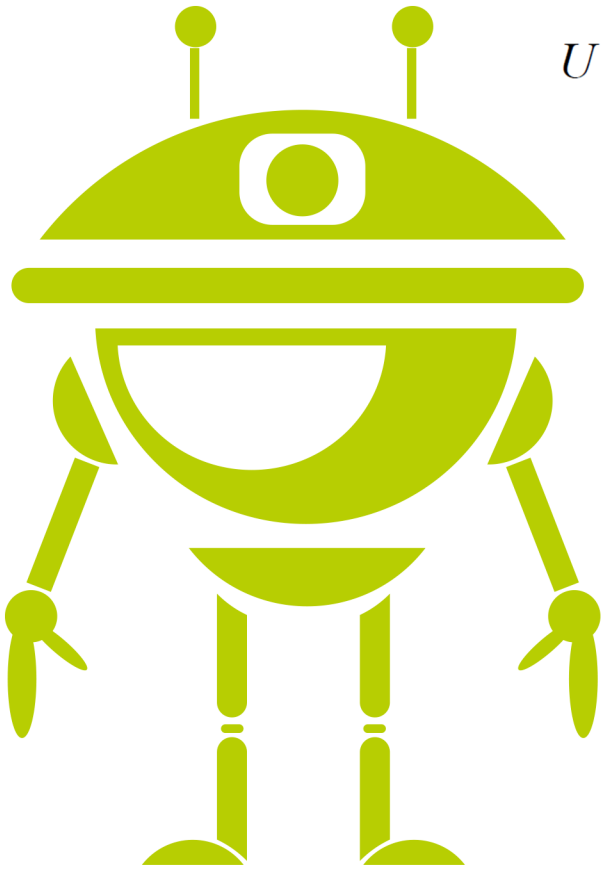
STATISTIKA II

ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA

PŘÍPRAVA NA ZÁPOČTOVÝ TEST

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

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



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

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

$$E(X) = \pi$$

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

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

EDU FOR LIFE

VZDĚLÁNÍ, KTERÉ SE TI BUDE HODIT

Adriana Řeháčková

www.statistickyneklasicky.cz

STATISTIKA
ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA

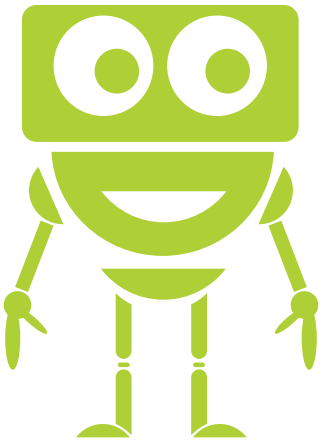
Ukážeme si, že statistika není žádná nuda a že má opravdu smysl! Cílem kurzu je, abys dokázal základní až lehce pokročilé statistické metody správně použít při řešení příkladů. Nemine Tě ani osvojení SPSS. Po kurzu bys měl vědět, co, kdy a jak použít a dokázat tyto znalosti využít na reálných datech. Tento kurz slouží jako příprava k zápočtovému testu.

OBSAH KURZU:

1) Testy dobré shody
2) Normální rozdělení
3) Analýza kategoriálních dat (Kontingenční tabulky)
4) Jednoduchá lineární regrese a korelace
5) Nelineární korelace a regrese
6) Vícenásobná regrese a korelace
7) Časová řada, základní charakteristiky, trendová funkce

O TOMTO MATERIÁLU:

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



STATISTICKY NEKLASICKY

TESTY DOBRÉ SHODY

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?

SPSS výpočet:

- počítáme, buď jestli je rozdělení rovnoměrné nebo máme předem dané rozdělení a ověřujeme pravdivost tohoto tvrzení, když je příklad zadán jen tabulkou četností (leden – 45; únor – 20, atd.), musíme zadat podle čeho se má četnost určovat (vážit) -> **Data -> Weight cases** – to, podle čeho se váží. Váhá nám číslo (x=45) převede na četnost (n=45).

Výpočet testu:
ANALYZE -> NONPARAMETRIC TESTS -> LEGACY DIALOGS -> > CHI-SQUARE
- zobrazí se tabulka, do části „test variable list“ vložíme testovanou proměnnou (musí mít vlastnosti – numeric a ordinal/nominal) a potvrdíme. Dále se dá nastavit zda budou rozděleny rovnoměrně či nikoli, ukážeme si na příkladech.

1) 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?

Ověřte na 5% hladině významnosti, zda průzkum odpovídá očekávání.

2) Pošta zjišťovala, zda zákazníci rozeselají balíčky rovnoměrně v rámci celého týdne.
V následující tabulce je uveden průměrný počet odeslaných balíčků pro jednotlivé dny v týdnu.

