeme si jak pomocí kalkulačky a ručních výpočtu řešit typové úlohy z předmětu statistika II pro obor PAE.

OBSAH KURZU:

1) Regresní modely
2) Vícenásobná lineární regrese
3) Kontingenční tabulky
4) Úvod do časových řad
5) Časové řady - Lineární trendová funkce a odhady

O TOMTO MATERIÁLU:

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



STATISTICKY NEKLASICKY

JEDNODUCHÁ LINEÁRNÍ REGRESE

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

Hledáme vysvětlení chování Y při dané hodnotě X.

- $x_1, \ldots, x_n$ - dané hodnoty proměnné X
- $y_1, \ldots, y_n$ - 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: $y = \beta_0 + \beta_1 x + \varepsilon,$

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:

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

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

Linear **Přímka**	$y' = b_0 + b_1 \cdot x$	lin. v prom. a lin. v para
Quadratic **Parabola**	$y' = b_0 + b_1 \cdot x + b_2 \cdot x^2$	nelin. v prom. a lin. v para
Inverse **Hyperbola** (Lomená)	$y' = b_0 + b_1 / x$	nelin. v prom. a lin. v para
Compound **Obecná exponenciála**	$y' = b_0 \cdot b_1^{\,x}$ $\ln(y) = \ln(b_0) + \ln(b_1)x$	nelin. v prom. a nelin. v para

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 koeficient korelace sílu závislosti.
d) Interpretujte věcně hodnotu regresního koeficientu b1.
e) Odhadněte střední hodnotu nákladů u domů za 80 tis. dolarů

2) V tabulce jsou uvedeny zisky firmy (v korunách) a počty zákazníků.

a) Modelujte závislost zisku firmy na počtu zákazníků pomocí regresní hyperboly.
b) Popište kvalitu daného regresního modelu.
c) Odhadněte střední hodnotu zisku firmy s 10 zákazníky.

Počet zákazníků	Zisk
1	500
2	746
2	760
3	810
5	917
6	961
6	980
8	1 028
9	1 052
10	1 070
12	1 084
13	1 082
17	1 060
21	800

3) 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) Vypočtěte koeficient úplné vícenásobné korelace.
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

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

Testování závislosti kvalitativních znaků (kontingenční tabulky)

Závislost 2 a více kvalitativních znaků analyzujeme opět statistickou analýzou četnostních tabulek a testujeme pomocí Chí-testu. Výpočet vychází z empirických a teoretických četností současného výskytu sledovaných znaků v souboru.

Pozorované empirické četnosti sestavíme do tzv. kontingenční tabulky, jejíž velikost se řídí počtem sledovaných znaků.

Pokud hodnota **spadá (nerovnost platí) do kritického oboru**, tak mezi sledovanými jevy existuje **statisticky významná závislost**.
Pokud hodnota **nespadá (nerovnost neplatí) do kritického oboru**, tak závislost mezi sledovanými jevy **není statisticky významná**.

1) Nemocniční zařízení nabízí několik druhů zdravotnických služeb. Ředitel nemocnice si přeje ověřit, zda spokojenost pacientů s nabízenými službami nemocnice závisí na jejich pohlaví. Bylo osloveno 750 náhodně vybraných bývalých pacientů, které již služeb nemocnice využili, bylo zaznamenáno jejich pohlaví a byli dotázáni, zda byli či nebyli spokojeni s nabízenými službami. Výsledky jsou zobrazeny v tabulce níže.

   Na hladině významnosti 0,05 na základě odpovídajícího testu určete, zda v dané nemocnici existuje souvislost mezi pohlavím a spokojeností s nabízenými službami nemocnice.

Spokojenost/pohlaví	Muži	Ženy
Spokojen(a)	288	312
Nespokojen(a)	60	90

2) Na základě dat v následující tabulce určete, zda existuje závislost mezi pohlavím a počty kuřáků. (alfa = 0,05).

Pohlaví/kouření	Kuřák	Nekuřák
muž	6	2
žena	5	7

3) Pro vhodný přístup v personální politice potřebuje vedení podniku vědět, zda spokojenost v práci závisí na tom, jedná-li se o pražský závod či závody mimopražské. Výsledky šetření jsou v následující tabulce, Případnou závislost změřte pomocí známých měr kontingence.

Místo/stupeň spokojenosti	Velmi nespokojen	Spíše nespokojen	Spíše spokojen	Velmi spokojen
Praha	10	25	50	15
Venkov	20	10	130	40

4) Ve firmě Anderson s.r.o. se hodnotí spokojenost zaměstnanců. Na základě výsledků v tabulce, rozhodněte zda spokojenost závisí na prac. pozici zaměstnance.

