

## Jednoduchá lineární regrese

**Motivace:** Cíl regresní analýzy - popsat závislost hodnot veličiny Y na hodnotách veličiny X.

Nutnost vyřešení dvou problémů:

- jaký typ funkce se použije k popisu dané závislosti;
- jak se stanoví konkrétní parametry daného typu funkce?

**ad a)** Při určení typu funkce je třeba provést teoretický rozbor zkoumané závislosti. Teoretická analýza může upozornit například na to, že

s růstem hodnot veličiny X budou mít hodnoty veličiny Y tendenci monotónně růst či klesat,

tato tendence má charakter zrychlujícího se či zpomalujícího se růstu či poklesu,

jde o závislost, kdy s růstem hodnot veličiny X dochází zpočátku k růstu hodnot veličiny Y, který je po dosažení určitého maxima vystřídán poklesem,

apod.

Můžeme např. zkoumat závislost ceny ojetého auta (veličina Y) na jeho stáří (veličina X). Je zřejmé, že s rostoucím stářím bude klesat cena, ale není jasné, zda lineárně, kvadraticky či dokonce exponenciálně.

Vždy se snažíme o to aby regresní model byl jednoduchý, tj. aby neobsahoval příliš mnoho parametrů. Příklad-li v úvahu více funkcí, posuzujeme jejich vhodnost pomocí různých kritérií – viz dále.

Často však nemáme dostatek informací k provedení teoretického rozboru. Pak se snažíme odhadnout typ funkce pomocí dvourozměrného tečkového diagramu.

Zde se omezíme na funkce, které závisejí lineárně na parametrech  $\beta_0, \beta_1, \dots, \beta_p$ .

Zvláštní pozornost budeme věnovat polynomiální funkci 1. stupně  $y = \beta_0 + \beta_1 x$ .

**ad b)** Odhady  $b_0, b_1, \dots, b_p$  neznámých parametrů  $\beta_0, \beta_1, \dots, \beta_p$  získáme na základě dvourozměrného datového souboru  $\begin{pmatrix} x_1 & y_1 \\ \dots & \dots \\ x_n & y_n \end{pmatrix}$  me-

todou nejmenších čtverců, tj. z podmínky, aby součet čtverců odchylek zjištěných a odhadnutých hodnot byl minimální.

## Specifikace klasického modelu lineární regrese

$Y = m(X; \beta_0, \beta_1, \dots, \beta_p) + \varepsilon$ , kde

$m(X; \beta_0, \beta_1, \dots, \beta_p)$  - **teoretická regresní funkce**, která lineárně závisí na neznámých regresních parametrech  $\beta_0, \beta_1, \dots, \beta_p$  a

známých funkcích  $f_1(X), \dots, f_p(X)$ , které již neobsahují neznámé parametry, tj.  $m(X; \beta_0, \beta_1, \dots, \beta_p) = \sum_{j=0}^p \beta_j f_j(X)$ , přičemž  $f_0(X) \equiv 1$ .

Jde o **deterministickou složku** modelu.

Složka  $\varepsilon$  - **náhodná složka** modelu. Je to náhodná odchylka od deterministické závislosti  $Y$  na  $X$ . Popisuje závislost vysvětlované proměnné na neznámých nebo nepozorovaných proměnných a popisuje i vliv náhody. Nelze ji funkčně vyjádřit.

Veličina  $Y$  - **závisle proměnná (též vysvětlovaná) veličina**.

Veličina  $X$  - **nezávisle proměnná (též vysvětlující) veličina**.

Pořídíme  $n$  dvojic pozorování  $(x_1, y_1), \dots, (x_n, y_n)$ , tj. dvourozměrný datový soubor  $\begin{pmatrix} x_1 & y_1 \\ \dots & \dots \\ x_n & y_n \end{pmatrix}$ .

Pro  $i = 1, \dots, n$  platí:  $y_i = m(x_i; \beta_0, \beta_1, \dots, \beta_p) + \varepsilon_i$ .

O náhodných odchylkách  $\varepsilon_1, \dots, \varepsilon_n$  předpokládáme, že

- $E(\varepsilon_i) = 0$  (odchylky nejsou systematické)
- $D(\varepsilon_i) = \sigma^2 > 0$  (všechna pozorování jsou prováděna s touž přesností)
- $C(\varepsilon_i, \varepsilon_j) = 0$  pro  $i \neq j$  (mezi náhodnými odchylkami neexistuje žádný lineární vztah)
- $\varepsilon_i \sim N(0, \sigma^2)$

V tomto případě hovoříme o **klasickém modelu lineární regrese**.

## Označení

$b_0, b_1, \dots, b_p$  - odhady regresních parametrů  $\beta_0, \beta_1, \dots, \beta_p$  (nejčastěji je získáme metodou nejmenších čtverců, tj. z podmínky, že výraz

$\sum_{i=1}^n \left( y_i - \sum_{j=0}^p \beta_j f_j(x_i) \right)^2$  nabývá svého minima pro  $\beta_j = b_j, j = 0, 1, \dots, p$ )

$\hat{m}(x; b_0, \dots, b_p)$  - empirická regresní funkce

$\hat{y}_i = \hat{m}(x_i; b_0, \dots, b_p) = \sum_{j=0}^p b_j f_j(x_i)$  - regresní odhad  $i$ -té hodnoty veličiny  $Y$  ( $i$ -tá predikovaná hodnota veličiny  $Y$ )

$e_i = y_i - \hat{y}_i$  -  $i$ -té reziduum

$S_E = \sum_{i=1}^n (y_i - \hat{y}_i)^2$  - reziduální součet čtverců

$s^2 = \frac{S_E}{n - p - 1}$  - odhad rozptylu  $\sigma^2$

$S_R = \sum_{i=1}^n (\hat{y}_i - m_2)^2$  - regresní součet čtverců ( $m_2 = \frac{1}{n} \sum_{i=1}^n y_i$ )

$S_T = \sum_{i=1}^n (y_i - m_2)^2$  - celkový součet čtverců ( $S_T = S_R + S_E$ )