Den v týdnu	PO	UT	ST	ČT	PA	SO	NE
Průměrný počet odeslaných balíčků	18	16	14	20	19	7	6

a) Otestujte, zda jsou balíčky rozesílány rovnoměrně (na 5% hl. významnosti):
b) Otestujte, zda pošta resílá balíčky v následujícím poměru (opět na 5% hl. význ.):

Den v týdnu	PO	UT	ST	ČT	PA	SO	NE
Průměrný počet odeslaných balíčků	17,0%	17,0%	17,0%	17,0%	17,0%	7,5%	7,5%

3) Výrobce udává, že jsou jeho čokoládové pralinky zastoupeny v tomto poměru:

Pralinka	Podíl
oříšková	15%
hořká	20%
mléčná	45%
karamelová	20%

a) na vzorku 120 pralinek různých příchutí tvrzení výrobce ověřte.
b) Ověřte, že hmotnost pralinek má normální rozdělení.

(použijte soubor pralinky.sav)

TESTOVÁNÍ NORMALITY
Testování normality je nezbytný předpoklad pro výběr vhodného typu testu (parametrické = mají normální rozdělení, neparametrické = nemají normální rozdělení).
Postup je následující: -> **ANALYZE -> DESCRIPTIVE STATISTICS -> EXPLORE**
A do pole **DEPENDENT LIST** vybereš proměnnou, kterou chceš testovat a následně do **FACTOR LIST** můžeš zvolit zde ji chceš podle něčeho třídit (například vyberu v dependent list výsledky z testů a ve factor list pohlaví, tím získám testy za každé pohlaví zvlášť.

U normality platí, že pokud je sig (p-hodnota) **menší než 0,05**, tak data **nemají normální rozdělení** (zamítáme H0, která hovoří o normality). Pokud je **p-hodnota větší než 0,05** potom data **mají normální rozdělení.**

Rovněž si zobrazíme histogram, kde se můžeme i vizuálně vidět, zda se data chovají podle Gaussovy křivky (mají normální rozdělení) nebo ne.

KONTINGENČNÍ TABULKY

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:

Asociační tabulka (2x2 tab.)
- používá se chí kvadrát test nebo Fischerův test, určuje se to dle četností:

chí kvadrát -> **pokud je očekávaná četnost u všech znaků větší než 5.**
Hodnot by mělo být více jak 20.

Fischerův test -> pokud je alespoň 1 očekávaná četnost menší než 5.
Pro 20 a méně hodnot.

Kontingenční tabulka
- používá se pouze chí kvadrát test, ale tabulka musí splňovat dva předpoklady:
-> **podíl teoretických četností menších než 5 nesmí překročit 20%**
-> **žádná z teoretických četností nesmí být menší než 1**
- pokud jsou tyto podmínky porušeny, je nutné spojit slabé skupiny:
-> slučujeme řádky nebo sloupce – musí být logické, věcně správné a dobře interpretovatelné
-> poté opět vyjádříme teoretických četnosti a kontrolujeme podmínku

Výpočet závislosti znaků
-> ANALYZE -> DESCRIPTIVE STATISTICS -> CROSSTABS
-> vložit proměnné do řádku (Row) a sloupce (Column)
-> kliknout na Statistics a zaškrtnout Chi-square

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**.

KONTINGENČNÍ TABULKY

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	4	3
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:

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

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:

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

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

JEDNODUCHÁ LINEÁRNÍ REGRESE

1. Graf
- nejprve určíme, co je Y (závislá proměnná) a X (nezávislá proměnná)
-> GRAPHS -> LEGACY DIALOGS -> SCATTER/DOT -> SIMPLE SCATTER
- vložíme do Y a X proměnné, které jsme si určili
- můžeme vidět těsnou/volnou a přímou/nepřímou závislost

2. Proložení korelačního pole regresní přímkou
-> rozkliknout graf a dát -> ELEMENTS -> FIT LINE AT TOTAL
- graf se proloží regresní přímkou a vpravo můžeme vybrat její typ (linear, cubic, etc.)

3. Výpočet
-> ANALYZE -> REGRESSION -> LINEAR
- vloží se závisle (dependent) a nezávisle (independent) proměnná
- po potvrzení vyjedou v outputu čtyři tabulky

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) Vytvořte a zapište lineární regresní model závislost nákladů na údržbu na ceně.
b) Určete z kolika % jsou změny v nákladech na udržbu ovlivněny cenou domu.
c) Proveďte test významnosti regresního koeficientu a jeho intervalový odhad (alfa = 0,05).
d) Interpretujte věcně hodnotu regresního koeficientu b1.
e) Odhadněte střední hodnotu nákladů u domů za 80 tis. dolarů
f) Vypočítejte a interpretujte koeficient korelace.

