Popis časových řad

Pojem časové řady: Časovou řadou rozumíme řadu hodnot $y_{t_1}, \ldots, y_{t_n}$ určitého ukazatele uspořádanou podle přirozené časové posloupnosti $t_1 < \ldots < t_n$. Jsou-li časové intervaly (t_1, t_2), ..., (t_{n-1}, t_n) stejně dlouhé (ekvidistantní), zjednodušeně zapisujeme časovou řadu jako $y_1, \ldots, y_n$. Přitom ukazatel je veličina, která charakterizuje nějaký jev v určitém prostoru a určitém čase (okamžiku či intervalu).

Druhy časových řad

a) Časová řada okamžiková: příslušný ukazatel udává, kolik jevů existuje v daném časovém okamžiku (např. počet obyvatelstva k určitému dnu).

b) Časová řada intervalová: příslušný ukazatel udává, kolik jevů vzniklo či zaniklo v určitém časovém intervalu (např. počet sňatků během roku). Nejsou-li jednotlivé časové intervaly ekvidistantní, musíme provést očištění časové řady od důsledků kalendářních variací.

Příklad: Máme k dispozici údaje o tržbě obchodní organizace (v tis. Kč) v jednotlivých měsících roku 1995: 2400, 2134, 2407, 2445, 2894, 3354, 3515, 3515, 3225, 3063, 2694, 2600. Vypočtěte očištěné údaje.

Řešení: Průměrná délka měsíce je 365/12 dne. Očištěná hodnota

pro leden $y_1^{(o)} = 2400 \cdot \frac{365}{12 \cdot 31} = 2354{,}84$,

pro únor $y_2^{(o)} = 2134 \cdot \frac{365}{12 \cdot 28} = 2318{,}18$.

Pro ostatní měsíce analogicky dostaneme

2361,71; 2478,96; 2839,54; 3400,58, 3448,86; 3448,86; 3269,79; 3005,36; 2731,42; 2551,08.

Výpočet pomocí systému STATISTICA:

Vytvoříme nový datový soubor o třech proměnných: trzba, dm (délky jednotlivých měsíců) a ot (očištěná tržba) a 12 případech. Do proměnné trzba zapíšeme zjištěné hodnoty. Do proměnné dm vložíme délky jednotlivých měsíců, tj. 31, 28, 30, …, 31. Do Dlouhého jména proměnné ot napíšeme =trzba*365/(12*dm).

	1 trzba	2 dm	3 ot
1	2400	31	2354,839
2	2134	28	2318,185
3	2407	31	2361,707
4	2445	30	2478,958
5	2894	31	2839,543
6	3354	30	3400,583
7	3515	31	3448,858
8	3515	31	3448,858
9	3225	30	3269,792
10	3063	31	3005,363
11	2694	30	2731,417
12	2600	31	2551,075

Grafické znázornění okamžikové časové řady

Použijeme spojnicový diagram. Na vodorovnou osu vynášíme časové okamžiky $t_1, ..., t_n$, na svislou osu odpovídající hodnoty $y_1, ..., y_n$. Dvojice bodů (t_i, y_i), $i = 1, ..., n$ spojíme úsečkami.

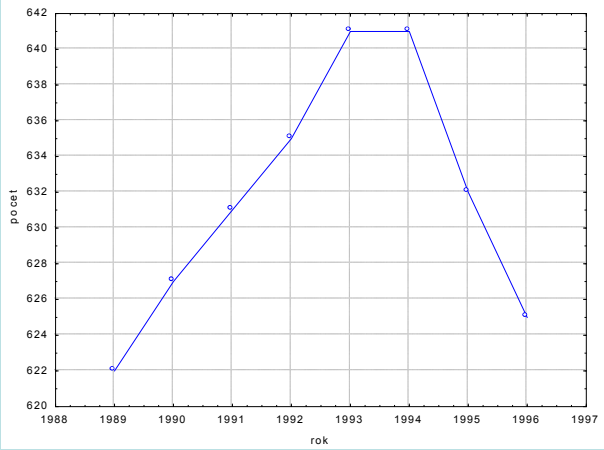
Příklad: Časová řada obsahuje údaje o počtu zaměstnanců určité akciové společnosti v letech 1989 – 1996 vždy k 31.12.

1989	1990	1991	1992	1993	1994	1995	1996
622	627	631	635	641	641	632	625

Znázorněte tuto časovou řadu graficky.

Řešení pomocí systému STATISTICA:

Vytvoříme datový soubor o dvou proměnných nazvaných rok a pocet a 8 případech.
Grafy – Bodové grafy – odškrtneme Lineární proložení – Proměnné X – rok, Y – počet – OK – OK. 2x klikneme na pozadí grafu – vybereme Graf: obecné – zaškrtneme Spojnice – OK.



Grafické znázornění intervalové časové řady

Použijeme sloupkový diagram. Je to soustava obdélníků, kde šířka obdélníku je rovna délce intervalu a výška odpovídá hodnotě ukazatele v daném intervalu. Ke znázornění intervalové časové řady lze použít i spojnicový diagram, přičemž na vodorovnou osu vynášíme středy příslušných intervalů.

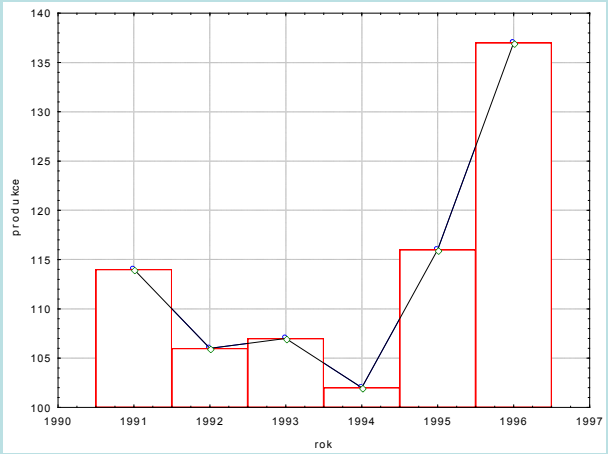
Příklad: Máme k dispozici údaje o produkci určitého podniku (v tisících výrobků) v letech 1991-1996.

1991	1992	1993	1994	1995	1996
114	106	107	102	116	137

Znázorněte tuto časovou řadu graficky.

Řešení pomocí systému STATISTICA:

Vytvoříme datový soubor o dvou proměnných nazvaných rok a produkce a 6 případech.
Grafy – Bodové grafy – odškrtneme Lineární proložení – Proměnné X – rok, Y – produkce – OK – OK. 2x klikneme na pozadí grafu – vybereme Graf: obecné – zaškrtneme Spojnice – Přidat nový graf – typ Sloupcový graf – OK. Do sloupců označených jako Nový1, Nový2 okopírujeme hodnoty proměnných rok a produkce. Ve Všech možnostech: Sloupce upravíme šířku sloupce na 1.