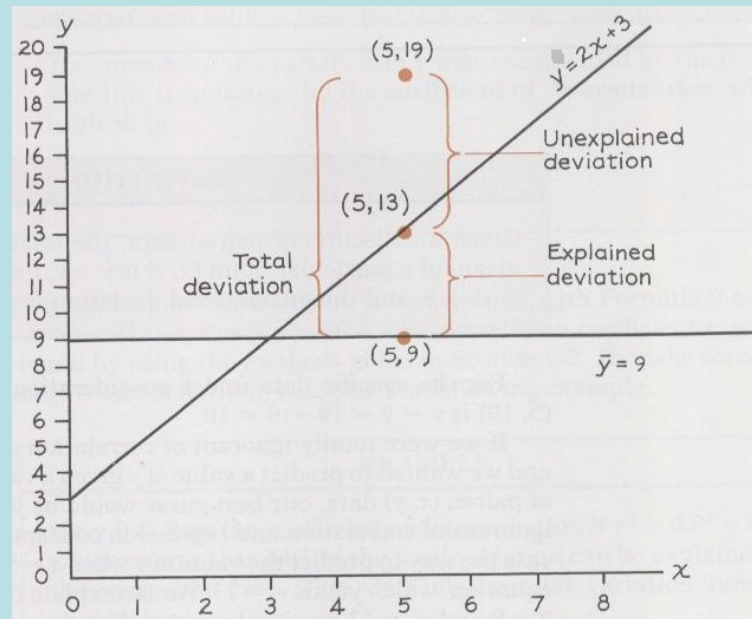
## Význam jednotlivých typů součtů čtverců

Předpokládejme, že máme dvourozměrný datový soubor, v němž průměr hodnot závisle proměnné veličiny Y je 9 a závislost veličiny Y na veličině X je popsána regresní přímkou  $y = 2x + 3$ . Dvourozměrný tečkový diagram obsahuje bod o souřadnicích (5, 19), který pochází z datového souboru. Na regresní přímce leží bod o souřadnicích (5, 13).

Odchylka zjištěné hodnoty 19 od průměru 9 je v obrázku označena „Total deviation“ a po umocnění je to jedna ze složek celkového součtu čtverců  $S_T$ , tj. složka  $y_i - \bar{y}$ .

Odchylka zjištěné hodnoty 19 od hodnoty 13 na regresní přímce je v obrázku označena „Unexplained deviation“ a po umocnění je to jedna ze složek reziduálního součtu čtverců  $S_E$ , tj. složka  $y_i - \hat{y}_i$ .

Odchylka hodnoty 13 na regresní přímce od průměru 9 je v obrázku označena „Explained deviation“ a po umocnění je to jedna ze složek regresního součtu čtverců  $S_R$ , tj. složka  $\hat{y}_i - \bar{y}$ .



## Maticový zápis klasického modelu lineární regrese

$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ , kde

$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$  - vektor pozorování závisle proměnné veličiny  $Y$ ,

$\mathbf{X} = \begin{bmatrix} 1 & f_1(\mathbf{x}_1) & \dots & f_p(\mathbf{x}_1) \\ \vdots & \vdots & \dots & \vdots \\ 1 & f_1(\mathbf{x}_n) & \dots & f_p(\mathbf{x}_n) \end{bmatrix}$  - regresní matice

(předpokládáme, že  $h(\mathbf{X}) = p+1 < n$ )

$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$  - vektor regresních parametrů,

$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$  - vektor náhodných odchylek.

Podmínky (a) až (d) lze zkráceně zapsat ve tvaru  $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I})$ .

Maticově zapsaná metoda nejmenších čtverců vede na rovnice

$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$  - systém normálních rovnic

$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$  - odhad vektoru  $\boldsymbol{\beta}$  získaný metodou nejmenších čtverců

$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$  - vektor regresních odhadů (vektor predikce)

$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$  - vektor reziduí

Vlastnosti odhadu  $\mathbf{b}$ :

- odhad  $\mathbf{b}$  je lineární, neboť je vytvořen lineární kombinací pozorování  $y_1, \dots, y_n$  s maticí vah  $\begin{bmatrix} \mathbf{x}'_1 \\ \vdots \\ \mathbf{x}'_n \end{bmatrix} \mathbf{X}'$ ;

- odhad  $\mathbf{b}$  je nestranný, neboť  $E(\mathbf{b}) = \boldsymbol{\beta}$ ;

- odhad  $\mathbf{b}$  má varianční matici  $\text{var } \mathbf{b} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ ;

- odhad  $\mathbf{b} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$  vzhledem k platnosti podmínky (d);

- pro odhad  $\mathbf{b}$  platí [Gaussova - Markovova věta](#): Odhad  $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$  je nejlepší nestranný lineární odhad vektoru  $\boldsymbol{\beta}$ .

## Příklad

U šesti obchodníků byla zjišťována poptávka po určitém druhu zboží loni (veličina X - v kusech) a letos (veličina Y - v kusech).

číslo obchodníka	1	2	3	4	5	6
poptávka loni (X)	20	60	70	100	150	260
poptávka letos (Y)	50	60	60	120	230	320

Předpokládejte, že závislost letošní poptávky na loňské lze vystihnout regresní přímkou. Sestavte regresní matici, vypočtěte odhady regresních parametrů a napište rovnici regresní přímky. Interpretujte parametry regresní přímky.

## Řešení:

Sestavíme regresní matici.

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \text{ tedy } \mathbf{X} = \begin{pmatrix} 1 & 20 \\ 1 & 60 \\ 1 & 70 \\ 1 & 100 \\ 1 & 150 \\ 1 & 260 \end{pmatrix}.$$

Podle vzorce  $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$  získáme odhady regresních parametrů.

Nejprve vypočítáme matici  $\mathbf{X}'\mathbf{X} = \begin{pmatrix} 6 & 660 \\ 660 & 109000 \end{pmatrix}$  a k ní inverzní matici  $(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0,499084 & -0,003022 \\ -0,003022 & 0,000027 \end{pmatrix}$ .

Dále získáme součin  $\mathbf{X}'\mathbf{y} = \begin{pmatrix} 840 \\ 138500 \end{pmatrix}$  a nakonec vektor odhadů regresních parametrů:  $\mathbf{b} = \begin{pmatrix} 0,499084 & -0,003022 \\ -0,003022 & 0,000027 \end{pmatrix} \begin{pmatrix} 840 \\ 138500 \end{pmatrix} = \begin{pmatrix} 0,6868 \\ 1,2665 \end{pmatrix}$ .

Regresní přímka má tedy rovnici

$$y = 0,6868 + 1,2665 x.$$

Znamená to, že při nulové loňské poptávce by letošní poptávka činila 0,6868 kusů a při zvýšení loňské poptávky o 10 kusů by se letošní poptávka zvedla o 12,665 kusů.

## Výpočet pomocí systému STATISTICA

Vytvoříme nový datový soubor se dvěma proměnnými X a Y a 6 případy:

Statistiky – Vícerozměrná regrese – Závisle proměnná Y, nezávisle proměnná X - OK – OK – Výpočet: Výsledky regrese.