2) Ve 12 náhodně vybraných prodejnách byly zjištěny výdaje na zeleniny a výdaje na zeleninové saláty (v tis. Kč.). Charakterizujte závislost zeleninových salátů na zelenině.

x	y
17	19
23	30
14	20
18	23
45	56
34	44
45	48
18	23
28	33
76	81
31	40
44	51

1) Proveďte test významnosti regresního koeficientu (α=0,05).
2) Stanovte 95-ti procentní interval spolehlivosti pro regresní koeficient.
3) Proveďte test významnosti koeficientu korelace (α=0,05).
4) Stanovte 95-ti procentní interval spolehlivosti pro korelační koeficient.

Intervalový odhad korelačního koeficientu ρ_{yx}:

je-li n<100, potom Fischerova z-transformace

$z \in (z_r \pm u_\alpha . s_z)$

$$s_z = \frac{1}{\sqrt{n-3}}$$

je-li n>100, potom $r_{yx} \pm u_\alpha . s_r$

$$s_r = \frac{1-r_{yx}^2}{\sqrt{n-2}}$$

BONUS

Ve 12 náhodně vybraných prodejnách byly zjištěny výdaje na zeleniny a výdaje na zeleninové saláty (v tis. Kč.). Charakterizujte závislost zeleninových salátů na zelenině.

x	y
17	19
23	30
14	20
18	23
45	56
34	44
45	48
18	23
28	33
76	81
31	40
44	51

U předešlé úlohy určete:
a) Zobrazte pás spolehlivosti pro jednotlivá pozorování a pás spolehlivosti pro reg. přímku.
b) Ověřte předpoklady modelu - analýza reziduí.

Typy funkcí a jejich názvy v SPSS (linearita modelu a proměnných)

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
Logarithmic **Logaritmus**	$y' = b_0 + b_1 \cdot \ln(x)$	nelin. v prom. a lin. v para
Cubic **Kubická křivka**	$y' = b_0 + b_1 \cdot x + b_2 \cdot x^2 + b_3 \cdot x^3$	nelin. v prom. a lin. v para
Power **Obecná mocnina**	$y' = b_0 \cdot x^{b_1}$ $\ln(y) = \ln(b_0) + b_1 \cdot \ln(x)$	nelin. v prom. a nelin. v para
S **S - křivka**	$y' = e^{b_0 + b_1 x}$ $\ln(y) = b_0 + b_1 / x$	nelin. v prom. a nelin. 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
Growth **Růstová křivka**	$y' = e^{b_0 + b_1 x}$ $\ln(y) = b_0 + b_1 \cdot x$	nelin. v prom. a nelin. v para
Exponential **Exponenciála**	$y' = b_0 \cdot e^{b_1 x}$ $\ln(y) = \ln(b_0) + b_1 x$	nelin. v prom. a nelin. v para

NELINEÁRNÍ KORELACE A REGRESE

Nalezení nejvhodnější funkce, zápis funkce a vyjádření jejích charakteristik
-> ANALYZE -> REGRESSION -> CURVE ESTIMATION
- vloží se závislá a nezávislá proměnná
- zaškrtneme, které funkce chceme spočítat a potvrdíme

1) V tabulce jsou uvedeny údaje o dvanácti vozech značky Toyota vyrobené v roce 2019:

Cena v tis. Kč	Počet najetých km (tis.)
150	83
130	97
125	112
170	56
193	39
218	28
230	34
105	112
128	98
167	80
189	48
163	58

a) Rozhodněte, který regresní model lépe vystihuje průběh závislosti ceny na počtu najetých km a uveďte důvod vašeho rozhodnutí. Rovnici zapište.
b) Z kolika % lze variabilitu cenu vysvětlit pomocí počtu najetých km.
c) Otestujte významnost modelu jako celku (alfa=0,05)
d) Pomocí zvolené funkce odhadněte cenu auta, které má najeto 100 tis. km - proveďte bodový i intervalový odhad.

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

a) Modelujte závislost zisku firmy na počtu zákazníků pomocí vhodného reg. modelu.
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

. VÍCENÁSOBNÁ KORELACE A REGRESE
- jedna závislá proměnná (Y) a dvě či více nezávislých proměnných (X1, X2,… Xn)

ANALYZE -> REGRESSION -> LINEAR
Těsnost (síla závislosti; korelační koeficient):
a) celková (totální)
b) párová - měření těsnosti dvou proměnných na sobě
> ANALYZE -> CORRELATE -> BIVARIATE
c) parciální
- zjišťuje se těsnost závisle proměnné na jedné nezávisle proměnné, když ostatní nezávisle proměnné jsou konstantní
-> ANALYZE -> CORRELATE -> PARTIAL

1) 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) Interpretujte par. reg. koeficient u proměnné najeto.
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

2) V následující tabulce jsou uvedeny údaje pro vybrané telefonní služby společnosti U2.

