



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

## Přednášky

předmětu M6130 Výpočetní statistika

## Průzkumová analýza jednorozměrných dat, diagnostické grafy

### Motivace

Průzkumová analýza dat je odvětví statistiky, které pomocí různých postupů odhaluje zvláštnosti v datech. Při zpracování dat se často používají metody, které jsou založeny na předpokladu, že data pocházejí z nějakého konkrétního rozložení, nejčastěji normálního. Tento předpoklad nemusí být vždy splněn, protože data

- mohou pocházet z jiného rozložení
- mohou být zatížena hrubými chybami
- mohou pocházet ze směsi několika rozložení.

Proto je důležité provést průzkumovou analýzu dat, abychom se vyvarovali neadekvátního použití statistických metod.

Data zkoumáme pomocí **funkcionálních** a **číselných charakteristik** a pomocí **diagnostických grafů**.

### Osnova:

- datový soubor
- bodové a intervalové rozložení četností
- typy znaků, číselné charakteristiky znaků
- krabicový diagram, N-P plot, P-P plot, Q-Q plot, histogram

## Funkcionální charakteristiky datového souboru

### Označení

Na množině objektů  $\xi \dots \xi$  zjišťujeme hodnoty znaku  $X$  (např. u 6 domácností zjišťujeme počet členů).  
Hodnotu znaku  $X$  na objektu  $\xi$  označíme  $x_i$ ,  $i = 1, \dots, n$ .

Tyto hodnoty zaznameneáme do **jednorozměrného datového souboru**  $\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$  (např.  $\begin{pmatrix} 2 \\ 1 \\ 2 \\ 3 \\ 1 \\ 2 \end{pmatrix}$ ).

Uspořádané hodnoty  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$  tvoří **uspořádaný datový soubor**  $\begin{pmatrix} X_{(1)} \\ \vdots \\ X_{(n)} \end{pmatrix}$ , v našem případě  $\begin{pmatrix} 1 \\ 1 \\ 2 \\ 2 \\ 2 \\ 3 \end{pmatrix}$ .

Vektor  $\begin{pmatrix} x_{[1]} \\ \vdots \\ x_{[r]} \end{pmatrix}$ , kde  $x_{[1]} < \dots < x_{[r]}$  jsou navzájem různé hodnoty znaku  $X$ , se nazývá **vektor variant**, v našem případě  $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ .

## Bodové rozložení četností

Je-li počet variant znaku  $X$  malý, přiřazujeme četnosti jednotlivým variantám a hovoříme o bodovém rozložení četností.

$n_j$  – absolutní četnost varianty  $x_{[j]}$

$p_j = \frac{n_j}{n}$  – relativní četnost varianty  $x_{[j]}$

$N_j = n_1 + \dots + n_j$  – absolutní kumulativní četnost prvních  $j$  variant

$F_j = \frac{N_j}{n} = p_1 + \dots + p_j$  – relativní kumulativní četnost prvních  $j$  variant

Absolutní a relativní četnosti zapisujeme do tabulky rozložení četností nebo je znázorňujeme graficky např. pomocí sloupkového diagramu či polygonu četností.

Četnostní funkce:  $p(x) = \begin{cases} p_j & \text{pro } x = x_{[j]}, j = 1, \dots, r \\ 0 & \text{jinak} \end{cases}$

Empirická distribuční funkce:  $F(x) = \begin{cases} 0 & \text{pro } x < x_{[1]} \\ F_j & \text{pro } x_{[j]} \leq x < x_{[j+1]}, j = 1, \dots, r-1 \\ 1 & \text{pro } x \geq x_{[r]} \end{cases}$

**Příklad 1.:** U 30 domácností byl zjišťován počet členů.

Počet členů	1	2	3	4	5	6
Počet domácností	2	6	4	10	5	3

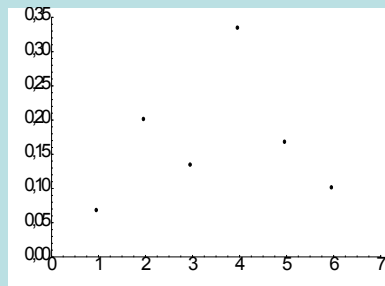
Vytvořte tabulku rozložení četností. Nakreslete grafy četnostní funkce a empirické distribuční funkce. Dále nakreslete sloupkový diagram a polygon četností počtu členů domácností.

**Řešení:**

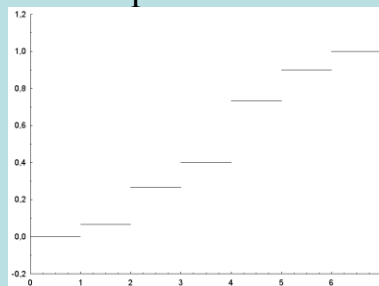
Tabulka rozložení četností

$x_{[j]}$	$n_j$	$p_i$	$N_i$	$F_i$
1	2	2/30	2	2/30
2	6	6/30	8	8/30
3	4	4/30	12	12/30
4	10	10/30	22	22/30
5	5	5/30	27	27/30
6	3	3/30	30	1

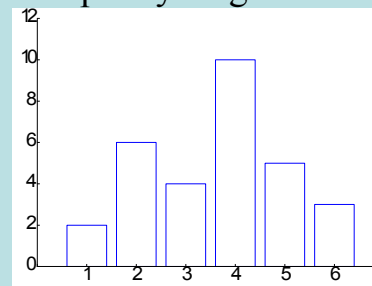
Graf četnostní funkce



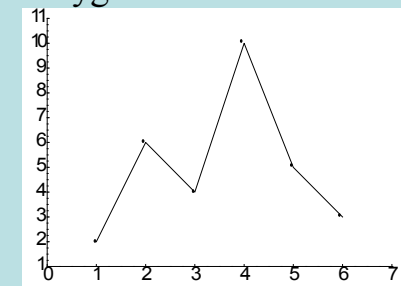
Graf empirické distribuční funkce



Sloupkový diagram



Polygon četností



## Intervalové rozložení četností

Je-li počet variant znaku  $X$  velký, přiřazujeme četnosti nikoli jednotlivým variantám, ale třídícím intervalům  $[u_{j-1}, u_j)$ , ...,  $[u_{r-1}, u_r)$  a hovoříme o intervalovém rozložení četností. Názvy četností jsou podobné jako u bodového rozložení četností, navíc zavádíme **četnostní hustotu**  $j$ -tého třídícího intervalu  $f_j = \frac{p_j}{d_j}$ , kde  $d_j = u_{j+1} - u_j$ . Stanovení počtu třídících intervalů je dosti subjektivní záležitost. Často se doporučuje volit  $r$  blízké  $\sqrt{n}$ .

