Téma č. 12: Úvod do analýzy časových řad

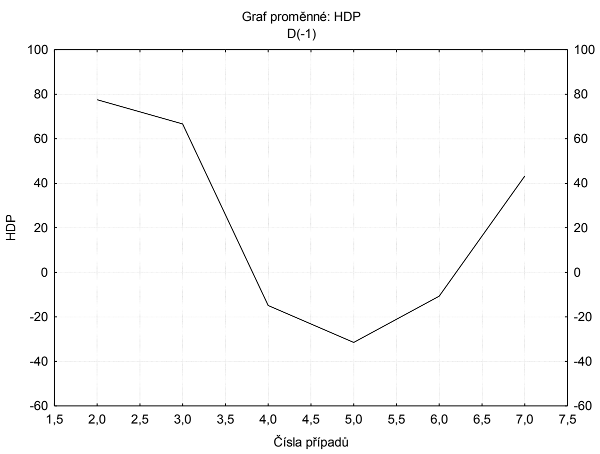
Příklad 1.: Časová řada vyjadřuje počet obyvatelstva ČSSR (v tisících) v letech 1965 až 1974 vždy ke dni 31.12.

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: Načteme datový soubor obyvatele_CSSR.sta o 11 proměnných a jednom případu. Do Dlouhého jména poslední proměnné napíšeme
=(v1/2+sum(v2:v9)+v10/2)/9
Dostaneme výsledek 14430,11.

Příklad 2.: Pro časovou řadu HDP ČR v letech 1994 až 2000 (v miliardách Kč) vypočtěte základní charakteristiky dynamiky a graficky znázorněte relativní přírůstky a koeficienty růstu.
Řešení: Načteme datový soubor HDP. sta.
Výpočet 1. diferencí: $\Delta_i = y_i - y_{i-}$ pro i = 2,...,n
Statistiky – Pokročilé lineární/nelineární modely – Časové řady/predikce – Proměnné Y – OK – OK (transformace, autokorelace, kříž. korelace, grafy) – Oddělit-sloučit - 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: $\delta_i = \dfrac{\Delta_i}{y_{i-1}}$ 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: $k_i = \dfrac{y_i}{y_{i-1}}$ pro i = 2,...,n

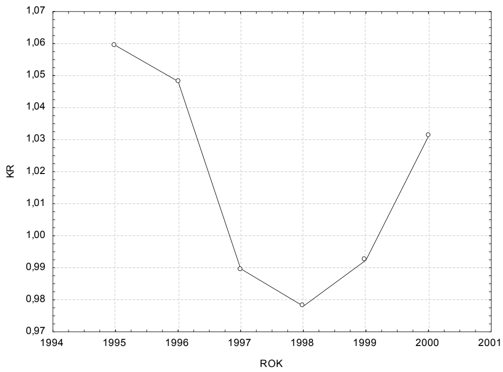
Do tabulky přidáme proměnnou KR a do jejího Dlouhého jména napíšeme vzorec =HDP/HDP_1.
Získáme tabulku

	HDP	HDP_2	HDP_1	RP	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	-0,010292	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.

Graf relativních přírůstků

Graf koeficientů růstu



Průměrný absolutní přírůstek a průměrný koeficient růstu vypočteme na kalkulačce pomocí vzorců $\bar{\Delta} = \dfrac{1433{,}8 - 1303{,}6}{6} = 21{,}7$ a $\bar{k} = \sqrt[6]{\dfrac{1433{,}8}{1303{,}6}} = 1{,}016$.

Příklad 3.: Je dána časová řada potratů (v tisících) v ČR v letech 1986 až 1996: 99,5 126,7 129,3 126,5 126,1 120,1 109,3 85,4 67,4 61,6 60.

Předpokládejte, že tato časová řada má kvadratický trend. Odhadněte parametry trendové funkce.

Vypočtěte index determinace ID^2.

Proveďte celkový F-test. (Popis celkového F- testu: Na hladině významnosti α testujeme H_0: $(\beta_1,\ldots,\beta_p)' = (0,\ldots,0)'$ proti H_1: $(\beta_1,\ldots,\beta_p)' \neq (0,\ldots,0)'$, přičemž p je počet odhadovaných regresních parametrů (bez parametru β_0)
(Nulová hypotéza říká, že dostačující je model konstanty.)