Výsledky regrese se závislou proměnnou : Y (Tabulka1)						
R= ,97197702 R2= ,94473932 Upravené R2= ,93092415						
F(1,4)=68,384 p<,00117 Směrod. chyba odhadu : 29,219						
N=6	Beta	Sm.chyba beta	B	Sm.chyba B	t(4)	Úroveň p
Abs.člen			0,686813	20,64236	0,033272	0,975052
X	0,971977	0,117538	1,266484	0,15315	8,269474	0,001167

Ve výstupní tabulce najdeme koeficient  $b_0$  ve sloupci B na řádku označeném Abs. člen, koeficient  $b_1$  ve sloupci B na řádku označeném X.

Rovnice regresní přímky:

$$y = 0,686813 + 1,266484 x.$$

Znamená to, že při nulové loňské poptávce by letošní poptávka činila 0,6868 kusů a při zvýšení loňské poptávky o 10 kusů by se letošní poptávka zvedla o 12,665 kusů.

## Testování významnosti modelu jako celku (celkový F-test)

Na hladině významnosti  $\alpha$  testujeme

$$H_0: \beta_1, \dots, \beta_p = 0, \dots, 0 \text{ proti } H_1: \beta_1, \dots, \beta_p \neq 0, \dots, 0.$$

(Nulová hypotéza říká, že dostačující je model konstanty.)

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

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

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

Výsledky F-testu zapisujeme do tabulky analýzy rozptylu:

zdroj variability	součet čtverců	stupně volnosti	podíl	statistika F
model	$S_R$	$p$	$S_R/p$	$\frac{S_R/p}{S_E/(n-p-1)}$
reziduální	$S_E$	$n-p-1$	$S_E/(n-p-1)$	-
celkový	$S_T$	$n-1$	-	-

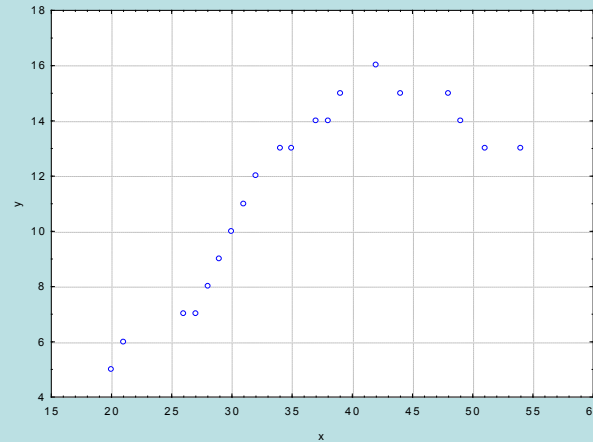


### Příklad:

Majitelé prodejny počítačových her nechali své prodavače absolvovat kurz prodejních dovedností. Poté zjišťovali po dobu 20 dnů, kolik osob navštíví během otevírací doby prodejnu (proměnná X) a jaká je v tento den tržba (proměnná Y, udává se v tisících Kč a je zaokrouhlená).

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
$x_i$	20	21	2	27	28	29	30	31	32	34	35	37	38	39	42	44	48	49	51	54
$y_i$	5	6	7	7	8	9	10	11	12	13	13	14	14	15	16	15	15	14	13	13

### Dvourozměrný tečkový diagram



Z grafu závislosti Y na X vyplývá, že s rostoucím počtem zákazníků se tržby zvyšují, avšak při denním počtu zákazníků asi 42 dosahují svého maxima a pak už zase klesají (vyšší počet zákazníků obsluha prodejny nezvládá a zákazníci odcházejí, aniž by nakoupili). Zdá se tedy, že vhodným modelem závislosti tržeb na počtu zákazníků bude regresní parabola

$$y = 3_0 + 3_1x + 3_2x^2 + \varepsilon.$$

Odhadněte parametry regresního modelu a proveďte celkový F-test.

## Řešení:

Vytvoříme nový datový soubor se třemi proměnnými X, Xkv, Y a o 20 případech. Do proměnných X a Y napíšeme zjištěné hodnoty a do Dlouhého jména proměnné Xkv napíšeme  $= X^2$ .

Získání odhadů  $b_0$ ,  $b_1$ ,  $b_2$ :

Statistiky – Vícerozměrná regrese – Závisle proměnná Y, nezávisle proměnné X, Xkv - OK – OK – Výpočet: Výsledky regrese.

Výsledky regrese se závislou proměnnou : y (prodejna_sof tware.sta)						
R= ,95519276 R2= ,91239322 Upravené R2= ,90208653						
F(2,17)=88,524 p<,00000 Směrod. chyba odhadu : 1,0623						
N=20	b*	Sm.chyba z b*	b	Sm.chyba z b	t(17)	p-hodn.
Abs.člen			-20,7723	3,373256	-6,15792	0,000011
x	4,52641	0,548220	1,5651	0,189559	8,25655	0,000000
xkv	-3,73838	0,548220	-0,0173	0,002535	-6,81912	0,000003

Regresní parabola má tedy tvar:  $y = -20,7723 + 1,5651x - 0,0173x^2$ .

Výsledky celkového F-testu jsou uvedeny v záhlaví výstupní tabulky. Testová statistika F nabývá hodnoty 88,524, odpovídající p-hodnota je blízká 0, tedy na hladině významnosti 0,05 zamítáme hypotézu, že dostačující je model konstanty.

Podrobnější výsledky získáme v tabulce analýzy rozptylu:

Aktivujeme Výsledky–vícenásobná regrese – Detailní výsledky – ANOVA

Analýza rozptylu (prodejna_sof tware.sta)					
Efekt	Součet čtv erců	sv	Průměr čtv erců	F	p-hodn.
Regres.	199,8141	2	99,90706	88,52445	0,000000
Rezid.	19,1859	17	1,12858		
Celk.	219,0000				

## Testování významnosti regresních parametrů (dílní t-testy)

Na hladině významnosti  $\alpha$  pro  $j = 0, 1, \dots, p$  testujeme hypotézu  $H_0: \beta_j = 0$  proti  $H_1: \beta_j \neq 0$ .

Testová statistika:  $T_j = \frac{b_j}{s_{b_j}}$  má rozložení  $t(n-p-1)$ , pokud  $H_0$  platí.

Kritický obor:  $w = (-\infty, -t_{1-\alpha/2, n-p-1}] \cup [t_{1-\alpha/2, n-p-1}, \infty)$ .  
 $T_j \in w \Rightarrow H_0$  zamítáme na hladině významnosti  $\alpha$ .