**Hustota četnosti:**  $f(x) = \begin{cases} f_j & \text{pro } u_{j-1} < x \leq u_j, j=1, \dots, r \\ 0 & \text{jinak} \end{cases}$  (grafem hustoty četnosti je histogram)

**Intervalová empirická distribuční funkce:**  $F(x) = \sum_{t \leq x} f(t)$ .

**Příklad 2.:** U 70 domácností byly zjišťovány týdenní výdaje na nealkoholické nápoje (v Kč).

Výdaje	3,56\$	6,99\$	9,512\$	12,55\$	15,58\$	18,21\$
Počet dom.	7	16	27	14	4	2

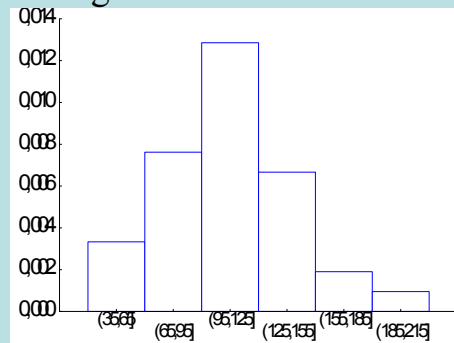
Sestavte tabulku rozložení četností, nakreslete histogram a graf intervalové empirické distribuční funkce.

**Řešení:**

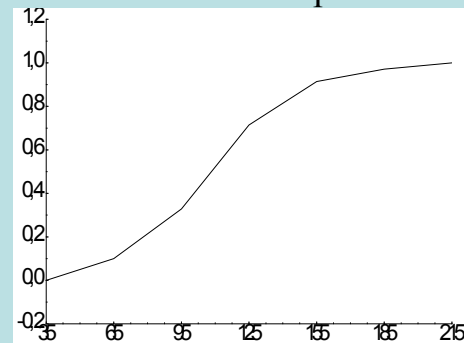
Tabulka rozložení četností

$(u_j, u_{j+1}]$	$n_j$	$p_j$	$f_j$	$N_j$	$F_j$
3,56\$	7	7/70	7/2100	7	7/70
6,99\$	16	16/70	16/2100	23	23/70
9,512\$	27	27/70	27/2100	50	50/70
12,55\$	14	14/70	14/2100	64	64/70
15,58\$	4	4/70	4/2100	68	68/70
18,21\$	2	2/70	2/2100	70	1

Histogram



Graf intervalové empirické distribuční funkce



## Číselné charakteristiky datového souboru

### Znaky nominálního typu

Tyto znaky umožňují obsahovou interpretaci pouze u relace rovnosti.

Příklady nominálních znaků: lékařská diagnóza, typ profese, barva očí, rodinný stav, národnost, ...

Charakteristikou polohy je **modus**, tj. nejčetnější varianta či střed nejčetnějšího intervalu.

### Znaky ordinálního typu

Lze u nich navíc obsahově interpretovat relaci uspořádání.

Příklad ordinálního znaku: školní klasifikace vyjadřuje menší nebo větší znalosti zkušných žáků – jedničkář je lepší než dvojkař, ale intervaly mezi známkami nemají obsahovou interpretaci. Nelze tvrdit, že rozdíl ve znalostech mezi jedničkářem a dvojkařem je stejný jako mezi trojkařem a čtyřkařem.

Další příklady: Různá bodování ve sportovních a uměleckých soutěžích, posuzování různých rysů sociálního chování, posuzování stavu pacientů, hodnocení postojů respondentů k různým otázkám, ...

Charakteristikou polohy je  **$\alpha$ -kvantil**. Je-li  $\alpha \in [0, 1]$ , pak  $\alpha$ -kvantil  $x_\alpha$  je číslo, které rozděluje uspořádaný datový soubor na dolní úsek, obsahující aspoň podíl  $\alpha$  všech dat a na horní úsek obsahující aspoň podíl  $1 - \alpha$  všech dat. Pro výpočet  $\alpha$ -kvantilu slouží algoritmus:

$$n\alpha = \begin{cases} \text{celé číslo} \Rightarrow x_{(c)} \\ \text{ne celé číslo} \Rightarrow \text{okrouhlá hodnota nejblížešlého číslo} \Rightarrow x_{(c)} \end{cases}$$

Pro speciálně zvolená  $\alpha$  užíváme názvů:

$x_{0,50}$  – **medián**,  $x_{0,25}$  – **dolní kvartil**,  $x_{0,75}$  – **horní kvartil**,  $x_{0,1}, \dots, x_{0,9}$  – **decily**,  $x_{0,01}, \dots, x_{0,99}$  – **percentily**.

Jako charakteristika variability slouží **kvartilová odchylka**:  $q = x_{0,75} - x_{0,25}$ .



**Příklad 3:** Během semestru se studenti podrobili písemnému testu z matematiky, v němž bylo možno získat 0 až 10 bodů. Výsledky jsou uvedeny v tabulce:

Počet bodů	0	1	2	3	4	5	6	7	8	9	10
Počet studentů	1	4	6	7	11	15	19	17	12	6	3

Zjistěte modus, medián, 1. decil, 9. decil a kvartilovou odchylku počtu bodů.

### Řešení:

Modus je nejčetnější varianta znaku, v tomto případě tedy 6.

Pro výpočet kvantilů musíme znát rozsah datového souboru:  $n = 1 + 4 + \dots + 3 = 101$ . Výpočty uspořádáme do tabulky.

$\alpha$	$n\alpha$	c	$x_\alpha = X_{(c)}$
0,50	50,5	51	6
0,10	10,1	11	2
0,90	90,9	91	8
0,25	25,25	26	4
0,75	75,75	76	7

$$q = 7 - 4 = 3$$

### Výpočet pomocí systému STATISTICA:

Otevřeme nový datový soubor o 2 proměnných a 11 případech. První proměnnou nazveme X, druhou četnost a zapíšeme do nich počet bodů a odpovídající absolutní četnosti.

Statistiky – Základní statistiky/tabulky – Popisné statistiky – zapneme proměnnou vah četnost – OK – OK – Proměnné X – OK – Detailní výsledky – vybereme Medián, Dolní a horní kvartily, Kvantilové hranice – Výpočet – ve výstupní tabulce upravíme počet desetinných míst.