Testová statistika $F = \dfrac{S_R/p}{S_E/(n-p-1)}$ má rozložení F(p, n-p-1), pokud H_0 platí. Přitom

$S_E = \sum_{t=1}^{n} (y_t - \hat{y}_t)^2$ je reziduální součet čtverců a $S_R = \sum_{t=1}^{n} (\hat{y}_t - \bar{y})^2$ je regresní součet čtverců,

kde $\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t$.

Kritický obor: $W = \langle F_{1-\alpha}(p, n-p-1), \infty)$.

$F \in W \Rightarrow$ H_0 zamítáme na hladině významnosti α.

Proveďte dílčí t-testy. (Popis dílčích t-testů: Na hladině významnosti α pro j = 0,1, ..., p testujeme hypotézu
H_0: $\beta_j = 0$ proti H_1: $\beta_j \neq 0$.

Testová statistika: $T_j = \dfrac{b_j}{s_{b_j}}$ má rozložení t(n-p-1), pokud H_0 platí. Přitom s_{b_j} je směrodatná chyba odhadu b_j.

Kritický obor: $W = (-\infty, -t_{1-\alpha/2}(n-p-1)\rangle \cup \langle t_{1-\alpha/2}(n-p-1), \infty)$.

$T_j \in W \Rightarrow$ H_0 zamítáme na hladině významnosti α.)

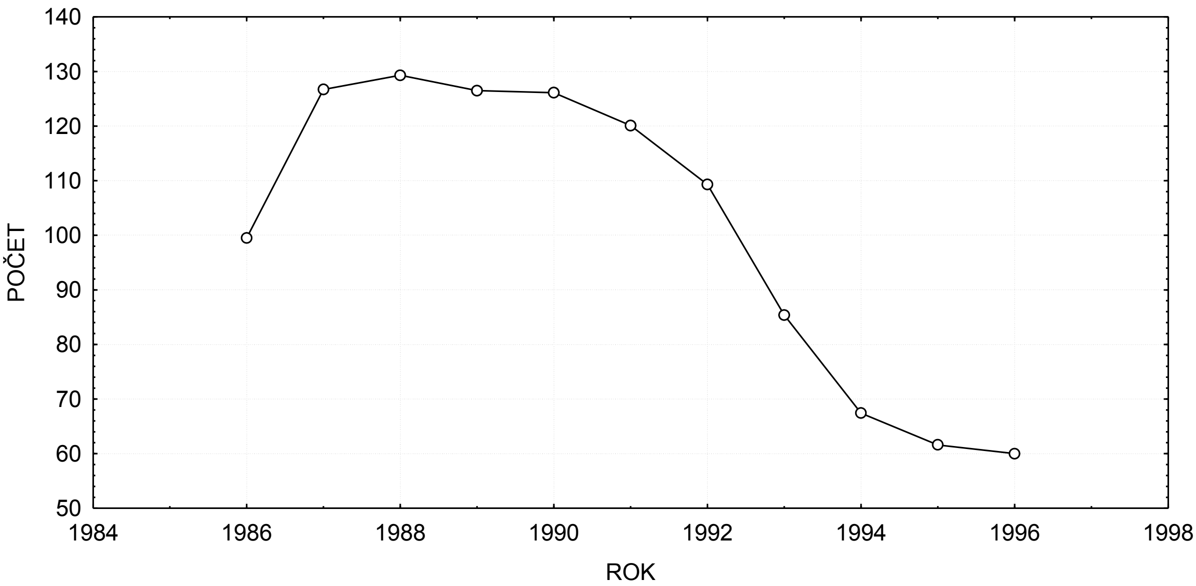
Ověřte normalitu reziduí.

Sestrojte 95% intervaly spolehlivosti pro parametry trendové funkce. (Vzorec pro meze 100(1- α)% intervalu spolehlivosti pro β_j: $b_j \pm t_{1-\alpha/2}(n-p-1)s_{b_j}$)

Stanovte střední absolutní procentuální chybu predikce (MAPE). MAPE se počítá podle vzorce $MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|$.

Graficky znázorněte průběh časové řady s odhadnutým trendem, 95% pásem spolehlivosti a 95% predikčním pásem.

Řešení:
Načteme datový soubor potraty.sta. Pro lepší orientaci znázorníme časovou řadu graficky.
Grafy – Bodové grafy – Proměnné X ROK, Y POCET – OK – vypneme Lineární proložení– OK.
Formát – Všechny možnosti – Graf: Obecné – zaškrtneme Spojnice – OK. Vznikne spojnicový diagram.