Průměr okamžikové časové řady

Nejprve vypočteme průměry pro jednotlivé dílčí intervaly $(t_1, t_2), (t_2, t_3), \ldots, (t_{n-1}, t_n)$: $\frac{y_1 + y_2}{2}, \frac{y_2 + y_3}{2}, \ldots, \frac{y_{n-} + y_n}{2}$. Jsou-li všechny tyto intervaly stejně dlouhé, vypočteme prostý chronologický průměr okamžikové časové řady:

$$\bar{y} = \frac{1}{n-1} \sum_{i=2}^{n} \frac{y_{i-1} + y_i}{2} = \frac{1}{n-1} \left(\frac{y_1}{2} + \sum_{i=2}^{n-1} y_i + \frac{y_n}{2} \right).$$

Nemají-li intervaly stejnou délku, vypočteme $d_i = t_i - t_{i-1}$, $i = 2, \ldots, n$ a použijeme vážený chronologický průměr okamžikové časové řady:

$$\bar{y} = \frac{1}{\sum_{i=2}^{n} d_i} \sum_{i=2}^{n} \frac{y_{i-1} + y_i}{2} \cdot d_i .$$

Příklad: Časová řada vyjadřuje počet obyvatelstva ČSSR (v tisících) v letech 1965 až 1974 vždy ke dni 31.12.

Rok	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974
počet	14194	14271	14333	14387	14443	14345	14419	14576	14631	14738

Charakterizujte tuto časovou řadu chronologickým průměrem.

Řešení: $\bar{y} = \frac{1}{9} \left(\frac{14194}{2} + 14271 + \ldots + 14631 + \frac{14738}{2} \right) = 14430$.

Průměr intervalové časové řady

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i .$$

Příklad: Vypočtěte průměrnou hodnotu roční časové řady HDP ČR (v miliardách Kč) v letech 1994 až 2000.

1994	1995	1996	1997	1998	1999	2000
1303,6	1381,1	1447,7	1432,8	1401,3	1390,6	1433,8

Řešení: $\bar{y} = \frac{1}{7}(1303,6 + \ldots + 1433,8) = 1398,7$.

Dynamické charakteristiky časových řad

Absolutní přírůstky

1. diference: $\Delta y_i = y_i - y_{i-1}, i = 2, \dots, n$

2. diference: $\Delta^2 y_i = \Delta y_i - \Delta y_{i-1} = y_i - 2y_{i-1} + y_{i-2}, i = 3, \dots, n$

atd.

(Diferencování má velký význam při odhadu trendu časové řady regresními metodami.)

Průměrný absolutní přírůstek: $\Delta = \frac{\sum_{i=2}^{n} \Delta y_i}{n-1} = \frac{y_n - y_1}{n-1}$

Relativní přírůstek

$\delta_i = \frac{\Delta y_i}{y_{i-1}}, i = 2, \dots, n$

(Relativní přírůstek po vynásobení 100 udává, o kolik procent se změnila hodnota v čase t_i oproti času t_{i-1}.)

Koeficient růstu (tempo růstu)

$k_i = \frac{y_i}{y_{i-1}}, i = 2, \dots, n$

(Koeficient růstu po vynásobení 100 udává, na kolik procent hodnoty v čase t_{i-1} vzrostla či poklesla hodnota v čase t_i.)

Průměrný koeficient růstu

$\bar{k} = \sqrt[n-1]{k_2 \cdot k_3 \cdot \dots \cdot k_n} = \sqrt[n-1]{\frac{y_n}{y_1}}$

Průměrný relativní přírůstek

$\delta = \bar{k} - 1$

Příklad: Pro časovou řadu HDP ČR v letech 1994 až 2000 (v miliardách Kč) vypočtěte základní charakteristiky dynamiky a graficky znázorněte 1. diference a koeficienty růstu.

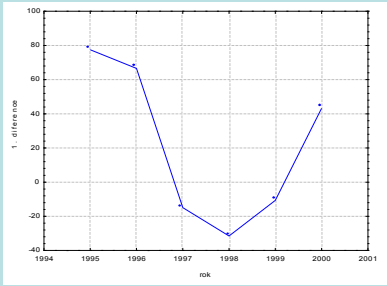
Řešení:

rok	HDP	Δy_i	k_i	δ_i
1994	1303,6	x	x	x
1995	1381,1	77,5	1,059	0,059
1996	1447,7	66,6	1,048	0,048
1997	1432,8	-14,7	0,990	-0,010
1998	1401,3	-31,5	0,978	-0,022
1999	1390,6	-10,7	0,992	-0,008
2000	1433,8	43,2	1,031	0,031

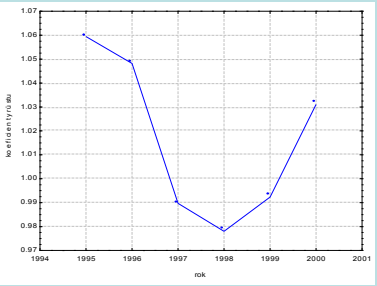
Průměrný absolutní přírůstek: $\Delta = \frac{1433,8 - 1303,6}{6} = 21,7$, tzn., že v období 1994 – 2000 rostl HDP průměrně o 21,7 miliard Kč ročně.

Průměrný koeficient růstu: $\bar{k} = \sqrt[6]{\frac{1433,8}{1303,6}} = 1,016$, tzn., že v období 1994 – 2000 rostl HDP průměrně o 1,6% ročně.

Graf 1. diferencí:

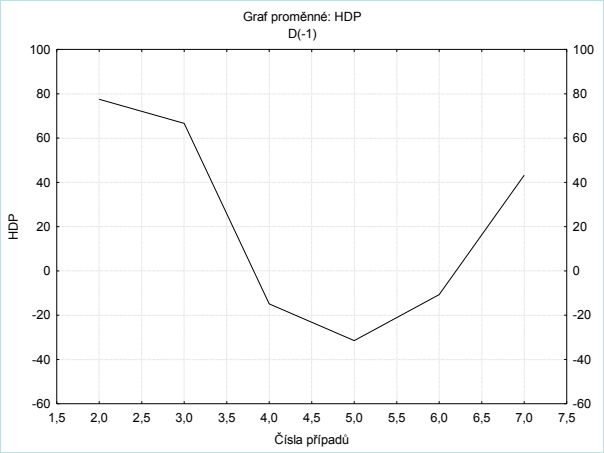


Graf koeficientů růstu:



Výpočet pomocí systému STATISTICA

Statistiky – Pokročilé lineární/nelineární modely – Časové řady/predikce – Proměnné HDP – OK – OK (transformace, autokorelace, kříž. korelace, grafy) – Diferencování - OK (transformovat vybrané řady) – vykreslí se graf.



Vrátíme se do Transformace proměnných – Uložit proměnné. Otevře se nové datové okno, kde v proměnné HDP_1 jsou uloženy 1. diference.

	HDP	HDP_1
1	1303,600	
2	1381,100	77,500
3	1447,700	66,600
4	1432,800	-14,900
5	1401,300	-31,500
6	1390,600	-10,700
7	1433,800	43,200

Výpočet relativních přírůstků: $\delta_i = \frac{\Delta_{y_i}}{y_{i-}}$ pro i = 2,...,n

Vrátíme se do Transformace proměnných – označíme proměnnou, kterou chceme transformovat (HDP) – vybereme Posun – OK, (Transformovat vybrané řady) – vykreslí se graf.
Vrátíme se do Transformace proměnných – Uložit proměnné. Tato transformovaná veličina se uloží do tabulky pod názvem HDP_1 (proměnná s 1. diferencemi se přejmenuje na HDP_2). Přidáme novou proměnnou RP a do jejího Dlouhého jména napíšeme vzorec =HDP_2/HDP_1.