Promě	Popisné statistiky (pocet bodu.sta)					
	N platn	Mediá	Spod kvart	Horn kvart	Kvani 10,00	Kvani 90,00
X	10	6	2	7	2	8

## Znaky intervalového a poměrového typu

U těchto znaků lze navíc obsahově interpretovat operaci rozdílu resp. podílu.

Příklad intervalového znaku: teplota měřená ve stupních Celsia. Např. naměříme-li ve čtyřech po sobě jdoucích dnech polední teploty 0, 2, 4, 6 °C, znamená to, že každým dnem stouply teploty o 2 °C. Nelze však říci, že z druhého na třetí den vzrostla teplota dvojnásobně, kdežto ze třetího na čtvrtý den pouze jeden a půl krát.

Další příklady: kalendářní systémy, směr větru, inteligenční kvocient, ...

Společný znak intervalových znaků: nula byla stanovena uměle, pouhou konvencí.

Příklad poměrového znaku: délka předmětu měřená v cm. Má-li jeden předmět délku 8 cm a druhý 16 cm, má smysl prohlásit, že druhý předmět je dvakrát delší než první předmět.

Další příklady: počet dětí v rodině, výška kapesného v Kč, hmotnost osoby, ...

Společný znak poměrových znaků: poměrový znak má přirozený počátek, ke kterému jsou vztahovány všechny další hodnoty znaku.

Charakteristika polohy: **aritmetický průměr**  $m = \frac{1}{n} \sum_{i=1}^n x_i$ .

U poměrových znaků, které nabývají pouze kladných hodnot, lze použít **geometrický průměr**  $\sqrt[n]{x_1 \cdot \dots \cdot x_n}$ .

Pomocí průměru zavedeme **i-tou centrovanou hodnotu**  $x_i - m$  (podle znaménka poznáme, zda i-tá hodnota je podprůměrná či nadprůměrná).

Znázornění rozložení četností dvou datových souborů, které se liší aritmetickým průměrem



## Vlastnosti aritmetického průměru

- Aritmetický průměr si lze představit jako těžiště dat – součet podprůměrných hodnot je stejný jako součet nadprůměrných hodnot – oba součty jsou v rovnováze.
- Průměr centrovaných hodnot je nulový, protože  $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) = \frac{1}{n} \sum_{i=1}^n x_i - \frac{1}{n} \sum_{i=1}^n \bar{x} = \bar{x} - \bar{x} = 0$ .
- Výraz  $\sum_{i=1}^n (x_i - a)^2$  (tzv. kvadratická odchylka) nabývá svého minima pro  $a = m$ . Uvedený výraz charakterizuje celkovou chybu, které se dopustíme, když datový soubor nahradíme jedinou hodnotou  $a$ . Tato chyba je tedy nejmenší, když datový soubor nahradíme aritmetickým průměrem, přičemž za míru chyby považujeme kvadratickou odchylku.
- Pokud každou hodnotu  $x_i$  podrobíme lineární transformaci  $y_i = a + bx_i$ , pak průměr transformovaných hodnot je roven lineární transformaci původního průměru, tj.  $m_2 = a + bm_1$ .
- Mají-li znaky  $X, Y$  průměry  $m_1, m_2$ , pak znak  $Z = X + Y$  má průměr  $m_1 + m_2$ .
- Aritmetický průměr je silně ovlivněn extrémními hodnotami.
- Aritmetický průměr je vhodné použít, pokud je rozložení dat přibližně symetrické.

### Příklad na vlastnosti aritmetického průměru:

U skupiny 20 pracovníků v určité dílně byly zjišťovány měsíční mzdy. Průměr mezd činil 15 500 Kč. Určete průměr mezd, jestliže mzdy všech pracovníků se zvýší

a) o 300 Kč, b) 1,1 krát, c) o 20%.

### Řešení:

Označme  $m_1$  průměr hodnot  $x_1, \dots, x_n$  a  $m_2$  průměr hodnot  $y_1, \dots, y_n$ , přičemž  $y_i = a + bx_i$ ,  $i = 1, \dots, n$ . Pak  $m_2 = a + bm_1$ .

ad a)  $m_2 = 300 + m_1 = 15\ 800$

Průměr se zvýšil o 300 Kč na 15 800 Kč.

ad b)  $m_2 = 1,1 \cdot m_1 = 17\ 050$

Průměr se zvýšil na 17 050 Kč.

ad c)  $m_2 = 1,2 \cdot m_1 = 18\ 600$

Průměr se zvýšil na 18 600 Kč.

## Charakteristiky variability intervalových a poměrových znaků

**Variační rozpětí**  $R = x_{(n)} - x_{(1)}$  (nevýhoda – bere v úvahu pouze nejmenší a největší hodnotu datového souboru),

**průměrná odchylka**  $O = \frac{1}{n} \sum_{i=1}^n |x_i - m|$  (udává, o kolik jednotek se data liší od průměru)

**rozptyl**  $s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - m)^2$  (nevýhoda – vychází ve druhých mocninách jednotek, v nichž byl měřen znak X)

**směrodatná odchylka**  $s = \sqrt{s^2}$ .

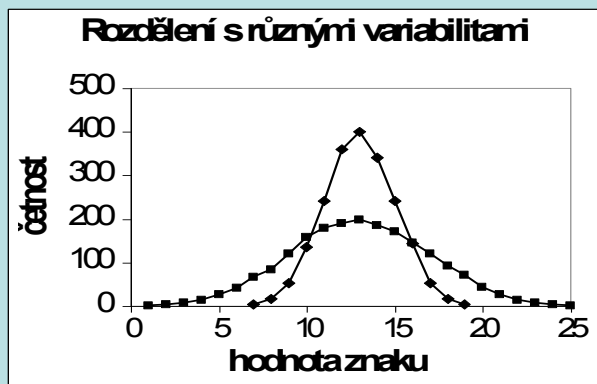
Pomocí směrodatné odchylky zavedeme **i-tou standardizovanou hodnotu**  $\frac{x_i - m}{s}$  (vyjadřuje, o kolik směrodatných odchylek se i-tá hodnota odchytila od průměru).

U poměrových znaků se jako charakteristika variability používá též:

**koeficient variace**  $\frac{s}{m}$  (často se udává v procentech a udává, kolika procent průměru dosahuje směrodatná odchylka),

**relativní průměrná odchylka**  $\frac{O}{m}$  (při vyjádření v procentech udává, kolika procent průměru dosahuje průměrná odchylka)

Znázornění rozložení četností dvou datových souborů, které se liší rozptylem:



### Vlastnosti rozptylu:

- Rozptyl je nulový pouze tehdy, když jsou všechny hodnoty stejné, jinak je kladný.