	Silně nespokojen	spíše nespokojen	spíše spokojen	velmi spokojen
služby	3	4	8	3
administrativa	16	20	12	6
IT oddělení	0	5	2	8
marketing	3	5	5	5

Časová řada - Posloupnost hodnot sledovaného ukazatele, která je jednoznačně uspořádána z hlediska času. Časové řady slouží k popsání a porozumění zákonitostem ekonomických, finančních či jiných skutečností, jenž časovými řadami zachycujeme, zároveň toto porozumění můžeme využít i k budoucím předpovědím.

Dělení: a) Podle rozhodného časového hlediska: **intervalové vs okamžikové**.

Intervalová časová řada obsahuje řadu hodnot sledovaného ukazatele za určitý interval, tj. obsahuje tokové veličiny (př. HDP/rok, průměrný příjem/měsíc, zisk, tržby/rok,měsíc)

Okamžikové časové řada obsahuje řadu hodnot sledovaného ukazatele k určitému okamžiku, tj. stavové veličiny (př. stav bankovnímu účtu k určitému dni, počet zaměstnanců k určitému dni, cena akcie).

b) Podle rekvence s jakou získáváme data: **Dlouhodobé a krátkodobé**
Dlouhodobé - Roční (perioda mezi dvěma po sobě jdoucími záznamy je roční a delší).
Krátkodobé - Čtvrtletní, měsíční, týdenní, denní, atd. (perioda mezi dvěma po sobě jdoucími záznamy je kratší než 1 rok)

1) Firma zabývající se provozováním internetového portálu zaznamenala za posledních 8 let prudký rozvoj, který dokumentuje tabulka dosaženého zisku před zdaněním (v tis. Kč). Určete pro tuto řadu absolutní a relativní přírůstky, koeficienty růstu, průměrný absolutní přírůstek a průměrný koeficient růstu.

Rok	2000	2001	2002	2003	2004	2005	2006	2007
Zisk	958	1002	1281	1569	1899	2222	2855	3544

Rok	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**
Zisk	**958**	**1002**	**1281**	**1569**	**1899**	**2222**	**2855**	**3544**
Absolutní přírustek	-	44	279	288	330	323	633	689
Relativní přírustek		0,0459	0,2784	0,224824	0,210325	0,17009	0,284878	0,24133
Koeficient růstu		1,0459	1,2784	1,224824	1,210325	1,17009	1,284878	1,24133

Dekompozice (rozklad) časové řady:

$$y_t = T_t + C_t + S_t + \varepsilon_t \qquad y_t = T_t \cdot C_t \cdot S_t \cdot \varepsilon_t$$

V případě **aditivního modelu** jsou jednotlivé složky uvažovány ve svých skutečných napozorovaných hodnotách.

V případě **multiplikativního modelu** je uvažována ve své skutečně napozorované hodnotě pouze trendová složka. Ostatní složky se většinou uvádějí v relativních hodnotách vůči trendu a jsou tedy bezrozměrné

Zdroj: Tomáš Karel

INTERPOLACE A EXTRAPOLACE ČŘ
-interpolace
Přibližné určení chybějící hodnoty sledovaného ukazatele uvnitř ČŘ za předpokladu, že známe sousední hodnoty 1. prostřednictvím dvou sousedních hodnot (pomocí aritmetického průměru těchto hodnot * průměrný koeficient růstu)
2. prostřednictvím všech hodnot v ČŘ (pomocí trendové funkce)

-extrapolace
Určení hodnot ČŘ za interval známých hodnot, z pravidla do budoucnosti
1. statické prognózování – pomocí trendových funkcí a sezónních indexů odhadujeme budoucí úroveň ukazatele. Uvedený postup však naráží na předpoklad neměnnosti dosavadního vývoje
Uvedený nedostatek (neměnnost) odstraňují metody adaptivního prognózování a ARIMA.

Autokorelace - Hodnoty ukazatele v řadě se vzájemně ovlivňují, jedna má vliv na druhou.
Durbin-Watsonův koeficient autokorelace - Keficientu se pohybuje v rozmezí < 0 , 4 >.
Pokud je tato statistika rovna číslu 2, rezidua **nevykazují žádnou autokorelaci (to chceme),** hodnoty D menší než 2 značí pozitivní autokorelaci a hodnoty větší než 2 značí autokorelaci negativní.