Výpočet koeficientů růstu: $k_i = \frac{y_i}{y_{i-}}$ pro i = 2,...,n

Do tabulky přidáme proměnnou KR a do jejího Dlouhého jména napíšeme vzorec =HDP/HDP_1. Získáme tabulku

	1 HDP	2 HDP_2	3 HDP_1	4 RP	5 KR
1	1303,600				
2	1381,100	77,500	1303,600	0,059451	1,059451
3	1447,700	66,600	1381,100	0,048222	1,048222
4	1432,800	-14,900	1447,700	-0,01029	0,989708
5	1401,300	-31,500	1432,800	-0,02198	0,978015
6	1390,600	-10,700	1401,300	-0,00764	0,992364
7	1433,800	43,200	1390,600	0,031066	1,031066
8			1433,800		

Pomocí Grafy - 2D Grafy – Spojnicové grafy (Proměnné) vykreslíme průběh relativních přírůstků a koeficientů růstu.

Průměrný absolutní přírůstek a průměrný koeficient růstu vypočteme na kalkulačce pomocí vzorců

$\Delta = \frac{1433,8 - 1303,6}{6} = 21,7$ a $\bar{k} = \sqrt[6]{\frac{1433,8}{1303,6}} = 1,016$.

Aditivní model časové řady

Předpokládejme, že pro časovou řadu $y_1, ..., y_n$ platí model

$y_t = f(t) + \varepsilon_t$, $t = 1, ..., n$, kde

$f(t)$ je neznámá trendová funkce (trend), kterou považujeme za systematickou (deterministickou) složku časové řady (popisuje hlavní tendenci dlouhodobého vývoje časové řady),

ε_t je náhodná složka časové řady zahrnující odchylky od trendu. Náhodná složka splňuje předpoklady

$E(\varepsilon_t) = 0$,

$D(\varepsilon_t) = \sigma^2$,

$C(\varepsilon_t, \varepsilon_{t+h}) = 0$,

$\varepsilon_t \sim N(0, \sigma^2)$ (říkáme, že ε_t je bílý šum).

Odhad trendu časové řady pomocí klouzavých průměrů

Podstata klouzavých průměrů

Předpokládáme, že časová řada se řídí aditivním modelem

$y_t = f(t) + \varepsilon_t$, t = 1, ..., n.

Odhad trendu v bodě t získáme určitým zprůměrováním původních pozorování z jistého okolí uvažovaného časového okamžiku t. Můžeme si představit, že podél dané časové řady klouže okénko, v jehož rámci se průměruje. Nechť toto okénko zahrnuje d členů nalevo od bodu t a d členů napravo od bodu t. Hovoříme pak o vyhlazovacím okénku šířky h = 2d + 1. Prvních a posledních d hodnot trendu neodhadujeme, protože pro $t \in \{1, \dots, d\} \cup \{n-d+1, \dots, n\}$ není vyhlazovací okénko symetrické. Odhad trendu ve středu vyhlazovacího okénka je dán vztahem:

$$\hat{f}(t) = \frac{1}{2d+1}\left(y_{t-d} + y_{t-d+1} + \dots + y_{t+d}\right) = \frac{1}{2d+1}\sum_{k=0}^{2d} y_{t-d+k}, \quad t = d+1, \dots, n-d.$$

Šířka vyhlazovacího okénka

Velmi důležitou otázkou je stanovení šířky vyhlazovacího okénka. Je-li okénko příliš široké, bude se odhad trendu blížit přímce (říkáme, že je přehlazen) a zároveň se ztratí velký počet členů na začátku a na konci časové řady. Je-li naopak okénko úzké, bude se odhad trendu blížit původním hodnotám (říkáme, že odhad je podhlazen). Nejčastěji se volí šířka okénka h = 3, 5, 7, pro čtvrtletní hodnoty pak 4.

Příklad: Časová řada 215, 219, 222, 235, 202, 207, 187, 204, 174, 172, 201, 272 udává roční objemy vývozu piva (v miliónech litrů) z Československa v letech 1980 až 1991.

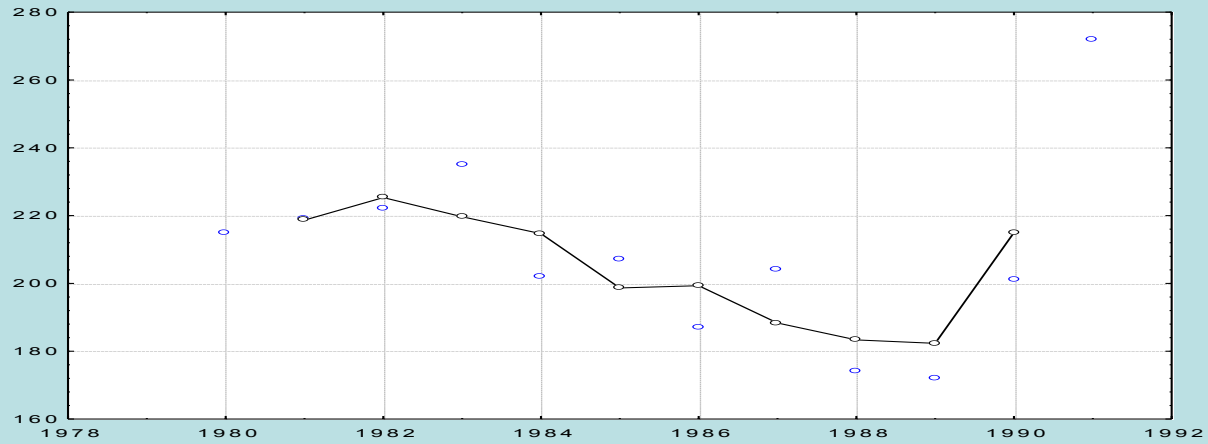
a) Odhadněte trend této časové řady pomocí klouzavých průměrů s vyhlazovacím okénkem šířky 3 a poté 5.
b) Graficky znázorněte průběh časové řady s odhadnutým trendem.

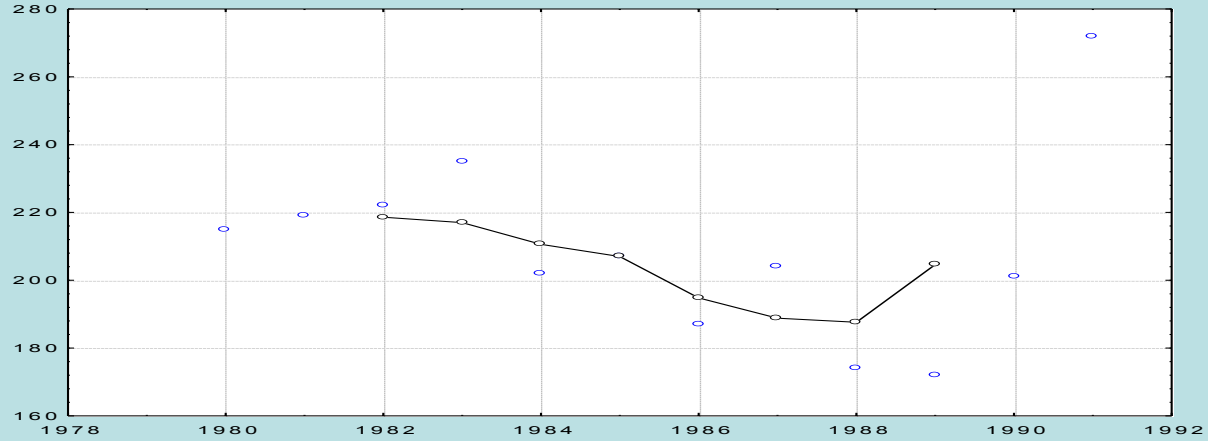
Řešení pomocí systému STATISTICA:

Vytvoříme datový soubor export_piva.sta o dvou proměnných ROK a VYVOZ a dvanácti případech.
Statistiky – Pokročilé lineární/nelineární modely – Časové řady/predikce – Proměnné Y – OK– OK (transformace, autokorelace, kříž. korelace, grafy) – Vyhlazování – zaškrtneme N-bod. klouzavý průměr, N = 3 – OK (Transformovat vybrané řady) – vykreslí se graf, vrátíme se do Transformace proměnných – Uložit proměnné. Otevře se nový spreadsheet, kde v proměnné VYVOZ_1 jsou uloženy klouzavé průměry pro N = 3. Totéž uděláme pro případ N = 5. Ve spreadsheetu se proměnná VYVOZ_1 přepíše na VYVOZ_2 a nová proměnná se uloží jako VYVOZ_1. Nově vzniklé proměnné nazveme KP3 a KP5. K datovému souboru přidáme proměnnou ROK, do jejíhož Dlouhého jména napíšeme =1979+v0.

	export_piva.sta			
	1 rok	2 VYVOZ	3 KP3	4 KP5
1	1980	215,000		
2	1981	219,000	218,667	
3	1982	222,000	225,333	218,600
4	1983	235,000	219,667	217,000
5	1984	202,000	214,667	210,600
6	1985	207,000	198,667	207,000
7	1986	187,000	199,333	194,800
8	1987	204,000	188,333	188,800
9	1988	174,000	183,333	187,600
10	1989	172,000	182,333	204,600
11	1990	201,000	215,000	
12	1991	272,000		

Grafické znázornění časové řady s odhadnutým trendem provedeme pomocí vícenásobných bodových grafů.





Cíl regresní analýzy trendu

Regresní analýza trendu má objasnit vztah mezi závisle proměnnou veličinou Y a časem t.

Předpokládáme, že trend f(t) závisí (lineárně či nelineárně) na neznámých parametrech $\beta_0, \beta_1, ..., \beta_k$ a známých funkcích $\varphi_0(t), \varphi_1(t), ..., \varphi_k(t)$, které již neobsahují žádné neznámé parametry, tj.

$f(t) = g(\beta_0, \beta_1, ..., \beta_k; \varphi_0(t), \varphi_1(t), ..., \varphi_k(t))$.

Odhady $b_0, b_1, ..., b_k$ neznámých parametrů $\beta_0, \beta_1, ..., \beta_k$ lze získat např. metodou nejmenších čtverců a pak vyjádřit odhad $\hat{f}(t)$ neznámého trendu v bodě t pomocí odhadů $b_0, b_1, ..., b_k$ a funkcí $\varphi_0(t), \varphi_1(t), ..., \varphi_k(t)$, tj.

$\hat{f}(t) = g(b_0, b_1, ..., b_k; \varphi_0(t), \varphi_1(t), ..., \varphi_k(t))$.

Nejdůležitější typy trendových funkcí

Volba typu trendové funkce se provádí

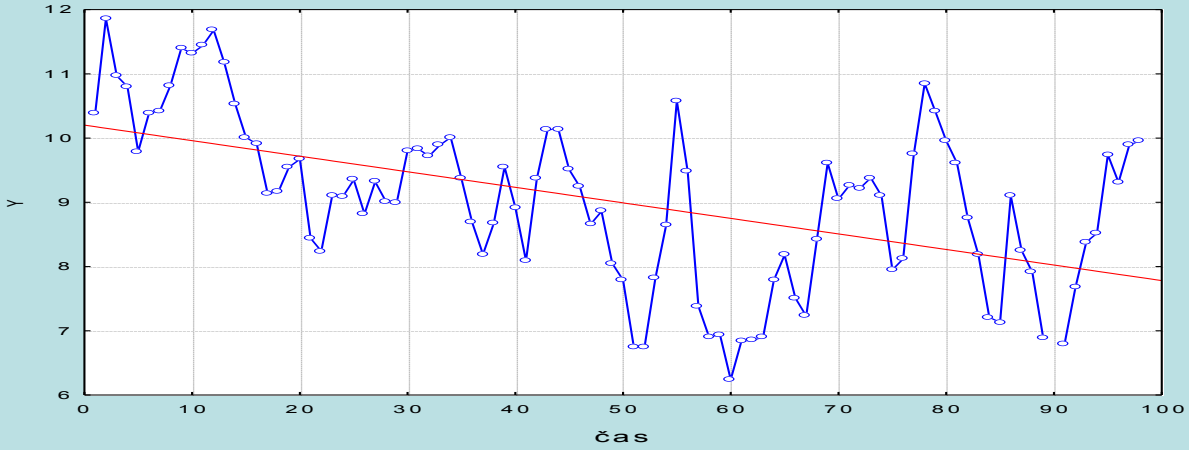
- na základě teoretických znalostí a zkušeností se zkoumanou veličinou Y_t
- pomocí grafu časové řady
- pomocí informativních testů založených na jednoduchých charakteristikách časové řady

a) Lineární trend

Analytické vyjádření: $f(t) = \beta_0 + \beta_1 t$

Informativní test: 1. diference ($\Delta y_t = y_t - y_{t-1}, t = 2, \ldots, n$) jsou přibližně konstantní.

Příklad lineárního trendu:

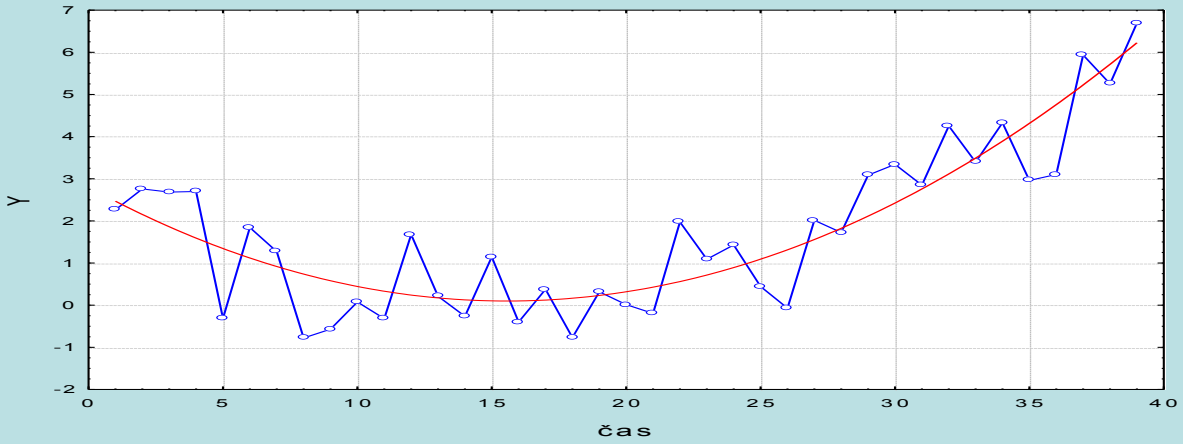


b) Kvadratický trend

Analytické vyjádření: $f(t) = \beta_0 + \beta_1 t + \beta_2 t^2$

Informativní test: 1. diference mají přibližně lineární trend, 2. diference ($\Delta^2 y_t = \Delta y_t - \Delta y_{t-1} = y_t - 2y_{t-1} + y_{t-2}, t = 3, \ldots, n$) jsou přibližně konstantní.