- Rozptyl centrovaných hodnot je roven původnímu rozptylu, neboť  $\frac{1}{n} \sum_{i=1}^n (x_i - m)^2 = \frac{1}{n} \sum_{i=1}^n (x_i - m + 0)^2 = \frac{1}{n} \sum_{i=1}^n (x_i - m)^2 + \frac{1}{n} \sum_{i=1}^n 0^2 = \frac{1}{n} \sum_{i=1}^n (x_i - m)^2$ .

- Rozptyl standardizovaných hodnot je 1, protože  $\frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - m}{s} - 0 \right)^2 = \frac{1}{s^2} \cdot \frac{1}{n} \sum_{i=1}^n (x_i - m)^2 = \frac{s^2}{s^2} = 1$ .

- Rozptyl se zpravidla počítá podle vzorce  $s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - m)^2$ .

- Pokud každou hodnotu  $x_i$  podrobíme lineární transformaci  $y_i = a + bx_i$ , pak rozptyl transformovaných hodnot je roven původnímu rozptylu vynásobenému  $b^2$ , tj.  $s_2^2 = b^2 s_1^2$ .

- Rozptyl je stejně jako průměr silně ovlivněn extrémními hodnotami.

- Rozptyl se nehodí jako charakteristika variability, je-li rozložení dat nesymetrické.

**Příklad 4:** Kurzy akcií společnosti AAA Auto Group v průběhu 23 dní v měsíci srpnu 2010 byly následující: 17,75; 17,74; 17,85; 17,59; 17,92; 17,98; 18,39; 18,25; 18,30; 18,00; 18,15; 18,15; 18,22; 18,40; 18,25; 17,95; 18,25; 18,23; 17,95; 17,90; 17,80; 17,87; 17,87. Vypočtěte charakteristiky variability.

### Řešení:

Nejprve vypočítáme variační rozpětí:  $R = x_{(n)} - x_{(1)} = 18,39 - 17,59 = 0,8$ .

Před výpočtem dalších charakteristik variability musíme získat aritmetický průměr:  $m = \frac{1}{n} (17,75 + 17,74 + \dots + 17,87) = 18,03$ .

Průměrná odchylka:  $O = \frac{1}{n} \sum_{i=1}^n |x_i - m| = \frac{1}{23} (17,75 - 18,03 + 17,74 - 18,03 + \dots + 17,87 - 18,03) = 0,196$  €

Relativní průměrná odchylka:  $\frac{O}{m} 100\% = \frac{0,196}{18,03} 100\% = 1,1\%$

Rozptyl:  $S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - m)^2 = \frac{1}{23} (17,75 - 18,03)^2 + (17,74 - 18,03)^2 + \dots + (17,87 - 18,03)^2 = 0,04$

Směrodatná odchylka:  $S = \sqrt{0,04} = 0,221$

Koeficient variace:  $\frac{S}{m} 100\% = \frac{0,221}{18,03} 100\% = 1,2\%$

## Výpočet pomocí systému STATISTICA:

Otevřeme nový datový soubor o jedné proměnné X a 23 případech. Do proměnné X zapíšeme zjištěné kurzy akcií. Statistiky – Základní statistiky/tabulky – Popisné statistiky – OK – Proměnné X – OK – Detailní výsledky – vybereme Průměr, Rozptyl, Rozpětí – Výpočet. Ve výstupní tabulce přidáme za proměnnou Rozptyl tři nové proměnné nazvané rozptyl, směr. odch. a koef. variace. Do Dlouhého jména proměnné rozptyl napíšeme  $=v3*22/23$ , Dlouhého jména proměnné směr. odch. napíšeme  $=\text{sqrt}(v4)$  a do Dlouhého jména proměnné koef. variace napíšeme  $=100*v5/v1$ .

Promě	Prum	Rozp	Rozp	rozpt $=v3*2$	směr. o $=\text{sqrt}(v4)$	koef. var $=100*v5$
x	18,03	0,810	0,051	0,049	0,22136	1,22751

Pro výpočet průměrné odchyly a relativní průměrné odchyly je zapotřebí přidat k původnímu datovému souboru dvě nové proměnné nazvané Průměr a Odchylna. Do Dlouhého jména proměnné Průměr napíšeme  $=18,033$  a do Dlouhého jména proměnné Odchylna napíšeme  $=\text{abs}(v1-v2)$ . Nyní spočteme průměr proměnné Odchylna: Statistiky – Základní statistiky/tabulky – Popisné statistiky – OK – Proměnné Odchylna – OK – Detailní výsledky – vybereme Průměr – Výpočet. Ve výstupní tabulce přejmenujeme proměnnou Průměr na prům. odch. a za tuto proměnnou přidáme proměnnou rel. prům. odch. Do jejího Dlouhého jména napíšeme  $=100*v1/18,033$ .

Promě	odchy	rel. prům. $=100*v1/v1$
Odchy	0,196	1,0895



## Vážené číselné charakteristiky

Známe-li absolutní četnosti  $n_1, \dots, n_r$  či relativní četnosti  $p_1, \dots, p_r$  variant  $x_{[1]}, \dots, x_{[r]}$ , můžeme spočítat

vážený průměr  $m = \frac{1}{n} \sum_{j=1}^r n_j x_{[j]} = \sum_{j=1}^r p_j x_{[j]}$ ,

vážený rozptyl  $s^2 = \frac{1}{n} \sum_{j=1}^r n_j x_{[j]}^2 - m^2 = \sum_{j=1}^r p_j x_{[j]}^2 - m^2$  (výpočetní vzorec:  $s^2 = \frac{1}{n} \sum_{j=1}^r n_j x_{[j]}^2 - \left( \frac{1}{n} \sum_{j=1}^r n_j x_{[j]} \right)^2$ ),

váženou průměrnou odchylku  $\sigma = \sqrt{\frac{1}{n} \sum_{j=1}^r n_j x_{[j]}^2 - m^2} = \sqrt{\sum_{j=1}^r p_j x_{[j]}^2 - m^2}$ .

**Příklad 5:** U 35 zaměstnanců byl zjištěn počet odpracovaných hodin za měsíc.

Počet odpracovaných hodin	184	185	186	187	188	189
Počet zaměstnanců	4	6	7	6	7	5

Vypočítejte průměr, průměrnou odchylku, relativní průměrnou odchylku, směrodatnou odchylku a koeficient variace počtu odpracovaných hodin.