### Příklad:

V předešlém příkladě, kde byla modelována závislost tržby na počtu zákazníků regresní parabolou, proved'te dílní t-testy o nevýznamnosti jednotlivých regresních parametrů

### Řešení:

Stačí interpretovat výstupní tabulku vícenásobné regrese:

Výsledky regrese se závislou proměnnou : y (prodejna_software.sta)						
R= ,95519276 R2= ,91239322 Upravené R2= ,90208653						
F(2,17)=88,524 p<,00000 Směrod. chyba odhadu : 1,0623						
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(17)	p-hodn.
N=20						
Abs.člen			-20,7723	3,373256	-6,15792	0,000011
x	4,52641	0,548220	1,5651	0,189559	8,25655	0,000000
xkv	-3,73838	0,548220	-0,0173	0,002535	-6,81912	0,000003

Sloupec označený t(17) obsahuje realizace testových statistik a sloupec p-hodn. pak odpovídající p-hodnoty. Ve všech třech případech jsou p-hodnoty menší než 0,05, tedy na hladině významnosti 0,05 zamítáme hypotézy o nevýznamnosti regresních parametrů  $\beta_0, \beta_1, \beta_2$ .

## Kritéria pro posouzení vhodnosti zvolené regresní funkce

### a) Index determinace

$$ID^2 = \frac{S_R}{S_T} = 1 - \frac{S_E}{S_T} \quad \text{- index determinace (} 0 \leq ID^2 \leq 1 \text{)}$$

- udává, jakou část variability závisle proměnné veličiny Y lze vysvětlit zvolenou regresní funkcí (často se udává v %);
- je zároveň mírou těsnosti závislosti proměnné Y na proměnné X;
- je to obecná míra, nezávislá na typu regresní funkce (lze použít i pro měření nelineární závislosti);
- je to míra, která nebere v úvahu počet parametrů regresní funkce. U regresních funkcí s více parametry vychází tedy obvykle vyšší než u regresních funkcí s méně parametry;
- tato míra není symetrická.

Za vhodnější se považuje ta regresní funkce, pro niž je index determinace vyšší. V případě, že porovnáváme několik modelů s rozdílným počtem parametrů, používáme adjustovaný index determinace:

$$ID_{adj}^2 = ID^2 \cdot \frac{n - ID^2 \cdot p}{n - p} \quad \text{- adjustovaný index determinace}$$

V příkladu s prodejem software najdeme index determinace ve výstupní tabulce regrese:

Výsledky regrese se závislou proměnnou : y (prodejna_software.sta)						
R= ,95519276 R2= ,91239322 Upravené R2= ,90208653						
F(2,17)=88,524 p<,00000 Směrod. chyba odhadu : 1,0623						
N=20	b*	Sm.chyba z b*	b	Sm.chyba z b	t(17)	p-hodn.
Abs.člen			-20,7723	3,373256	-6,15792	0,000011
x	4,52641	0,548220	1,5651	0,189559	8,25655	0,000000
xkv	-3,73838	0,548220	-0,0173	0,002535	-6,81912	0,000003

Index determinace je zde označen jako R2, nabývá hodnoty 0,9124 a říká nám, že 91,24% variability tržeb je vysvětleno regresní parabolou. Adjustovaný index determinace je označen Upravené R2.

## b) Testové kritérium F

Za vhodnější je považována ta regresní funkce, u níž je hodnota testové statistiky  $F = \frac{S_R/p}{S_E/(n-p-1)}$  pro test významnosti modelu jako celku vyšší.

Ve výstupní tabulce regrese je testová statistika F uvedena v záhlaví:

Výsledky regrese se závislou proměnnou : y (prodejna_software.sta)						
R= ,95519276 R2= ,91239322 Upravené R2= ,90208653						
F(2,17)=88,524 p<,00000 Směrod. chyba odhadu : 1,0623						
N=20	b*	Sm.chyba z b*	b	Sm.chyba z b	t(17)	p-hodn.
Abs.člen			-20,7723	3,373256	-6,15792	0,000011
x	4,52641	0,548220	1,5651	0,189559	8,25655	0,000000
xkv	-3,73838	0,548220	-0,0173	0,002535	-6,81912	0,000003

V našem příkladě je označena F(2,17) a nabývá hodnoty 88,524.

### c) Reziduální součet čtverců a reziduální rozptyl

Reziduální součet čtverců:  $S_E = \sum_{i=1}^n y_i - \hat{y}_i$

Za vhodnější považujeme funkci, která má reziduální součet čtverců nižší. Reziduální součet čtverců lze použít pouze tehdy, když srovnáváme funkce se stejným počtem parametrů.

Reziduální rozptyl:  $s^2 = \frac{S_E}{n - p}$

Za vhodnější považujeme tu funkci, která má reziduální rozptyl nižší. Reziduální rozptyl můžeme použít vždy, bez ohledu na to, kolik parametrů mají srovnávané regresní funkce.

Obě charakteristiky najdeme v tabulce ANOVA:

Analýza rozptylu (prodejna_software.sta)					
Efekt	Součet čtverců	sv	Průměr čtverců	F	p-hodn.
Regres.	199,8141	2	99,90706	88,52445	0,000000
Rezid.	19,1859	17	1,12858		
Celk.	219,0000				

Reziduální součet čtverců je 19,1859 a reziduální rozptyl je 1,12858.

#### d) Střední absolutní procentuální chyba predikce (MAPE)

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

Za vhodnější považujeme tu funkci, která má MAPE nižší.

System STATISTICA MAPE neposkytuje, tuto chybu musíme vypočítat.

Statistiky – Vícerozměrná regrese – Závisle proměnná Y, nezávisle proměnné x, xkv - OK – OK – zvolíme Rezi-  
dua/předpoklady/předpovědi – Reziduální analýza – Uložit – Uložit rezidua & předpovědi – vybereme proměnnou y - OK.  
K vzniklému datovému souboru přidáme jednu novou proměnnou, nazveme ji chyba a do jejího Dlouhého jména napíšeme  
 $=100 * \text{abs}((v1-v2)/v1)$

Pomocí Statistiku – Základní statistiky/tabulky – Popisné statistiky zjistíme průměr proměnné chyba. V našem případě je  
MAPE 9,31%.

### e) Analýza reziduí

Rezidua považujeme za odhady náhodných odchylek a klademe na ně stejné požadavky jako na náhodné odchylky, tj. mají být nezávislá,

mají být normálně rozložená,

mají mít nulovou střední hodnotu,

mají mít konstantní rozptyl (tj. jsou homoskedastická).

Nezávislost reziduí (autokorelaci) posuzujeme např. pomocí Durbinovy – Watsonovy statistiky, která by se měla nacházet v intervalu  $\langle 1,4;2,6 \rangle$  (to je ovšem pouze orientační vodítko, korektní postup spočívá v porovnání této statistiky s tabelovanou kritickou hodnotou).

Normalitu reziduí ověřujeme pomocí testů normality (např. Lilieforsovou variantou Kolmogorovova – Smirnovova testu nebo Shapirovým – Wilksovým testem) či graficky pomocí N-P plotu.