Příklad kvadratického trendu:

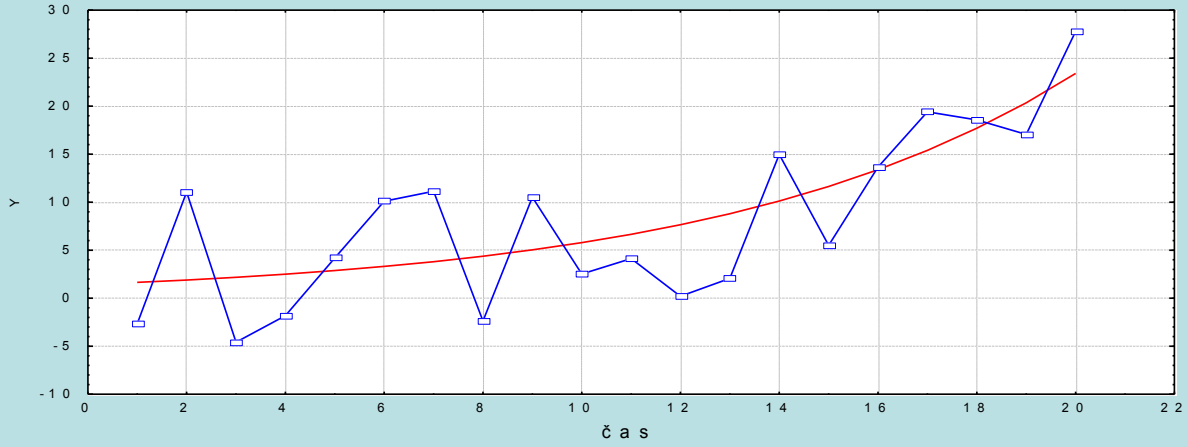


c) Exponenciální trend

Analytické vyjádření: $f(t) = \beta_0 \beta_1^t$.

Informativní test: koeficienty růstu ($k_t = \frac{y_t}{y_{t-1}}, t = 2, \ldots, n$) jsou přibližně konstantní.

Příklad exponenciálního trendu:

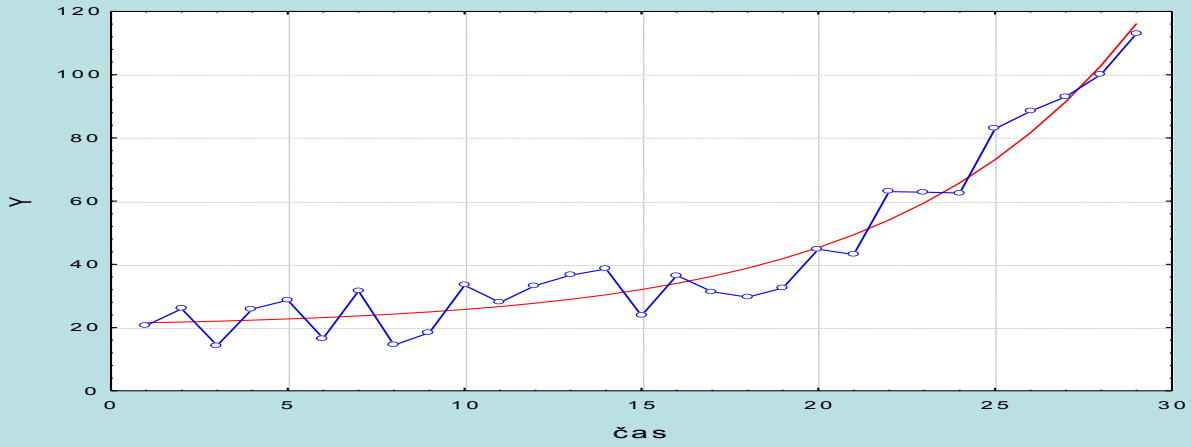


d) Modifikovaný exponenciální trend

Analytické vyjádření: $f(t) = \alpha + \beta_0 \beta_1^t$.

Informativní test: řada podílů sousedních 1. diferencí je přibližně konstantní.

Příklad modifikovaného exponenciálního trendu

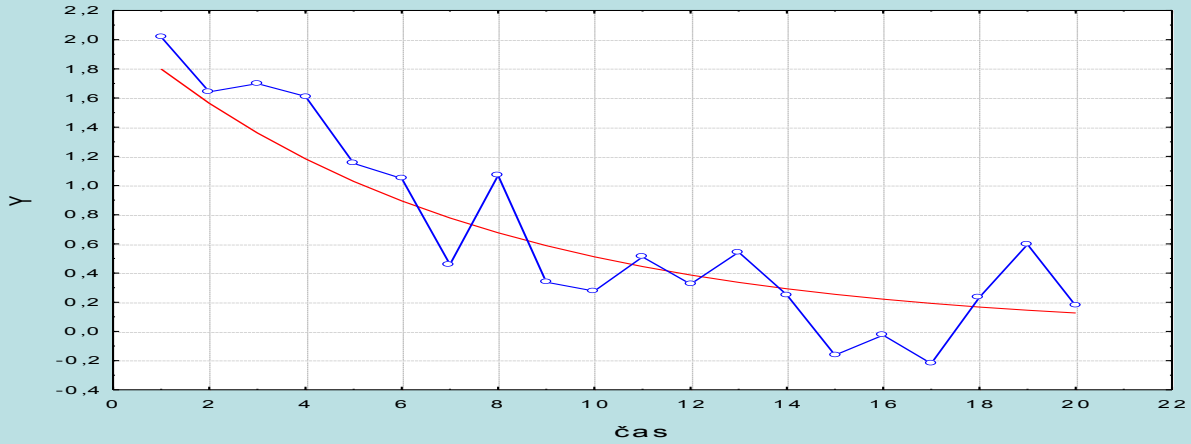


e) Logistický trend

Analytické vyjádření: $f(t) = \frac{\alpha}{1 + 3_{0} \beta_{1}^{t}}$

Informativní test: průběh 1. diferencí je podobný Gaussově křivce a podíly $\frac{1/y_{t+} - /y_{t+}}{1/y_{t+} - /y_{t}}$ jsou přibližně konstantní.