**Řešení:**

$$\text{Vážený průměr: } \bar{m} = \frac{1}{n} \sum_{i=1}^k x_i \cdot h_i = \frac{1}{35} \cdot (4 \cdot 184 + 6 \cdot 185 + 7 \cdot 186 + 6 \cdot 187 + 7 \cdot 188 + 5 \cdot 189) = 186,6$$

Vážená průměrná odchylka:

$$o = \frac{1}{n} \sum_{i=1}^k |x_i - \bar{m}| \cdot h_i = \frac{1}{35} \cdot (4 \cdot |184 - 186,6| + 6 \cdot |185 - 186,6| + 7 \cdot |186 - 186,6| + 6 \cdot |187 - 186,6| + 7 \cdot |188 - 186,6| + 5 \cdot |189 - 186,6|) = 0,74 \text{ h} = 1 \text{ h } 23 \text{ min}$$

$$\text{Vážený rozptyl: } s^2 = \frac{1}{n} \sum_{i=1}^k x_i^2 \cdot h_i - \bar{m}^2 = \frac{1}{35} \cdot (4 \cdot 184^2 + 6 \cdot 185^2 + 7 \cdot 186^2 + 6 \cdot 187^2 + 7 \cdot 188^2 + 5 \cdot 189^2) - 186,6^2 = 2525$$

$$\text{Vážená směrodatná odchylka: } s = \sqrt{2525} = 50,25 \text{ h} = 1 \text{ h } 35 \text{ min}$$

$$\text{Relativní průměrná odchylka: } \frac{o}{\bar{m}} \cdot 100\% = \frac{0,74}{186,6} \cdot 100\% = 0,39\%$$

$$\text{Koeficient variace: } \frac{s}{\bar{m}} \cdot 100\% = \frac{50,25}{186,6} \cdot 100\% = 26,9\%$$

Vidíme, že zaměstnanci odpracovali za měsíc v průměru 186,6 h, přičemž průměrná odchylka dosahuje 0,39 % průměrné odpracované doby a směrodatná odchylka dosahuje 26,9 % průměrné odpracované doby.

## Výpočet pomocí systému STATISTICA:

Otevřeme nový datový soubor o 2 proměnných a 6 případech. První proměnnou nazveme X, druhou četnost a zapíšeme do nich počet odpracovaných hodin a odpovídající počty zaměstnanců.

Statistiky – Základní statistiky/tabulky – Popisné statistiky – zapneme proměnnou vah četnost – OK – OK – Proměnné X – OK – Detailní výsledky – vybereme Průměr, Rozptyl – Výpočet. Ve výstupní tabulce přidáme za proměnnou Rozptyl dvě nové proměnné nazvané směr. odch. a koef. variace. Do Dlouhého jména proměnné směr. odch. napíšeme  $=\sqrt{v2*34/35}$  a do Dlouhého jména proměnné koef. variace napíšeme  $=100*v3/v1$ .

Promě	Prum	Rozp	směr.od $=\sqrt{v2*34/35}$	koef. var $=100*v3/v1$
X	186,6	2,	1,5892	0,85168

Pro výpočet průměrné odchyly a relativní průměrné odchyly je zapotřebí přidat k původnímu datovému souboru dvě nové proměnné nazvané Průměr a Odchylna. Do Dlouhého jména proměnné Průměr napíšeme  $=186,6$  a do Dlouhého jména proměnné Odchylna napíšeme  $=\text{abs}(v1-v3)$ . Nyní spočteme průměr proměnné Odchylna: Statistiky – Základní statistiky/tabulky – Popisné statistiky – zapneme proměnnou vah četnost – OK – OK – Proměnné Odchylna – OK – Detailní výsledky – vybereme Průměr – Výpočet. Ve výstupní tabulce přejmenujeme proměnnou Průměr na prům. odch. a za tuto proměnnou přidáme proměnnou rel. prům. odch. Do jejího Dlouhého jména napíšeme  $=100*v1/186,6$ .

Promě	prum odch	rel. prům odch $=100*v1/v3$
Odchy	1,382	0,74108

Převod desetinných částí hodiny na minuty můžeme provést např. pomocí aplikace na adrese <http://www.prevody-jednotek.cz/>.

## Počáteční a centrální momenty

Aritmetický průměr a rozptyl jsou speciální případy momentů. Zavedeme

**k-tý počáteční moment**  $m_k = \frac{1}{n} \sum_{i=1}^n x_i^k$ ,  $k = 1, 2, \dots$ ,

**k-tý centrální moment**

$m_k = \frac{1}{n} \sum_{i=1}^n (x_i - m)^k$ ,  $k = 1, 2, \dots$

Pomocí 3. a 4. počátečního momentu se definuje šikmost a špičatost.

**Šikmost:**  $\alpha_3$  - měří nesouměrnost rozložení četností kolem průměru.

Je-li rozložení dat symetrické kolem aritmetického průměru, pak  $\alpha_3 = 0$ .

Má-li rozložení dat prodloužený pravý konec, jde o **kladně zešikmené rozložení**,  $\alpha_3 > 0$ .

Má-li rozložení dat prodloužený levý konec, jde o **záporně zešikmené rozložení**,  $\alpha_3 < 0$ .

Znázornění rozložení četností dvou datových souborů, které se liší aritmetickým průměrem a šikmostí



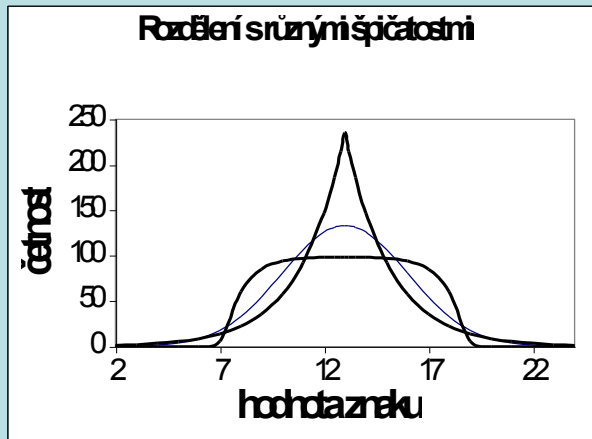
**Špičatost:**  $\alpha_4$  - měří koncentraci rozložení četností kolem průměru.

Je-li rozložení dat normální (Gaussovo), pak  $\alpha_4 = 0$ .

Je-li rozložení dat strmé, pak  $\alpha_4 > 0$ .

Je-li rozložení dat ploché, pak  $\alpha_4 < 0$ .