Trendová funkce $\hat{f}(t) = \beta_0 + \beta_1 t + \beta_2 t^2$

Odhady parametrů:

Statistiky – Vícenásobná regrese – Proměnné Závislé, Nezávislé t, tkv - OK

N=11	Výsledky regrese se závislou proměnnou : POCET (potraty.sta) R= ,94015284 R2= ,88388736 Upravené R2= ,85485920 F(2,8)=30,449 p<,00018 Směrod. chyba odhadu : 10,629					
	Beta	Sm.chyba beta	B	Sm.chyba B	t(8)	Úroveň p
Abs.člen			103,2418	11,67235	8,84499	0,000021
t	1,30140	0,531476	10,9470	4,47060	2,44866	0,040020
tkv	-2,16020	0,531476	-1,4748	0,36285	-4,06453	0,003611

Odhadnutá trendová funkce má tedy tvar:

$\hat{f}(t) = 103{,}2418 + 10{,}947\,t - 1{,}4748\,t^2$, kde t = 1, …, 11.

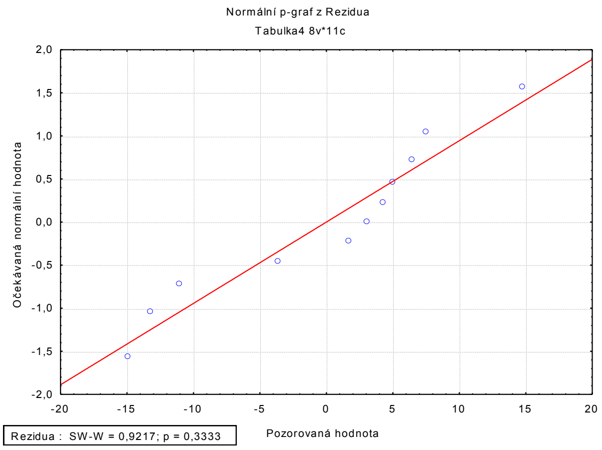
Index determinace je 0,884, tedy kvadratická trendová funkce vysvětluje variabilitu dané časové řady z 88,4%.

Testová statistika celkového F-testu je 30,449, p-hodnota je blízká 0, tedy na hladině významnosti 0,05 zamítáme hypotézu o nevýznamnosti modelu jako celku.

Všechny tři dílčí t-testy mají p-hodnoty menší než 0,05, tedy na hladině významnosti 0,05 zamítáme hypotézy o nulovosti parametrů β_0, β_1, β_2.

Ověření normality reziduí:

Na záložce Rezidua/předpoklady/předpovědi zvolíme Reziduální analýza – Uložit – Uložit rezidua & předpovědi. Sestrojíme N-P plot reziduí a současně provedeme S-W test:



S-W test poskytuje p-hodnotu 0,333, tedy na hladině významnosti 0,05 nezamítáme hypotézu o normalitě reziduí.

Sestrojení 95% intervalů spolehlivosti pro parametry trendu:

Ve výstupní tabulce výsledků regrese přidáme za proměnnou Úroveň p dvě nové proměnné dm (pro dolní meze 95% intervalů spolehlivosti) a hm (pro horní meze 95% intervalů spolehlivosti). Do Dlouhého jména proměnné dm resp. hm napíšeme:
=v3-v4*VStudent(0,975;8) resp. =v3+v4*VStudent(0,975;8)

N=11	Výsledky regrese se závislou proměnnou : POCET (potraty.sta) R= ,94015284 R2= ,88388736 Upravené R2= ,85485920 F(2,8)=30,449 p<,00018 Směrod. chyba odhadu : 10,629							
	Beta	Sm.chyba beta	B	Sm.chyba B	t(8)	Úroveň p	dm =v3-v4*	hm =v3+v4
Abs.člen			103,2418	11,67235	8,84499	0,000021	76,32533	130,1583
t	1,30140	0,531476	10,9470	4,47060	2,44866	0,040020	0,637767	21,25622
tkv	-2,16020	0,531476	-1,4748	0,36285	-4,06453	0,003611	-2,31156	-0,63809