Příklad logistického trendu:

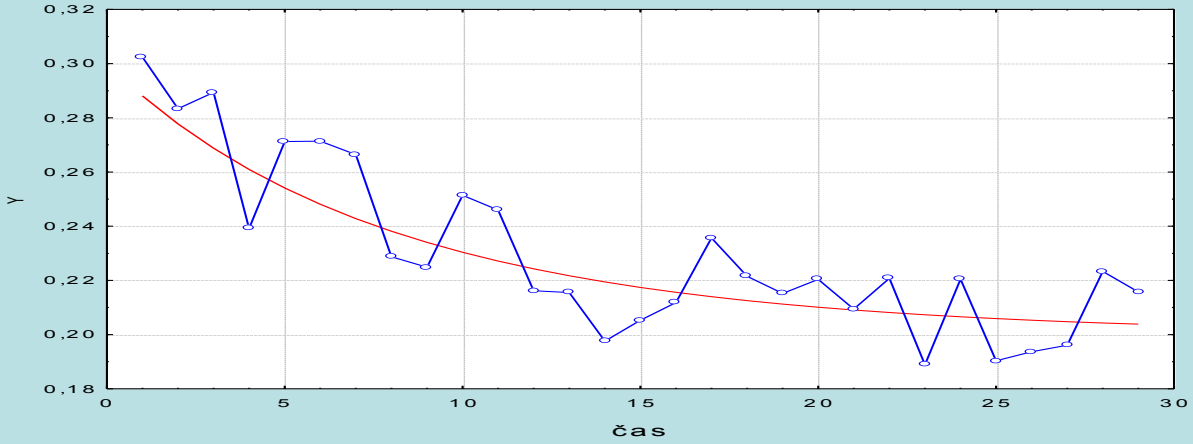


f) Gompertzova křivka

Analytické vyjádření: $f(t) = \alpha\beta_0^{\beta_1^t}$

Informativní test: podíly $\frac{\ln y_{t+2} - \ln y_{t+1}}{\ln y_{t+1} - \ln y_t}$ jsou přibližně konstantní.

Příklad Gompertzovy křivky



Modely (a), (b), (c) jsou lineární nebo se dají linearizovat a odhady parametrů získáme metodou nejmenších čtverců.
Modely (d), (e), (f) jsou nelineární a odhady parametrů se získávají speciálními numerickými metodami.

Orientační ověřování kvality modelu

- Index determinace (tj. podíl vysvětlené a celkové variability závisle proměnné veličiny) by měl být blízký 1.
- Body grafu $[f(t), \hat{f}(t)]$, t = 1, 2, ..., n by se měly řadit do přímky se směrnicí 1.

Příklad: Časová řada 112, 149, 238, 354, 580, 867 udává zisk (v tisících dolarů) jisté společnosti v prvních šesti letech její existence.

a) Graficky znázorněte průběh této časové řady.

b) Vypočtěte koeficienty růstu

c) Z grafu časové řady a chování koeficientů růstu lze usoudit, že časová řada má exponenciální trend $f(t) = \beta_0 \beta_1^t$.

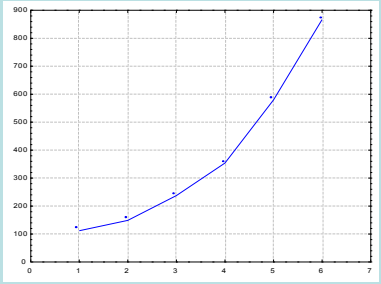
Odhadněte jeho parametry.

d) Najděte odhad zisku společnosti v 7. a 8. roce její existence.

e) Zjistěte index determinace a sestrojte graf $(f(t), \hat{f}(t))$, t = 1, ..., 6.

Řešení: Znovu uveďme hodnoty časové řady: 112, 149, 238, 354, 580, 867

ad a)

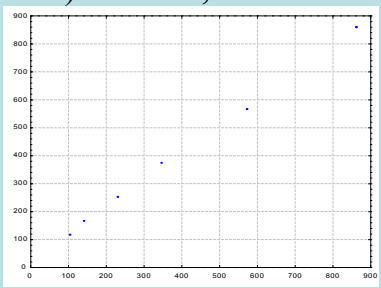


ad b) Koeficienty růstu: 149/112 = 1,33, 238/149 = 1,597, 354/238 = 1,487, 580/354 = 1,628, 867/580 = 1,495. Vidíme, že koeficienty růstu jsou přibližně konstantní.

ad c) Model $f(t) = \beta_0 \beta_1^t$ linearizujeme a metodou nejmenších čtverců získáme odhady $\ln b_0 = 4,227983$, $\ln b_1 = 0,420199$. Odlogaritmováním dostaneme $b_0 = 68,57875$, $b_1 = 1,522265$.

ad d) $\hat{y}_7 = 68,57875 \cdot 1,522265^{7} = 1299$, $\hat{y}_8 = 68,57875 \cdot 1,522265^{8} = 1977$

ad e) $ID^2 = 0,996$



Jak index determinace, tak graf $(t, \hat{f}(t))$ svědčí o tom, že model byl zvolen správně.

Výpočet pomocí systému STATISTICA:
Vytvoříme datový soubor se dvěma proměnnými čas a Y a 6 případy.

ad a) Časovou řadu znázorníme graficky pomocí Grafy – Bodové grafy.

ad b) Koeficienty růstu získáme pomocí Statistiky – Pokročilé lineární/nelineární modely – Časové řady/predikce.

ad c) K datovému souboru přidáme novou proměnnou ln Y, kterou získáme zlogaritmováním proměnné Y, v níž jsou uloženy hodnoty zisku společnosti. Provedeme regresní analýzu se závisle proměnnou ln Y a nezávisle proměnnou čas. K výstupní tabulce přidáme novou proměnnou, do jejíhož Dlouhého jména napíšeme =exp(b)

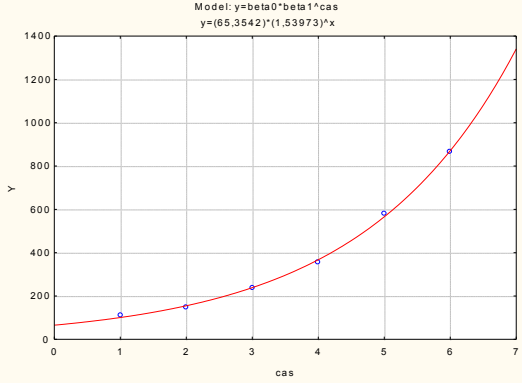
	Výsledky regrese se závislou proměnnou : lnY (zisk_spolecnosti.sta) R= ,99801042 R2= ,99602479 Upravené R2= ,99503099 F(1,4)=1002,2 p<,00001 Směrod. chyba odhadu : ,05553						
N=6	b*	Sm.chyba z b*	b	Sm.chyba z b	t(4)	p-hodn.	NProm =exp(b)
Abs.člen			4,227983	0,051691	81,79336	0,000000	68,57875
cas	0,998010	0,031525	0,420199	0,013273	31,65812	0,000006	1,522265

Vidíme, že $Y = 68{,}57875 \cdot 1{,}522265^{t}$.

ad d) Pro výpočet predikovaného zisku v 7. a 8. roce existence společnosti použijeme STATISTIKU jako kalkulačku.

ad e) Index determinace najdeme ve výstupní tabulce regrese pod označením R2. V našem případě je 0,996.
Pro získání grafu závislosti predikovaných hodnot na naměřených hodnotách přidám ek datovému souboru proměnnou predikce a do jejího Dlouhého jména napíšeme =68,57875*1,52265^cas. Pak vytvoříme Bodový graf.

Můžeme též nakreslit dvourozměrný tečkový diagram s odhadnutou regresní křivkou:
Na liště Detaily vybereme Proložení Exponenciální.



Porovnání empirického a teoretického rozložení

Motivace: Možnost použití statistických testů je podmíněna nějakými předpoklady o datech. Velmi často je to předpoklad o typu rozložení, z něhož získaná data pocházejí. Mnoho testů je založeno na předpokladu normality. (Testování normality bylo probráno ve 2. kapitole.) Opomíjení předpokladů o typu rozložení může v praxi vést i ke zcela zavádějícím výsledkům, proto je nutné věnovat tomuto problému patřičnou pozornost.