Cena	90	52	89	65	45	63	76	62
Počet minut	100	50	110	100	50	70	78	65
počet SMS	4	3	3	2	3	2	3	4

a) Sestrojte dvojnásobný reg. model pro závislost celkové ceny na počtu provolaných minut a odeslaných SMS.
b) Interpretujte parciální reg. koeficient u proměnné počet provolaných minut.
c) Proveďte test významnosti parciálního reg. koeficientu u proměnné počet provolaných minut.
d) Odhadněte o kolik by se změnila celková cena, při zvýšení počet minut o 10 v případě, že počet odeslaných SMS zůstane konstantní.
e) Proveďte celkový test významnosti zvoleného modelu.

ČASOVÉ ŘADY

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 frekvence 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)

ČASOVÉ ŘADY
- u časových řad se pro nezávisle proměnnou místo x používá „t“
- funkce jsou jinak stejné jako pro regresní modely
- když počítáme časové řady, VŽDY se musí nadefinovat, že se o časové řady jedná.

1.Definice datové proměnné
- nejdříve se musí v SPSS nadefinovat, že se jedná o časovou řadu
DATA -> DEFINE DATES
-> zde se zvolí, zda máme časové řady roční, měsíční, apod.

2. Grafické zobrazení časové řady
ANALYZE -> FORECASTING -> SEQUENCE CHARTS

3. Výpočet a interpretace 1. diference
TRANSFORM -> CREATE TIME SERIES

4. Vyrovnání klouzavým průměrem
TRANSFORM -> CREATE TIME SERIES a následně Centered moving average

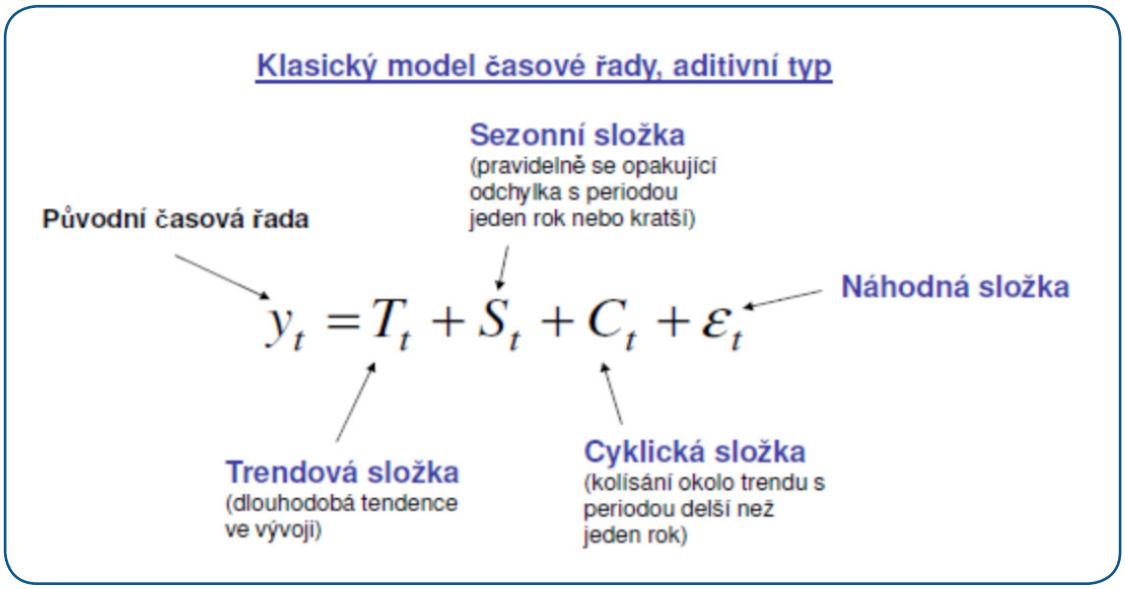
5. Vyrovnání lineární trendovou funkcí
ANALYZE -> REGRESSION -> CURVE ESTIMATION

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. Průměrný absolutní přírustek a průměrný koeficient růstu. Ve kterém roce došlo k největšímu nárustu zisku?

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

2) Spočítejte bázické indexy, kde bází je rok 2000 a interpretujte první dva bázické indexy. Následně vyrovnejte časovou řadu klouzavými průměry délky 3 a 5.

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



Máme časovou řadu ročních údajů výroby tlačítkových telefonů v tis. ks v letech 2008-2017.

rok	Yt
2008	450
2009	342
2010	321
2011	295
2012	324
2013	270
2014	210
2015	306
2016	251
2017	221

a) Vyberte vhodnou trendovou funkci pro zachycení vývoje tlačítkových telefonů. Výběr zdůvodněte.
b) Proveďte celkový test významnosti modelu.
c) Odhadněte a správně zapište celkový počet prodaných tlačítkových telefonů v roce 2018 a 2020.