Testování nulovosti střední hodnoty reziduí provádíme pomocí jednovýběrového t-testu.

Homoskedasticitu reziduí posuzujeme pomocí grafu závislosti reziduí na predikovaných hodnotách. V tomto grafu by rezidua měla být rovnoměrně rozptýlena.



**Příklad:** Proveďte analýzu reziduí pro příklad s modelováním závislosti tržby na počtu zákazníků.

### Posouzení nezávislosti reziduí pomocí Durbinovy – Watsonovy statistiky:

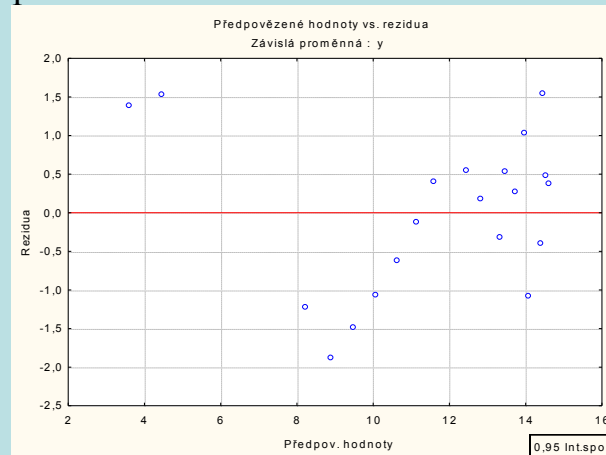
Statistiky – Vícenásobná regrese – proměnná Závislá: y, nezávislá x, xkv – OK – na záložce Residua/předpoklady/předpovědi vybereme Reziduální analýza - Detaily – Durbin-Watsonova statistika:

	Durbin-Watson.d	Sériové korelace
Odhad	0,702506	0,599248

Hodnota této statistiky je nízká, svědčí o tom, že rezidua jsou kladně korelovaná.

### Posouzení homoskedasticity reziduí

Reziduální analýza – Bodové grafy – Předpovědi vs. rezidua



Je vidět, že rezidua nejsou kolem 0 rozmístěna náhodně. Model s regresní parabolou tedy není úplně vhodný.

### Testování nulovosti střední hodnoty reziduí:

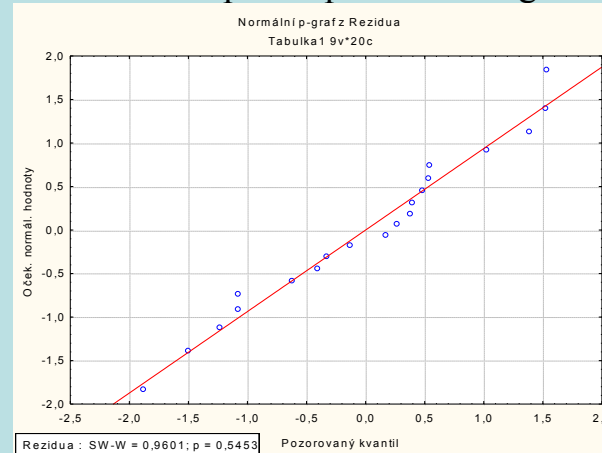
Pro proměnnou Rezidua z tabulky uložené pomocí Reziduální analýzy provedeme jednovýběrový t-test: Statistiky - Základní statistiky/tabulky – t-test, samost. vzorek – OK – proměnné Rezidua – OK.

Proměnná	Průměr	Sm.odch.	N	Sm.chyba	Referenční konstanta	t	SV	p
Rezidua	-0,000000	1,004880	20	0,224698	0,00	-0,000000	19	1,000000

Na hladině významnosti 0,05 nezamítáme hypotézu, že střední hodnota reziduí je 0.

### Posouzení normality reziduí:

Na záložce Pravděpodobnostní grafy zvolíme Normální pravděpodobnostní graf reziduí:



Rezidua se řadí kolem ideální přímky, lze tedy soudit, že se řídí normálním rozložením.

**Závěr:** V neprospěch regresní paraboly hovoří hodnota Durbinovy – Watsonovy statistiky a graf závislosti reziduí na predikovaných hodnotách.

## Popis časových řad

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

### Druhy časových řad

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

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

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

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

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 (očistěná 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  $=\text{trzba} * 365 / (12 * \text{dm})$ .

	1	2	3
	trzba	dm	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  $t_1, \dots, t_n$ , na svislou osu odpovídající hodnoty  $y_1, \dots, y_n$ . Dvojice bodů  $(t_i, y_i)$ ,  $i = 1, \dots, n$  spojíme úsečkami.

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

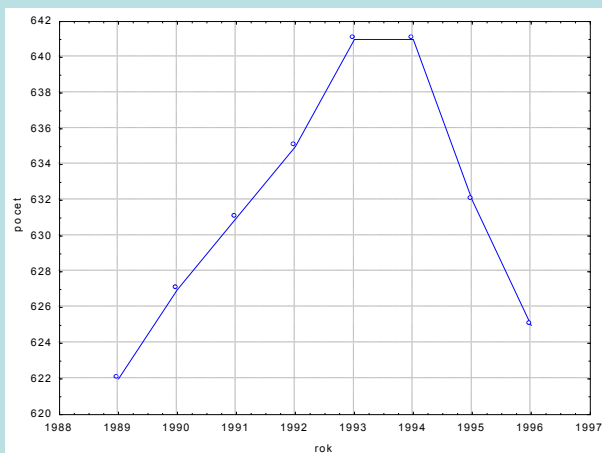
1989	1990	1991	1992	1993	1994	1995	1996
622	627	631	635	641	641	632	625

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 – pocet – 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.

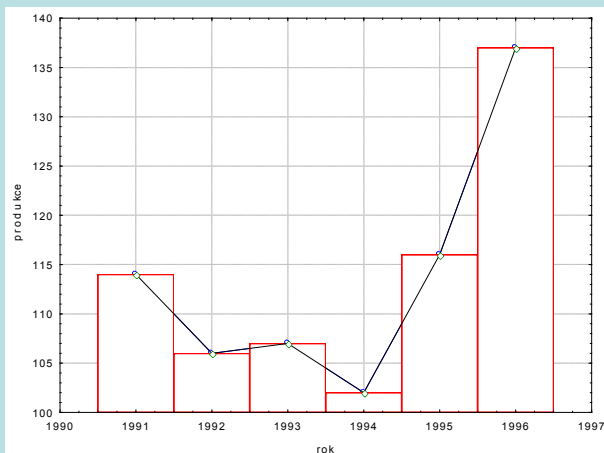
1991	1992	1993	1994	1995	1996
114	106	107	102	116	137

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

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

Nemají-li intervaly stejnou délku, vypočteme  $d_i = t_i - t_{i-1}$ ,  $i = 2, \dots, n$  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:**  $\bar{y} = \frac{1}{9} \left( \frac{14194 + 14738}{2} + 14271 + \dots + 14631 \right) = 14430$ .