Znázornění rozložení četností dvou datových souborů, které se liší špičatostí

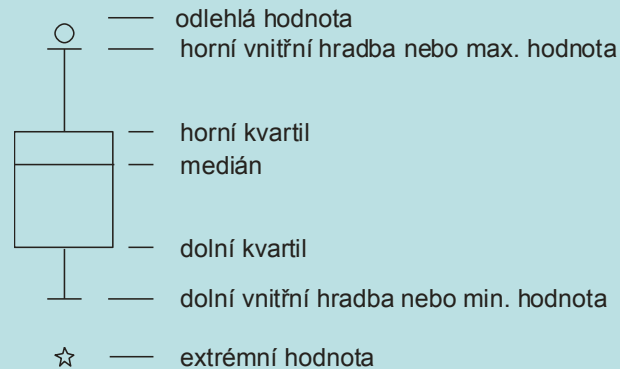


## Diagnostické grafy

### Krabicový diagram

Umožňuje posoudit symetrii a variabilitu datového souboru a existenci odlehlých či extrémních hodnot.

Způsob konstrukce



**Odlehlá hodnota** leží mezi **vnějšími a vnitřními hradbami**, tj. v intervalu

$(x_{0,75} + 1,5q, x_{0,75} + 3q)$  či v intervalu  $(x_{0,25} - 3q, x_{0,25} - 1,5q)$ .

**Extrémní hodnota** leží za vnějšími hradbami, tj. v intervalu  $(x_{0,75} + 3q, \infty)$  či v intervalu  $(-\infty, x_{0,25} - 3q)$ .

**Příklad 6:** Pro údaje z příkladu 1 sestrojte krabicový diagram.

**Řešení:**

Počet členů	1	2	3	4	5	6
Počet domácností	2	6	4	10	5	3

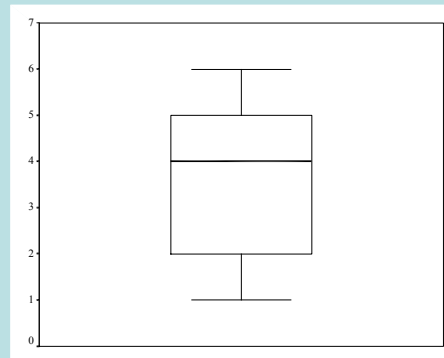
Rozsah souboru  $n = 30$ . Výpočty potřebných kvantilů uspořádáme do tabulky.

$\alpha$	$n\alpha$	c		$x_\alpha$
0,25	7,5	8	$x_{(c)}=x_{(8)}$	2
0,50	15	15	$\frac{x_{(15)} + x_{(16)}}{2}$	4
0,75	22,5	23	$x_{(c)}=x_{(23)}$	5

$$q = 5 - 2 = 3$$

Dolní vnitřní hradba:  $x_{0,25} - 1,5q = 2 - 1,5 \cdot 3 = -2,5$

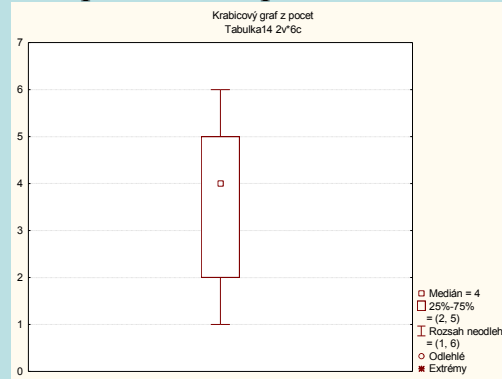
Horní vnitřní hradba:  $x_{0,75} + 1,5q = 5 + 1,5 \cdot 3 = 9,5$



Vidíme, že datový soubor vykazuje určitou nesymetrii – medián je posunut směrem k hornímu kvartilu, soubor je tedy záporně sešikmen. V souboru se nevyskytují žádné odlehle ani extrémní hodnoty.

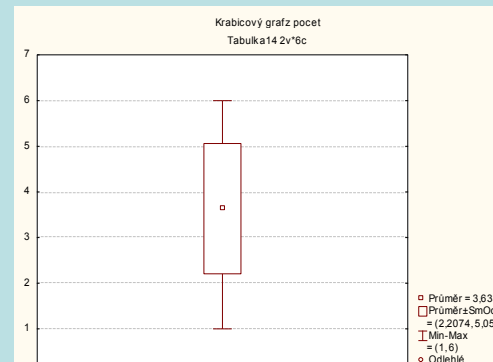
## Výpočet pomocí systému STATISTICA:

Otevřeme nový datový soubor o 2 proměnných a 6 případech. První proměnnou nazveme počet, druhou četnost a zapíšeme do nich počet členů domácnosti a odpovídající absolutní četnosti. Zvolíme Grafy – 2D Grafy – Krabicové grafy. Zapneme proměnnou vah četnost, zadáme závisle proměnnou pocet a dostaneme krabicový diagram:



**Upozornění:** Máme-li data intervalového či poměrového charakteru, o nichž lze předpokládat, že pocházejí z nějakého symetrického rozložení (například normálního), je možné použít jinou variantu krabicového diagramu: bod či čára uvnitř krabice reprezentuje průměr, vodorovné hrany krabice jsou ve výšce průměr  $\pm$  směrodatná odchylka a svorky končí v minimu či maximu.

V našem případě dostaneme krabicový diagram:





Před uvedením dalších diagnostických grafů je nutné zavést pojem pořadí čísla v posloupnosti čísel.

### Pojem pořadí

Nechť  $x_1, \dots, x_n$  je posloupnost reálných čísel.

- Jsou-li čísla navzájem různá, pak pořadím  $R_i$  čísla  $x_i$  rozumíme počet těch čísel  $x_1, \dots, x_n$ , která jsou menší nebo rovna číslu  $x_i$ .
- Vyskytují-li se mezi danými čísly skupinky stejných čísel, pak každé takové skupince přiřadíme průměrné pořadí.

### Příklad na stanovení pořadí

- Jsou dána čísla 9, 4, 5, 7, 3, 1. Stanovte pořadí těchto čísel.
- Jsou dána čísla 6, 7, 7, 9, 6, 10, 8, 6, 6, 9.

### Řešení

ad a)

usp. čísla	1	3	4	5	7	9
pořadí	1	2	3	4	5	6

ad b)

usp. čísla	6	6	6	6	7	7	8	9	9	10
pořadí	1	2	3	4	5	6	7	8	9	10
prům. pořadí	2,5	2,5	2,5	2,5	5,5	5,5	7	8,5	8,5	10

## Normální pravděpodobnostní graf (N-P plot)

N- P plot umožňuje graficky posoudit, zda data pocházejí z normálního rozložení.