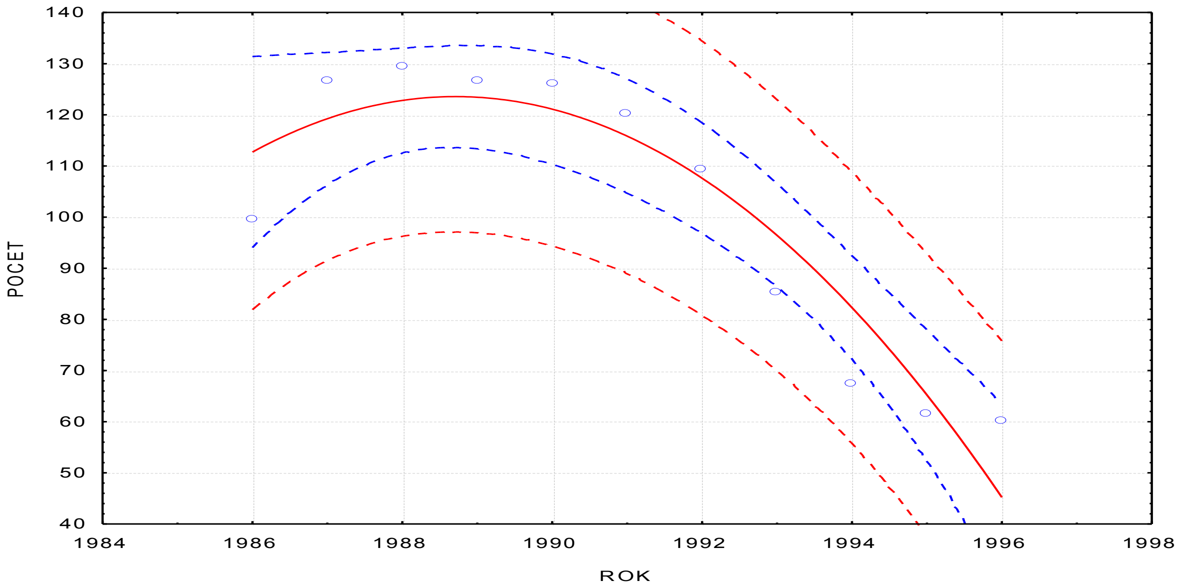
Vidíme, že $76{,}32 < \beta_0 < 130{,}16$ s pravděpodobností aspoň 0,95, $0{,}64 < \beta_1 < 21{,}26$ a $-2{,}31 < \beta_2 < -0{,}64$ s pravděpodobností aspoň 0,95.

Výpočet MAPE:

Ve výsledcích Vícenásobné regrese zvolíme záložku Rezidua / předpoklady / předpovědi – Reziduální analýza – Uložit – Uložit rezidua a předpovědi – Vybrat vše – OK. Ve vzniklé tabulce odstraníme proměnné 7 – 12, přidáme proměnnou chyby a do jejího Dlouhého jména napíšeme =100*abs(v6/v2). Pak spočteme průměr této proměnné a zjistíme, že MAPE = 9,21%.

Graf časové řady s proloženým kvadratickým trendem získáme takto:

Grafy – Bodové grafy – Proměnné X ROK, Y POCET – OK – Detaily Proložení Polynomiální. Ve vytvořeném grafu 2x klikneme na pozadí, vybereme Graf: Regresní pásy – Přidat nový pár pásů – Typ Spolehlivostní – OK. Totéž provedeme ještě jednou a nyní zaškrtneme Typ Predikční.