V této kapitole se seznámíme s testem dobré shody, který je (po splnění určitých předpokladů) použitelný k ověření shody empirického rozložení s jakýmkoliv teoretickým rozložením.

Testy dobré shody

Popis testu

Testujeme hypotézu, která tvrdí, že náhodný výběr $X_1, ..., X_n$ pochází z rozložení s distribuční funkcí $\Phi(x)$.

Spojitý případ:

- data rozdělíme do r třídicích intervalů $\langle u_j, u_{j+1} \rangle$, $j = 1, ..., r$
- zjistíme absolutní četnost n_j j-tého třídicího intervalu
- vypočteme pravděpodobnost p_j, že náhodná veličina X s distribuční funkcí $\Phi(x)$ se bude realizovat v j-tém třídicím intervalu. Platí-li nulová hypotéza, pak $p_j = \Phi(u_{j+1}) - \Phi(u_j)$.

Diskrétní případ:

- určíme varianty $x_{[j]}$, $j = 1, \ldots, r$
- pro variantu $x_{[j]}$ zjistíme absolutní četnost n_j
- vypočteme pravděpodobnost p_j, že náhodná veličina X s distribuční funkcí $\Phi(x)$ se bude realizovat variantou $x_{[j]}$. Platí-li nulová hypotéza, pak $p_j = \Phi(x_{[j]}) - \lim_{x \to x_{[j]}^-} \Phi(x) = P(X = x_{[j]})$.

Testová statistika: $K = \sum_{j=1}^{r} \frac{(n_j - np_j)^2}{np_j}$. Platí-li nulová hypotéza, pak $K \approx \chi^2(r-1-p)$, kde p je počet odhadovaných parametrů daného rozložení. (Např. pro normální rozložení p = 2, protože z dat odhadujeme střední hodnotu a rozptyl.) Pokud žádný parametr nemusíme odhadovat, hovoříme o úplně specifikovaném problému. Nulovou hypotézu zamítáme na asymptotické hladině významnosti α, když $K \geq \chi^2_{1-\alpha}(r-1-p)$. Aproximace se považuje za vyhovující, když $np_j \geq 5$, $j = 1, ..., r$.

Upozornění: Při nesplnění podmínky $np_j \geq 5$, $j = 1, ..., r$ je třeba některé intervaly resp. varianty slučovat, což vede ke ztrátě informace. Ve spojitém případě je hodnota testové statistiky K silně závislá na volbě třídicích intervalů

Příklad: (Testování shody empirického a teoretického rozložení při neúplně specifikovaném problému – diskrétní případ)
V tabulce jsou roztříděny fotbalové zápasy určité soutěže podle počtu vstřelených branek.

Počet branek	0	1	2	3	4 a víc
Počet zápasů	19	30	17	10	8

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že jde o výběr z Poissonova rozložení.

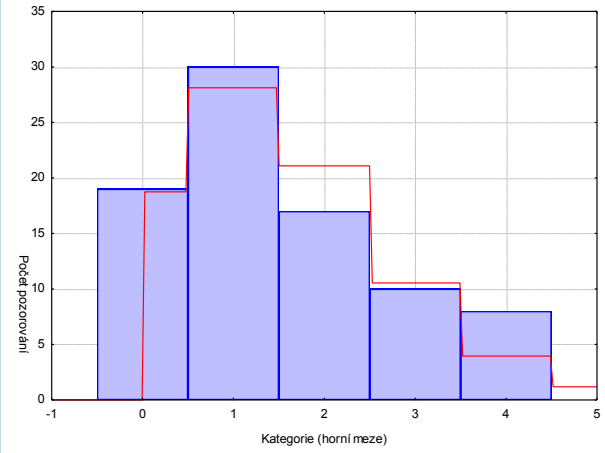
Výpočet pomocí systému STATISTICA:
Vytvoříme datový soubor s dvěma proměnnými a 5 případy. Proměnná POCET obsahuje počet vstřelených branek, proměnná CETNOST pak počet zápasů, v nichž bylo dosaženo zjištěného počtu branek.
Statistiky – Prokládání rozdělení – Diskrétní rozdělení – Poissonovo – OK – Proměnná POCET – klikneme na ikonu se závažím – Proměnná vah CETNOST – Stav Zapnuto – OK – Výpočet.

Proměnná: POCET , Rozdělení:Poissonovo, Lambda = 1,500 (branky.sta)
Chí-kvadrát = 2,07051, sv = 3, p = 0,55790

Kategorie	Pozorované Četnosti	Kumulativ. Pozorované	Procent Pozorované	Kumul. % Pozorované	Očekáv. Četnosti	Kumulativ. Očekáv.	Procent Očekáv.	Kumul. % Očekáv.
<= 0,00000	19	19	22,61905	22,6190	18,74294	18,74294	22,31302	22,3130
1,00000	30	49	35,71429	58,3333	28,11440	46,85733	33,46952	55,7825
2,00000	17	66	20,23810	78,5714	21,08580	67,94313	25,10214	80,8847
3,00000	10	76	11,90476	90,4762	10,54290	78,48603	12,55107	93,4358
< Nekonečno	8	84	9,52381	100,0000	5,51397	84,00000	6,56424	100,0000

V tomto případě je parametr λ Poissonova rozložení neznámý, je odhadnut pomocí výběrového průměru a odhad činí 1,5. Podmínky dobré aproximace jsou splněny, dokonce všechny teoretické četnosti jsou větší než 5. Dále je v záhlaví výstupní tabulky uvedena hodnota testového kritéria (2,07051), počet stupňů volnosti $r - p - 1 = 5 - 1 - 1 = 3$ a p-hodnota (0,5578). Nulová hypotéza se tedy nezamítá na asymptotické hladině významnosti 0,05.

Pro vytvoření grafu se vrátíme do Proložení diskrétních rozložení – Základní výsledky – Graf pozorovaného a očekávaného rozdělení.



Příklad: (Testování shody empirického a teoretického rozložení při neúplně specifikovaném problému – spojitý případ)
U 48 studentek VŠE v Praze byla zjišťována výška (v cm): 165 170 170 179 170 168 174 162 167 165 170 173 183 176 165 168 171 178 168 168 169 163 172 184 176 175 176 169 168 170 166 160 167162 162 166 170 168 155 162 169 166 160 169 165 163 168 163
Pomocí testu dobré shody testujte na hladině významnosti 0,05 hypotézu, že data pocházejí z normálního rozložení. Pomocí histogramu posuďte vizuálně předpoklad normality.