### Způsob konstrukce:

Na vodorovnou osu vynášíme uspořádané hodnoty  $x_{(1)} \leq \dots \leq x_{(n)}$ ,

na svislou osu kvantily  $U_{\alpha}$  standardizovaného normálního rozložení, kde  $\alpha = \frac{j}{n+1}$ , přičemž  $j$  je pořadí  $j$ -té uspořádané hodnoty (jsou-li některé hodnoty stejné, pak za  $j$  bereme průměrné pořadí odpovídající takové skupince).

Pocházejí-li data z normálního rozložení, pak všechny dvojice  $(x_{(j)}, U_{\alpha})$  budou ležet na přímce.

Pro data z rozložení s kladnou šikmostí se dvojice  $(x_{(j)}, U_{\alpha})$  budou řadit do **konkávní křivky**,

pro data z rozložení se zápornou šikmostí se dvojice  $(x_{(j)}, U_{\alpha})$  budou řadit do **konvexní křivky**.

### Příklad na konstrukci N – P plotu:

Desetkrát nezávisle na sobě byla změřena jistá konstanta. Výsledky měření: 2 1,8 2,1 2,4 1,9 2,1 2 1,8 2,3 2,2. Pomocí normálního pravděpodobnostního grafu posuďte, zda se tato data řídí normálním rozložením.

### Řešení:

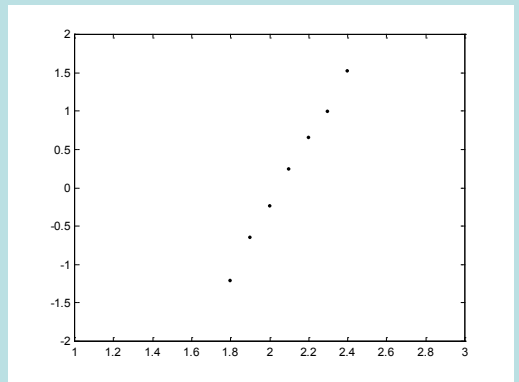
usp. hodnoty	1,8	1,8	1,9	2	2	2,1	2,1	2,2	2,3	2,4
pořadí	1	2	3	4	5	6	7	8	9	10
průměrné pořadí	1,5	1,5	3	4,5	4,5	6,5	6,5	8	9	10

Vektor hodnot průměrného pořadí:  $j = (1,5 \ 3 \ 4,5 \ 6,5 \ 8 \ 9 \ 10)$ ,

vektor hodnot  $\alpha : \begin{pmatrix} 1 \\ 1 \\ 2 \\ 2 \\ 5 \\ 8 \\ 10 \\ 10 \\ 3 \\ 5 \\ 5 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 2 \\ 2 \\ 5 \\ 8 \\ 10 \\ 10 \\ 3 \\ 5 \\ 5 \end{pmatrix}$ ,

vektor kvantilů  $u_{\alpha} = \begin{pmatrix} -2,1 \\ -1,1 \\ -0,49 \\ -0,24 \\ 0,24 \\ 0,64 \\ 0,98 \\ 1,17 \end{pmatrix}$ .

Normální pravděpodobnostní graf

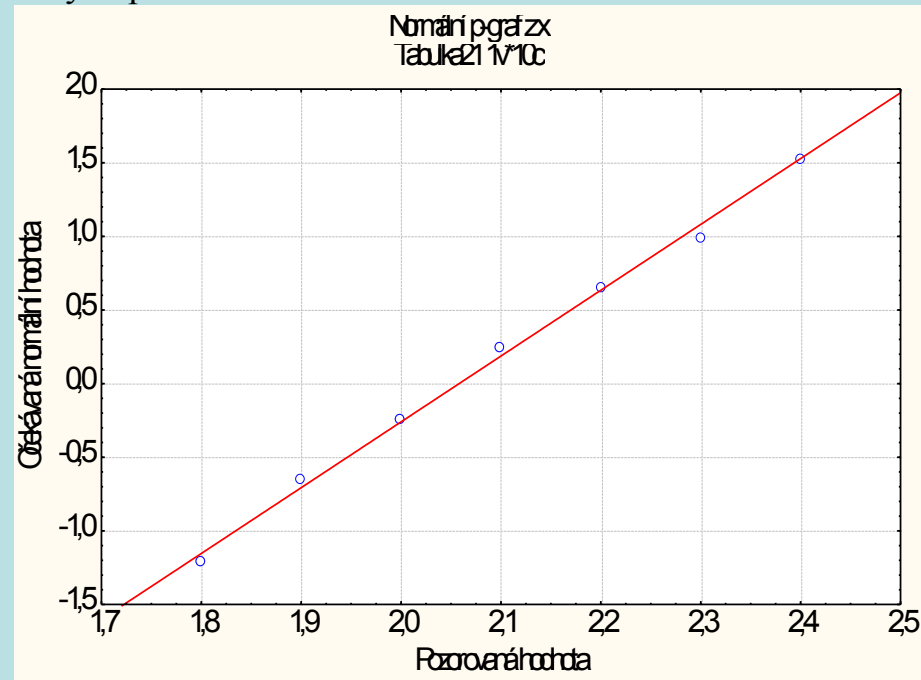


Protože dvojice  $(X_{(j)}, u_{\alpha})$  téměř leží na přímce, lze usoudit, že data pocházejí z normálního rozložení.

## Výpočet pomocí systému STATISTICA:

Otevřeme nový datový soubor o jedné proměnné a 10 případech. Zjištěné hodnoty zapíšeme do proměnné X.

Grafy – 2D Grafy – Normální pravděpodobnostní grafy – Proměnná X – OK - odškrtneme Neurčovat průměrnou pozici svázaných pozorování - OK.



## Quantile - quantile plot (Q-Q plot)

Umožňuje graficky posoudit, zda data pocházejí z nějakého známého rozložení (např. STATISTICA nabízí 8 typů rozložení: beta, exponenciální, Gumbelovo, gamma, log-normální, normální, Rayleighovo a Weibulovo).

### Způsob konstrukce:

na svislou osu vynášíme uspořádané hodnoty  $x_{(1)} \leq \dots \leq x_{(n)}$ ,

na vodorovnou osu kvantily  $K_{\alpha}^{(X)}$  vybraného rozložení, kde  $\alpha = \frac{1 - i_j}{n_{adj}}$ , přičemž  $r_{adj}$  a  $n_{adj}$  jsou korigující faktory  $\leq 0,5$ ,

implicitně  $r_{adj} = 0,375$  a  $n_{adj} = 0,25$ . (Jsou-li některé hodnoty  $x_{(1)} \leq \dots \leq x_{(n)}$  stejné, pak za  $j$  bereme průměrné pořadí odpovídající takové skupince.)