Volba vhodného modelu trendu (míra shody)
1. základem pro rozhodování o vhodném typu trendové funkce jsou věcně ekonomická kritéria
2. druhou možností volby je analýza grafu časové řady (nebezpečí: vizuální výběr funkce může být subjektivní a závisí na měřítku!)
3. elementární charakteristiky časové řady
4. Pomocí indexu determinace
5. Pomocí chyb odhadu (MSE, ME, MAE,...)
6. Pomocí reziduální směrodatné odchylky

2) Máte k dispozici časovou řadu počtu přínosných společenských inovací v podniku Elona Muska v letech 2009-2018.

rok	Yt
2009	12
2010	13
2011	17
2012	18
2013	22
2014	21
2015	28
2016	26
2017	33
2018	32

a) Vypočtěte lineární trendovou funkci.
b) Určete index korelace a determinace.
c) Určete odhad počtu inovací v roce 2020.
d) Vypočtěte hodnotu průměrného koeficientu růstu za celé sledované období a s jeho využitím vypočtěte rovněž prognózu na rok 2020.

ČASOVÉ ŘADY – KLOUZAVÉ PRŮMĚRY

Klouzavé průměry:
Pokud chceme očistit časovou řadu od náhodných nebo sezónních vlivů, můžeme použít klouzavé průměry. Pokud chceme z časové řady odstranit sezónnost liché délky a zachytit trend, používáme prosté klouzavé průměry té samé délky jako je délka sezónnosti.

1) Následující časovou řadu (počet bezdrátových sluchátek v tisících od nejmenované společnosti) očistěte od sezónních vlivů pomocí vhodných klouzavých průměrů.

Čtvrtletí	2017	2018	2019
1.	12	13	8
2.	23	17	12
3.	11	14	13
4.	8	11	15

Jaké klouzavé průměry jste použili k očištění od sezónních vlivů?
a) prosté klouzavé průměry
b) centrované klouzavé průměry délky 3
c) centrované klouzavé průměry délky 4
d) klouzavé průměry délky 4

I.17	12		
II.17	23	13,5	
III.17	11	13,75	13,625
IV.17	8	12,25	13
I.18	13	13	12,625
II.18	17	13,75	13,375
III.18	14	12,5	13,125
IV.18	11	11,25	11,875
I.19	8	11	11,125
II.19	12	12	11,5
III.19	13		
IV.19	15		

Kolik je hodnota výsledného klouzavého průměru pro 3. čtvrtletí 2017?
a) 13
b) nelze vypočítat
c) 13,625
d) 13,75

Kolik je hodnota výsledného klouzavého průměru pro 3. čtvrtletí 2019?
a) 11,67
b) 13
c) 11,5
d) nelze vypočítat

Kolik je hodnota výsledného klouzavého průměru pro 1. čtvrtletí 2018?
a) 13,375
b) 13
c) 12,625
d) nelze vypočítat

Individuální indexy (indexy stejnorodých ukazatelů):
a) jednoduché - Zkoumáme jeden výrobek na jedné pobočce
b) složené - Zkoumáme jeden výrobek na více pobočkách (za více poboček dohromady)

Souhrnné indexy (indexy nestejnorodých ukazatelů):
a) Paascheho – fixuje druhou veličinu v běžném období
b) Laspayresův – fixuje druhou veličinu v základním období
c) Fishrerův – geometrický průměr Paascheho a Laspayresova

Za produkci podniku, který vyrábí dva druhy automobilů, máte tyto údaje za rok 2017 a 2018:

Automobil	Hodnota výroby v mil. Kč		Produkce v tis. ks	
	2017	2018	2017	2018
Peugeot 407	100	145	15	23
Peugeot 308	120	105	32	38

Řešení:

1) Jak se změnil fyzický objem výroby podniku ve sledovaném období měřený Laspeyresovým indexem?
a) Zvýšil se o 0,1345 %
b) Zvýšil se o více než 50 %.
c) Zvýšil se o 34,5 %.
d) Zvýšil se o 3,45 %

2) Jak se změnil fyzický objem výroby podniku ve sledovaném období měřený Paascheho indexem?
a) Zvýšil se o 3,66 %.
b) Zvýšil se o 0,1366 %.
c) Zvýšil se o více než 50 %.
d) Zvýšil se o více % než v případě měření Laspayresovým indexem

3) Jak se změnil fyzický objem výroby podniku ve sledovaném období měřený Fisherovým indexem?
a) Zvýšil se méně než v případě měření Paascheho indexem.
b) Zvýšil se méně než v případě měření Laspayresovým indexem.
c) Zvýšil se o více než 50 %.
d) Zvýšil se o 3,24 %.