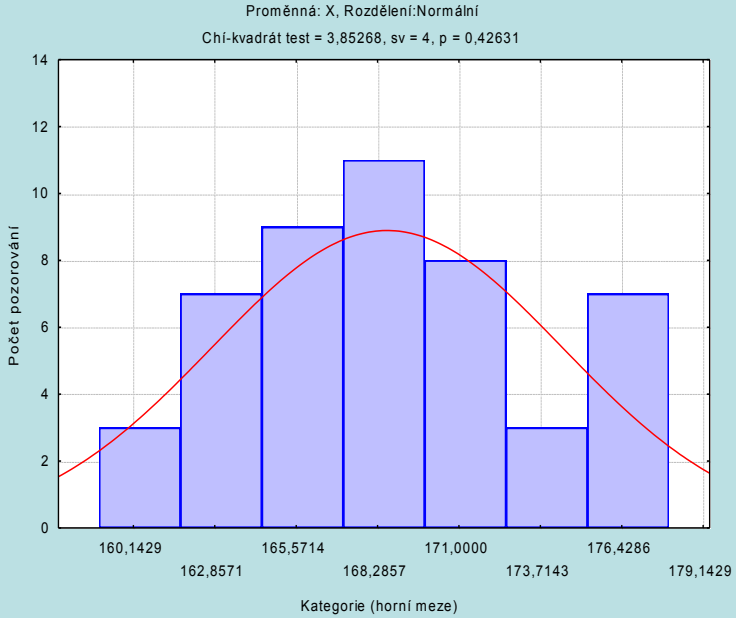
Výpočet pomocí systému STATISTICA:
Statistiky - Prokládání rozdělení – ponecháme implicitní nastavení na normální rozložení – OK – Proměnná X – OK – na záložce Parametry změníme Počet kategorií na 7 (podle Sturgesova pravidla) – Výpočet.

	Proměnná: X, Rozdělení: Normální (vyska.sta) Chí-kvadrát = 1,09280, sv = 1 (uprav.), p = 0,29585							
Horní hranice	Pozorované Četnosti	Kumulativ. Pozorované	Procent Pozorované	Kumul. % Pozorované	Očekáv. Četnosti	Kumulativ. Očekáv.	Procent Očekáv.	Kumul. % Očekáv.
<= 157,14286	1	1	2,08333	2,0833	1,19706	1,19706	2,49387	2,4939
162,28571	6	7	12,50000	14,5833	5,51484	6,71189	11,48924	13,9831
167,42857	12	19	25,00000	39,5833	13,46220	20,17409	28,04624	42,0293
172,57143	19	38	39,58333	79,1667	15,89146	36,06555	33,10721	75,1366
177,71429	6	44	12,50000	91,6667	9,07700	45,14255	18,91042	94,0470
182,85714	2	46	4,16667	95,8333	2,50365	47,64620	5,21594	99,2629
< Nekonečno	2	48	4,16667	100,0000	0,35380	48,00000	0,73708	100,0000

Při tomto roztřídění dat do 7 intervalů nejsou splněny podmínky dobré aproximace, ve třech intervalech jsou teoretické četnosti pod 5. Změníme tedy dolní mez na 159 a horní na 178.

Horní hranice	Proměnná: X, Rozdělení:Normální (vyska.sta) Chí-kvadrát = 3,85268, sv = 4, p = 0,42631 Pozorované Četnosti	Kumulativ. Pozorované	Procent Pozorované	Kumul. % Pozorované	Očekáv. Četnosti	Kumulativ. Očekáv.	Procent Očekáv.	Kumul. % Očekáv.
<= 161,71429	3	3	6,25000	6,2500	5,722996	5,72300	11,92291	11,9229
164,42857	7	10	14,58333	20,8333	5,675946	11,39894	11,82489	23,7478
167,14286	9	19	18,75000	39,5833	7,862633	19,26157	16,38048	40,1283
169,85714	11	30	22,91667	62,5000	8,812455	28,07403	18,35928	58,4876
172,57143	8	38	16,66667	79,1667	7,991516	36,06555	16,64899	75,1366
175,28571	3	41	6,25000	85,4167	5,863558	41,92910	12,21575	87,3523
< Nekonečno	7	48	14,58333	100,0000	6,070896	48,00000	12,64770	100,0000

V tomto případě jsou podmínky dobré aproximace splněny. Testová statistika se realizuje hodnotou 3,85268, p-hodnota je 0,42631, tedy na asymptotické hladině významnosti 0,05 hypotézu o normalitě nezamítáme. Podívejme se ještě na histogram s proloženou Gaussovou křivkou: Na záložce Základní výsledky zvolíme Graf pozorovaného a očekávaného rozdělení.



Upozornění: Test dobré shody může být použit i v těch případech, kdy rozložení, z něhož daný náhodný výběr pochází, neodpovídá nějakému známému rozložení (např. exponenciálnímu, normálnímu, Poissonovu, ...), ale je určeno intuitivně nebo na základě zkušenosti.

Příklad: Ve svých pokusech pozoroval J.G. Mendel 10 rostlin hrachu a na každé z nich počet žlutých a zelených semen. Výsledky pokusu:

č.rostliny	1	2	3	4	5	6	7	8	9	10
počet žlutých	25	32	14	70	24	20	32	44	50	44
počet zelených	11	7	5	27	13	6	13	9	14	18
celkem	36	39	19	97	37	26	45	53	64	62

Z genetických modelů vyplývá, že pravděpodobnost výskytu žlutého semene by měla být 0,75 a zeleného 0,25. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že výsledky Mendelových pokusů se shodují s modelem.

Řešení:
Výpočty potřebné pro stanovení testové statistiky K uspořádáme do tabulky.

j	n_j	p_j	np_j	$\frac{(n_j - np_j)^2}{np_j}$
1	25	0,75	36.0,75=27	0,148148
2	32	0,75	39.0,75=29,25	0,258547
⋮	⋮	⋮	⋮	⋮
10	44	0,75	62.0,75=46,5	0,134409

K = 0,148148 + 0,258547 + ... + 0,134409 = 1,797495, r = 10, $\chi^2_{0,95}(9) = 16,9$.
Protože 1,797495 < 16,9, nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.

Výpočet pomocí systému STATISTICA:

Vytvoříme datový soubor se třemi proměnnými celkem, X, Y a 10 případy. Do proměnné celkem zapíšeme celkový počet semen, do X počet žlutých semen, do proměnné Y vypočítané teoretické četnosti (tj. celkem*0,75)
Statistiky – Neparametrická statistika – Pozorované vs. Očekávané – OK – Proměnné – Pozorované X, Očekávané Y – OK – Výpočet.

Pozorované vs. očekávané četnosti (Mendel hrach.sta)
Chi-Kvadr. = 1,797495 sv = 9 p = ,994280
POZN.: Nestejné součty pozor. a ček. četností

Případ	pozorov. X	očekáv. Y	P - O	(P-O)^2 /O
C: 1	25,0000	27,0000	-2,00000	0,148148
C: 2	32,0000	29,2500	2,75000	0,258547
C: 3	14,0000	14,2500	-0,25000	0,004386
C: 4	70,0000	72,7500	-2,75000	0,103952
C: 5	24,0000	27,7500	-3,75000	0,506757
C: 6	20,0000	19,5000	0,50000	0,012821
C: 7	32,0000	33,7500	-1,75000	0,090741
C: 8	44,0000	39,7500	4,25000	0,454403
C: 9	50,0000	48,0000	2,00000	0,083333
C: 10	44,0000	46,5000	-2,50000	0,134409
Sčt	355,0000	358,5000	-3,50000	1,797495

Ve výstupní tabulce najdeme hodnotu testové statistiky (Chi-Kvadr. = 1,797495) a odpovídající p-hodnotu = 0,99428, kterou porovnáme se zvolenou hladinou významnosti. V našem případě je p-hodnota větší než 0,05, takže nulová hypotéza se nezamítá na asymptotické hladině významnosti 0,05. Neprokázali jsme rozpor mezi skutečností a Mendelovým modelem.