## Průměr intervalové časové řady

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

**Příklad:** Vypočtete 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:**  $\bar{y} = \frac{1}{7} (1303,6 + \dots + 1433,8) = 1398,7 .$



## Dynamické charakteristiky časových řad

### Absolutní přírůstky

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

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

atd.

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

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

### 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

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

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

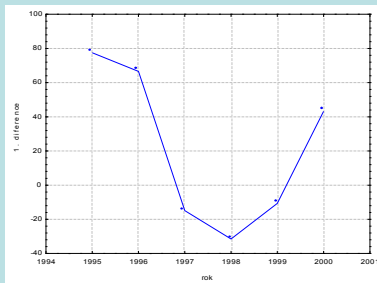
**Řešení:**

rok	HDP	$\Delta y_i$	$k_i$	$\delta_i$
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

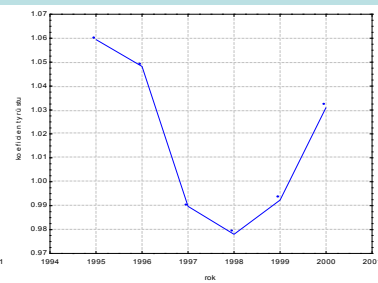
Průměrný absolutní přírůstek:  $\bar{\Delta} = \frac{1433,8 - 1303,6}{6} = 21,7$ , tzn., že v období 1994 – 2000 rostl HDP průměrně o 21,7 miliard Kč ročně.

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:

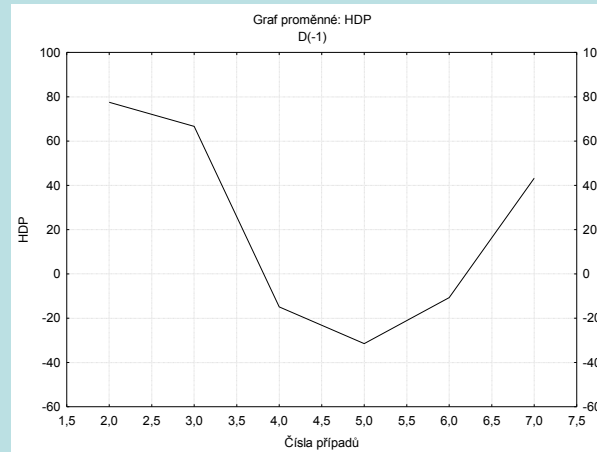


Graf koeficientů růstu:



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

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

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

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

## Aditivní model časové řady

Předpokládejme, že pro časovou řadu  $y_1, \dots, y_n$  platí model

$$y_t = f(t) + \varepsilon_t, \quad t = 1, \dots, n, \text{ kde}$$

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

$\varepsilon_t$  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  $\varepsilon_t$  je **bílý šum**).

## Odhad trendu časové řady pomocí klouzavých průměrů

### Podstata klouzavých průměrů

Předpokládáme, že časová řada se řídí aditivním modelem

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

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. Necht' 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:

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

### 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.

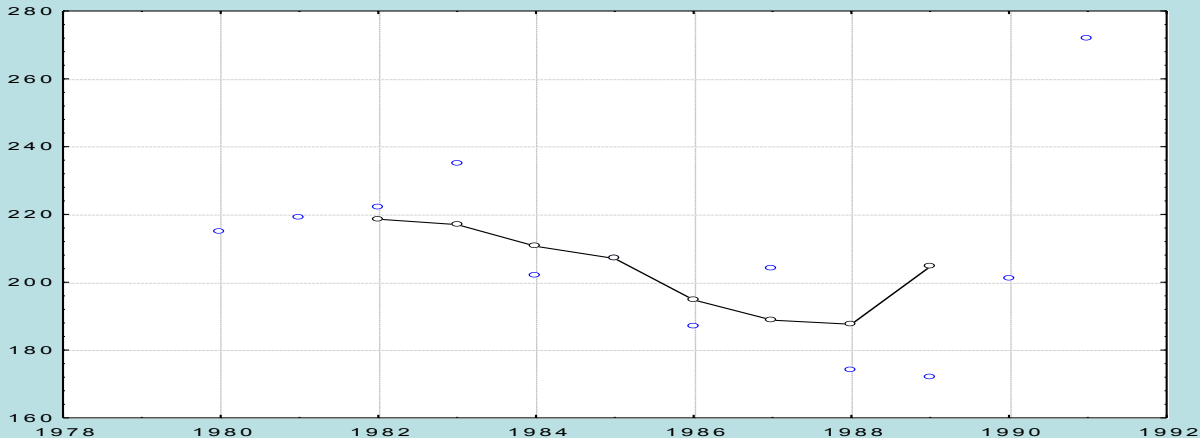
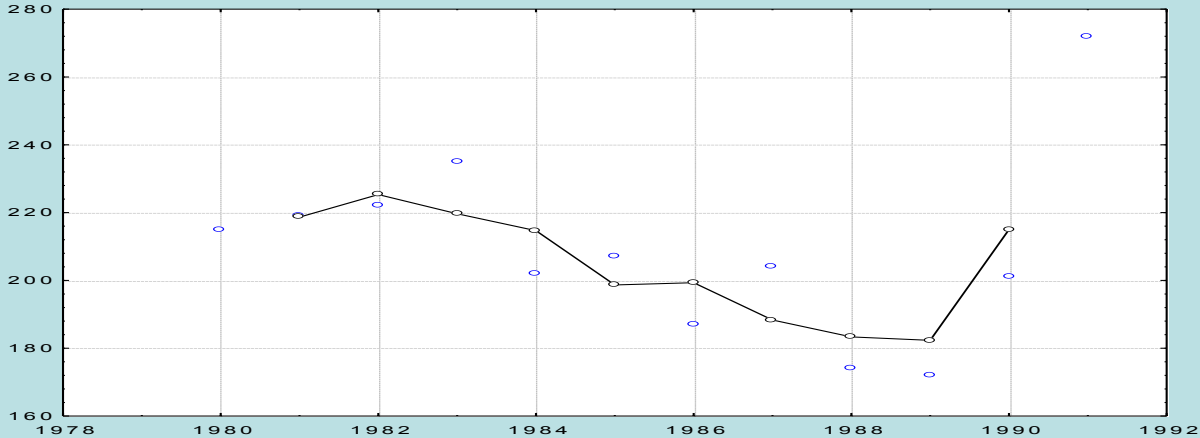
- Odhadněte trend této časové řady pomocí klouzavých průměrů s vyhlazovacím okénkem šířky 3 a poté 5.
- Graficky znázorněte průběh časové řady s odhadnutým trendem.