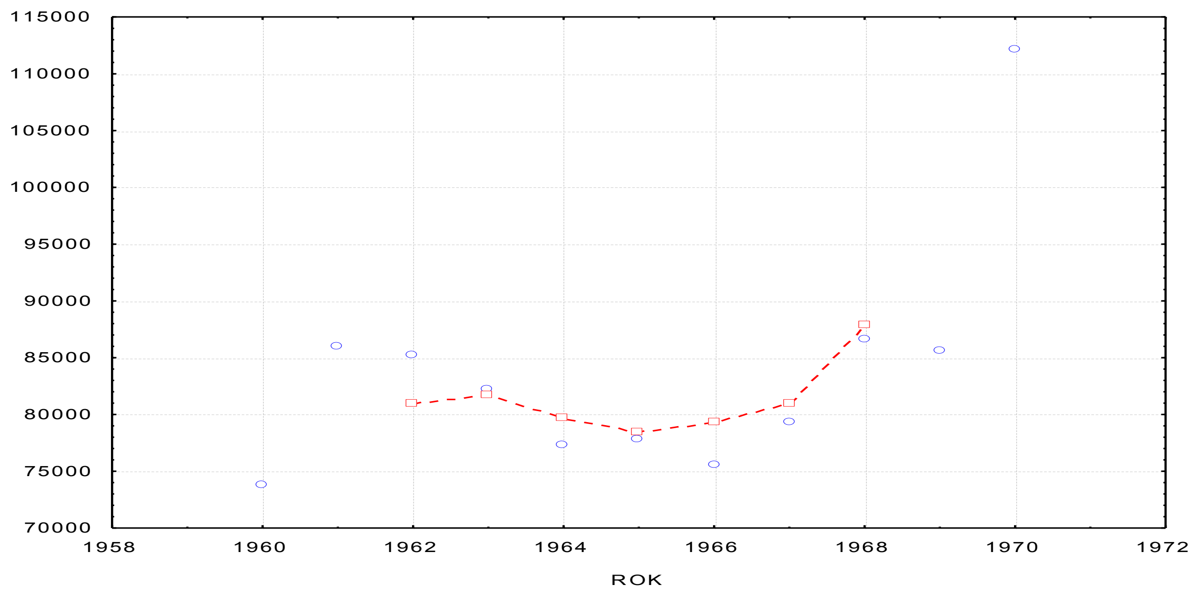


Příklad 4.: Máme k dispozici údaje o počtu bytů předaných do užívání v Československu v letech 1960 až 1970: 73 766   86 032   85 221   82 189   77 301   77 818   75 576   79 297   86 571   85 656   112 135. Odhadněte trend této časové řady pomocí klouzavých průměrů s vyhlazovacím okénkem šířky 5 a graficky znázorněte.

Řešení:
Načteme datový soubor byty.sta o dvou proměnných ROK a POCET a jedenácti případech. Statistiky – Pokročilé lineární/nelineární modely – Časové řady/predikce – Proměnné POCET – OK– OK (transformace, autokorelace, kříž. korelace, grafy) – Vyhlazování – zaškrtneme N-bod. klouzavý průměr, N = 5 – OK (Transformovat vybrané řady) – vykreslí se graf, vrátíme se do Transformace proměnných – Uložit proměnné. Otevře se nový spreadsheet, kde v proměnné POCET_1 jsou uloženy klouzavé průměry pro N = 5. Proměnnou POČET_1 okopírujeme do původního datového souboru do nové proměnné KP5 (pozor – roky 1960, 1961, 1969 a 1970 nemají přiřazený odhad).

	1 ROK	2 POCET	3 KP5
1	1960	73766	
2	1961	86032	
3	1962	85221	80901,8
4	1963	82189	81712,2
5	1964	77301	79621,0
6	1965	77818	78436,2
7	1966	75576	79312,6
8	1967	79297	80983,6
9	1968	86571	87847,0
10	1969	85656	
11	1970	112135	

Pomocí Grafy – Bodové grafy – Vícenásobný graf vytvoříme graf časové řady počtu bytů s odhadnutým trendem.



Porovnání empirického a teoretického rozložení

Příklad 1.: Firma, která vlastní několik supermarketů, se zajímá, zda zákazníci dávají přednost některému dnu v týdnu pro nákup. Náhodně bylo vybráno 300 zákazníků, kteří měli říci, který den v týdnu nejčastěji nakupují v supermarketu.
Výsledky:

Den	pondělí	úterý	středa	čtvrtek	pátek	sobota	neděle
Počet	10	20	40	40	80	60	50

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že žádný den v týdnu nemá při nakupování v supermarketu přednost před jinými dny.

Návod:
Vytvoříme nový datový soubor se dvěma proměnnými a sedmi případy. Proměnná X obsahuje pozorované absolutní četnosti a Y vypočítané teoretické četnosti (v našem případě 300/7).
Statistics – Nonparametrics – Observed versus expected χ^2 – Variables Observed X, Expected Y, OK – Summary. Dostaneme tabulku:

	Pozorované vs. očekávané četnosti (Tema10ukol3) Chi-Kvadr. = 78,00000 df = 6 p < ,000000			
Případ	pozorov. X	očekáv. Y	P - O	(P-O)^2 /O
C: 1	10,0000	42,8571	-32,8571	25,19048
C: 2	20,0000	42,8571	-22,8571	12,19048
C: 3	40,0000	42,8571	-2,8571	0,19048
C: 4	40,0000	42,8571	-2,8571	0,19048
C: 5	80,0000	42,8571	37,1429	32,19048
C: 6	60,0000	42,8571	17,1429	6,85714
C: 7	50,0000	42,8571	7,1429	1,19048
Sčt	300,0000	300,0000	0,0000	78,00000

Komentář: Ve výstupní tabulce najdeme hodnotu testové statistiky (Chi-Square = 78) a odpovídající p-hodnotu, kterou porovnáme se zvolenou hladinou významnosti. V našem případě je p-hodnota velmi malá, takřka nulová, takže nulová hypotéza se zamítá na asymptotické hladině významnosti 0,05. S rizikem omylu nejvýše 5% jsme tedy prokázali, že zákazníci nakupují během týdne nerovnoměrně.

Příklad 2.: Byl zjišťován počet poruch určitého zařízení za 100 hodin provozu ve 150 disjunktních 100 hodinových intervalech. Výsledky pozorování:

Počet poruch za 100 hodin provozu: 0 1 2 3 4 a víc

Absolutní četnosti: 52 48 36 10 4

Na hladině významnosti 0,05 testujte hypotézu, že uvedený náhodný výběr pochází z Poissonova rozložení s parametrem $\lambda = 1{,}2$.

Návod:

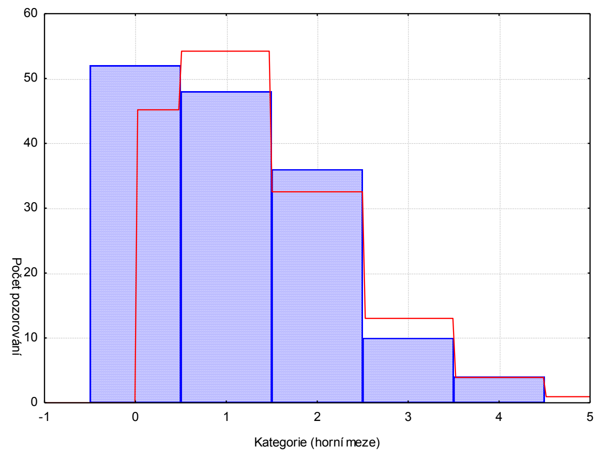
Vytvoříme nový datový soubor se dvěma proměnnými a 5 případy. Proměnná POCET obsahuje počet poruch, proměnná CETNOST pak absolutní četnosti zjištěného počtu poruch. Statistics – Distribution Fitting – Discrete Distributions – Poisson – OK – Variable POCET – Weight variable CETNOST – Status On – OK – Parameters Lambda 1,2, OK.

Proměnná: POCET, Rozdělení:Poissonovo, Lambda = 1,200 (Tema10ukol2)
Chí-kvadrát = 2,10955, sv = 2 (uprav.) , p = 0,34827

Kategorie	Pozorované Četnosti	Kumulativ. Pozorované	Procent Pozorované	Kumul. % Pozorované	Očekáv. Četnosti	Kumulativ. Očekáv.	Procent Očekáv.	Kumul. % Očekáv.	Pozorované - Očekáv.
<= 0,00000	52	52	34,66667	34,6667	49,59890	49,5989	33,06594	33,0659	2,40110
1,00000	48	100	32,00000	66,6667	54,88945	104,4884	36,59297	69,6589	-6,88945
2,00000	36	136	24,00000	90,6667	30,37216	134,8605	20,24810	89,9070	5,62784
3,00000	10	146	6,66667	97,3333	11,20395	146,0645	7,46930	97,3763	-1,20395
< Nekonečno	4	150	2,66667	100,0000	3,93554	150,0000	2,62369	100,0000	0,06446

Komentář: V záhlaví výstupní tabulky je uvedena hodnota testového kritéria (2,10955), počet stupňů volnosti = 2 a p-hodnota (0,34827). Nulová hypotéza se tedy nezamítá na asymptotické hladině významnosti 0,05.

Pro vytvoření grafu se vrátíme do Fitting Discrete Distributions – Quick – Plot of observed and expected distribution.



Komentář: V grafu jsou patrné určité rozdíly mezi hodnotami pravděpodobnostní a četnostní funkce, ale tyto rozdíly nejsou příliš velké.