### Řešení pomocí systému STATISTICA:

Vytvoříme datový soubor export\_piva.sta o dvou proměnných ROK a VYVOZ a dvanácti případech. Statistika – Pokročilé lineární/nelineární modely – Časové řady/predikce – Proměnné Y – OK – OK (transformace, autokorelace, kříž. korelace, grafy) – Vyhlažování – 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.





## 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. Tato univerzálnost je ovšem provázena poněkud sníženou silou testu. Proto byly pro některá rozložení vyvinuty speciální testy využívající charakteristických vlastností těchto rozložení. Zde uvedeme tzv. jednoduché testy exponenciálního a Poissonova rozložení.

## Testy dobré shody

### Popis testu

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

Spojitý případ:

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

Diskrétní případ:

- určíme varianty  $x_{[j]}, j = 1, \dots, r$
- 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  $p_j = \Phi(x_{[j]}) - \Phi(x_{[j-1]}) = P\{X = x_{[j]}\}$ .

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  $\alpha$ , když  $K \geq \chi^2_{1-\alpha}(r-1-p)$ . Aproximace se považuje za vyhovující, když  $np_j \geq 5, j = 1, \dots, r$ .

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

**Příklad:** (Testování shody empirického a teoretického rozložení při úplně specifikovaném problému)

Ze souboru rodin s pěti dětmi bylo náhodně vybráno 84 rodin a byl zjišťován počet chlapců:

Počet chlapců	0	1	2	3	4	5
Počet rodin	3	10	22	31	14	4

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že rozložení počtu chlapců se řídí binomickým rozložením  $Bi(5; 0,5)$ .

**Řešení:** Počet chlapců v náhodně vybrané rodině s 5 dětmi je náhodná veličina s rozložením  $Bi(5; 0,5)$ , její pravděpodobnostní funkce je

$$p_j = \binom{5}{j} \frac{1}{32}, j = 0, 1, \dots, 5.$$

Výpočty potřebné pro stanovení testové statistiky  $K$  uspořádáme do tabulky.

$j$	$n_j$	$p_j$	$np_j$
0	3	0,03125	$84 \cdot 0,03125 = 2,625$
1	10	0,15625	$84 \cdot 0,15625 = 13,125$
2	22	0,3125	$84 \cdot 0,3125 = 26,25$
3	31	0,3125	$84 \cdot 0,3125 = 26,25$
4	14	0,15625	$84 \cdot 0,15625 = 13,125$
5	4	0,03125	$84 \cdot 0,03125 = 2,625$

Podmínky dobré aproximace nejsou splněny, sloučíme tedy první dvě varianty a poslední dvě varianty.

$j$	$n_j$	$p_j$	$np_j$	$\frac{(n_j - np_j)^2}{np_j}$
0 a 1	13	0,1875	$84 \cdot 0,1875 = 15,75$	0,480159
2	22	0,3125	$84 \cdot 0,3125 = 26,25$	0,688095
3	31	0,3125	$84 \cdot 0,3125 = 26,25$	0,859524
4 a 5	18	0,1875	$84 \cdot 0,1875 = 15,75$	0,321429

Vypočteme realizaci testové statistiky:  $K = 0,48059 + 0,688095 + 0,859524 + 0,321429 = 2,3492$ , počet tříd  $r = 4$ , počet odhadovaných parametrů  $p = 0$ ,  $r - p - 1 = 3$ , kritický obor  $W = \langle \chi^2_{1-\alpha; r-p-1}, \infty \rangle = \langle \chi^2_{0,95; 3}, \infty \rangle = \langle 7,8147; \infty \rangle$ . Protože  $K \notin W$ , nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.

## Výpočet pomocí systému STATISTICA:

Vytvoříme datový soubor se dvěma proměnnými a čtyřmi případy. Proměnná nj obsahuje zjištěné četnosti (po sloučení variant), proměnná npj pak teoretické četnosti.

Statistiky – Neparametrická statistika – Pozorované vs. očekávané  $\chi^2$  – OK – Proměnné – Pozorované četnosti nj, očekávané četnosti npj – OK – Výpočet.

Pozorované vs. očekávané četnosti (Tabulka1)				
Chi-Kvadr. = 2,349206 sv = 3 p = ,503161				
Případ	pozorov. nj	očekáv. npj	P - O	(P-O) <sup>2</sup> /O
C: 1	13,00000	15,75000	-2,75000	0,480159
C: 2	22,00000	26,25000	-4,25000	0,688095
C: 3	31,00000	26,25000	4,75000	0,859524
C: 4	18,00000	15,75000	2,25000	0,321429
Sčt	84,00000	84,00000	0,00000	2,349206

V záhlaví výstupní tabulky je uvedena hodnota testového kritéria (2,349206), počet stupňů volnosti = 3 a p-hodnota (0,503161). Nulová hypotéza se tedy nezamítá na asymptotické hladině významnosti 0,05.

**Příklad:** (Testování shody empirického a teoretického rozložení při neúplně specifikovaném problému – diskretní případ)  
 V tabulce jsou rozříděny 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:

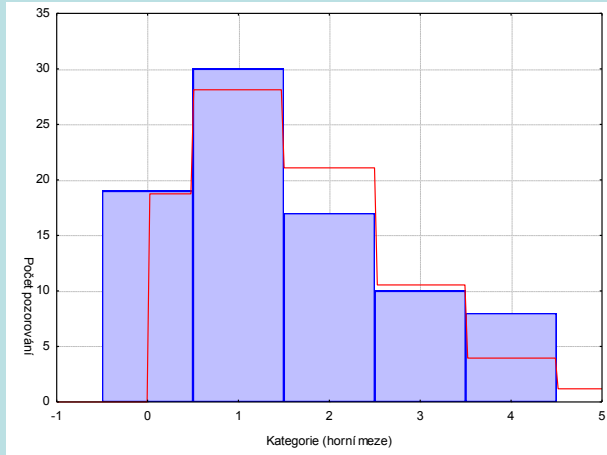
Vytvoříme datový soubor s dvěma proměnnými a 5 případy. Proměnná POCET obsahuje počet vstřelených branek, proměnná CETNOST pak počet zápasů, v nichž bylo dosaženo zjištěného počtu branek.

Statistiky – Prokládání rozdělení – Diskretní 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-kv adrát = 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  $\lambda$  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): 165 170 170 179 170 168 174 162 167 165 170 173 183 176 165 168 171 178 168 168 169 163 172 184 176 175 176 169 168 170 166 160 167 162 162 166 170 168 155 162 169 166 160 169 165 163 168 163

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:

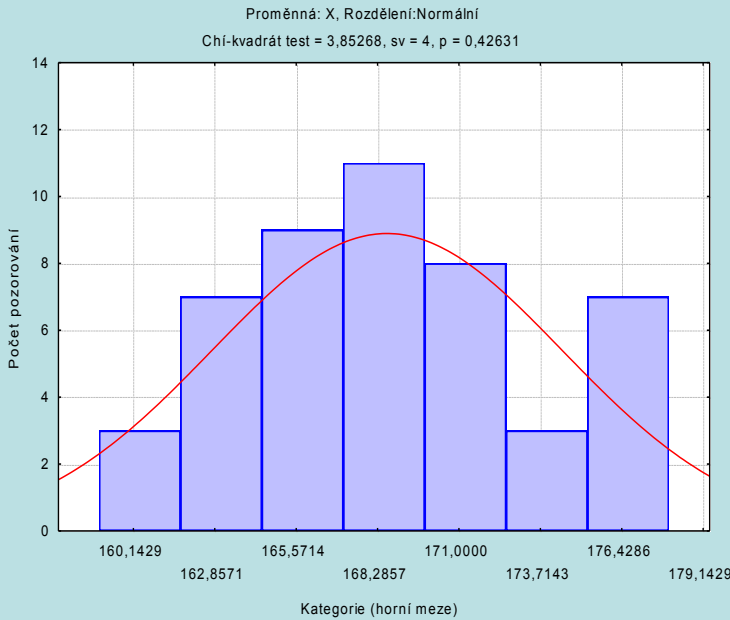
Statistiky - Prokládání rozdělení – ponecháme implicitní nastavení na normální rozložení – OK – Proměnná X – OK – na záložce Parametry změním Počet kategorií na 7 (podle Sturgesova pravidla) – Výpočet.

Horní hranice	Proměnná: X, Rozdělení: Normální (vyska.sta) Chi-kvadrát = 1,09280, sv = 1 (uprav.), p = 0,29585							
	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 roztrídění 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.





## Jednoduchý test exponenciálního a Poissonova rozložení

### Jednoduchý test exponenciálního rozložení

Testujeme hypotézu, která tvrdí, že náhodný výběr  $X_1, \dots, X_n$  pochází z exponenciálního rozložení. Označme  $M$  výběrový průměr a  $S^2$  výběrový rozptyl tohoto náhodného výběru. Víme, že střední hodnota náhodné veličiny  $X \sim \text{Ex}(\lambda)$  je  $E(X) = 1/\lambda$  a rozptyl je  $D(X) = 1/\lambda^2$ . Test založíme na statistice  $K = \frac{(n-1)S^2}{M^2}$ , která se v případě platnosti  $H_0$  asymptoticky řídí rozložením  $\chi^2(n-1)$ . Kritický obor:  $w = \langle 0, \chi^2_{\alpha/2}(n-1) \rangle \cup \langle \chi^2_{1-\alpha/2}(n-1), \infty \rangle$ . Jestliže  $K \in w$ ,  $H_0$  zamítáme na asymptotické hladině významnosti  $\alpha$ .

### Příklad

Byla zkoumána doba životnosti 45 součástek (v hodinách). Průměrná životnost byla  $m = 99,93$  a rozptyl  $s^2 = 7328,91$ . Na asymptotické hladině významnosti 0,05 testujte hypotézu, že daný náhodný výběr pochází z exponenciálního rozložení.

### Řešení:

Testovou statistiku  $K$  vypočteme podle vzorce  $K = \frac{(n-1)S^2}{M^2}$ . Kritický obor má tvar:  $w = \langle 0, \chi^2_{\alpha/2}(n-1) \rangle \cup \langle \chi^2_{1-\alpha/2}(n-1), \infty \rangle$ .

V našem případě  $K = 32,2924$ ,  $w = \langle 0; 27,575 \rangle \cup \langle 64,202; \infty \rangle$ ,  $H_0$  tedy nezamítáme na asymptotické hladině významnosti 0,05.

## Jednoduchý test Poissonova rozložení

Testujeme hypotézu, která tvrdí, že náhodný výběr  $X_1, \dots, X_n$  pochází z Poissonova rozložení. Označme  $M$  výběrový průměr a  $S^2$  výběrový rozptyl tohoto náhodného výběru. Víme, že střední hodnota náhodné veličiny  $X \sim \text{Po}(\lambda)$  je  $E(X) = \lambda$  a rozptyl je  $D(X) = \lambda$ . Test založíme na statistice  $K = \frac{(n-1)S^2}{M}$ , která se v případě platnosti  $H_0$  asymptoticky řídí rozložením  $\chi^2(n-1)$ . Kritický obor:  $W = \langle 0, \chi^2_{\alpha/2}(n-1) \rangle \cup \langle \chi^2_{1-\alpha/2}(n-1), \infty \rangle$ . Jestliže  $K \in W$ ,  $H_0$  zamítáme na asymptotické hladině významnosti  $\alpha$ .

## Příklad

Studujeme rozložení počtu pacientů, kteří během 75 dnů přijdou na pohotovost. Osmihodinovou pracovní dobu rozdělíme do půlhodinových intervalů a v každém intervalu zjistíme počet příchozích pacientů:

Počet pacientů	0	1	2	3	4	4	6	7	8	9	10
Pozrovaná četnost	79	188	282	275	196	114	45	10	7	3	1

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že daný náhodný výběr pochází z Poissonova rozložení.

## Řešení:

Celkový počet pacientů je  $n = 1200$ . Realizaci výběrového průměru  $M$  získáme jako vážený průměr počtu pacientů ( $m = 2,8033$ ) a realizaci výběrového rozptylu  $S^2$  získáme jako vážený rozptyl počtu pacientů ( $s^2 = 2,7086$ ). Testovou statistiku

vypočteme podle vzorce  $K = \frac{(n-1)S^2}{M}$ , tedy  $K = 1158,5$ , kritický obor

$$W = \langle 0, \chi^2_{\alpha/2}(n-1) \rangle \cup \langle \chi^2_{1-\alpha/2}(n-1), \infty \rangle = \langle 0, \chi^2_{0,025}(1199) \rangle \cup \langle \chi^2_{0,975}(1199), \infty \rangle = \langle 0; 1104,93 \rangle \cup \langle 1296,86; \infty \rangle$$

Protože testová statistika se nerealizuje v kritickém oboru,  $H_0$  nezamítáme na asymptotické hladině významnosti 0,05.