Pokud vybrané rozložení závisí na nějakých parametrech, pak se tyto parametry odhadnou z dat nebo je může zadat uživatel.

Body  $(K_{\alpha}^{(X)}, x_{(j)})$  se metodou nejmenších čtverců proloží přímkou. Čím méně se body odchylují od této přímky, tím je lepší soulad mezi empirickým a teoretickým rozložením.

**Příklad na konstrukci Q-Q plotu:** Desetkrát nezávisle na sobě byla změřena jistá konstanta. Výsledky měření: 2,1 1,8 2,1 2,4 1,9 2,1 2,1 1,8 2,3 2,2. Pomocí Q-Q plotu ověřte, zda se tato data řídí normálním rozložením.

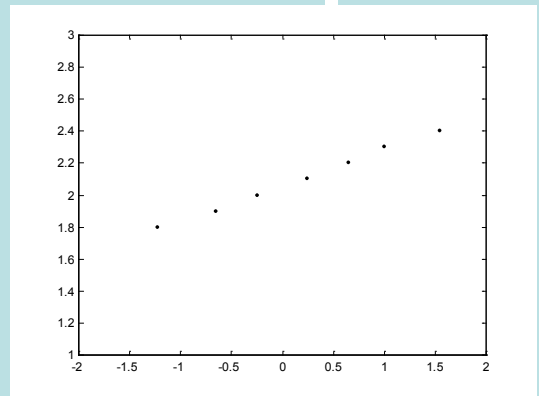
**Řešení:**

usp.hodnoty	1,8	1,8	1,9	2	2	2,1	2,1	2,2	2,3	2,4
pořadí	1	2	3	4	5	6	7	8	9	10
průměrné pořadí	1,5	1,5	3	4,5	4,5	6,5	6,5	8	9	10

Vektor hodnot průměrného pořadí:  $j = (1,5 \ 3 \ 4,5 \ 6,5 \ 8 \ 9 \ 10)$

vektor hodnot  $\alpha = (1 - \frac{1}{10}, \frac{1}{10} - \frac{1}{20}, \dots, \frac{1}{20} - \frac{1}{10}, \frac{1}{10}) = (0,9, 0,85, \dots, 0,1)$

vektor kvantilů  $u_\alpha = (-2,27, -1,55, -1,24, -1,02, -0,84, -0,7, -0,56, -0,4, -0,26)$

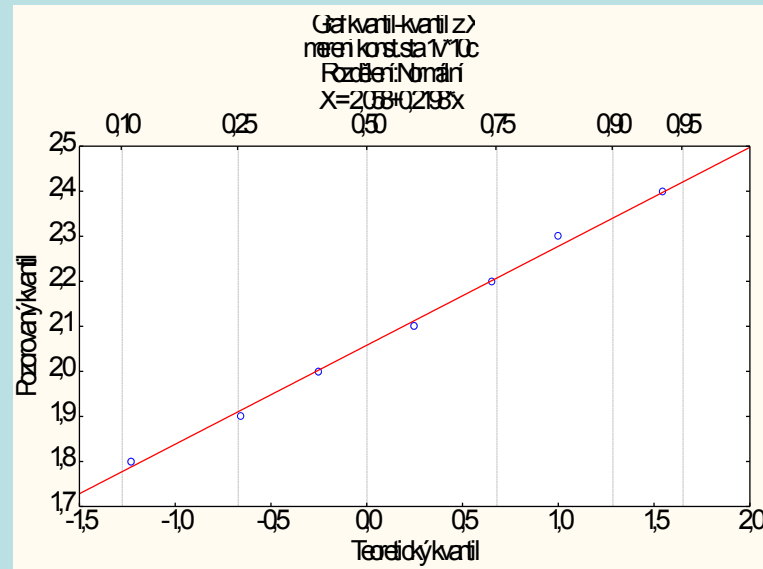


Vzhled grafu nasvědčuje tomu, že data pocházejí z normálního rozložení.

## Výpočet pomocí systému STATISTICA:

Otevřeme nový datový soubor o jedné proměnné a 10 případech. Zjištěné hodnoty zapíšeme do proměnné X.

Grafy – 2D Grafy – Grafy typu Q-Q– Proměnná X – OK - odškrtneme Neurčovat průměrnou pozici svázaných pozorování - OK.



## Probability - probability plot (P-P plot)

Používá se ke stejným účelům jako Q-Q plot, ale jinak se konstruuje.

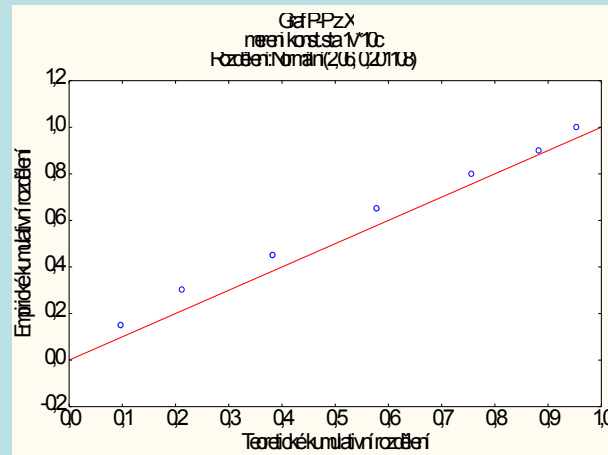
**Způsob konstrukce:** spočtou se standardizované hodnoty  $z_j = \frac{x_j - \bar{x}}{s}$ ,  $j = 1, \dots, n$ . Na vodorovnou osu se vynesou hodnoty teoretické distribuční funkce  $\Phi(z_{(j)})$  a na svislou osu hodnoty empirické distribuční funkce  $F(z_{(j)}) = j/n$ . (Jsou-li některé hodnoty  $x_{(1)} \leq \dots \leq x_{(n)}$  stejné, pak za  $j$  bereme průměrné pořadí odpovídající takové skupince.) Pokud se body  $(\Phi(z_{(j)}), F(z_{(j)}))$  řadí kolem hlavní diagonály čtverce  $[0,1] \times [0,1]$ , lze usuzovat na dobrou shodu empirického a teoretického rozložení.

**Příklad na konstrukci P-P plotu pomocí systému STATISTICA:** Desetkrát nezávisle na sobě byla změřena jistá konstanta. Výsledky měření: 2,1, 1,8, 2,1, 2,4, 1,9, 2,1, 2,1, 1,8, 2,3, 2,2. Pomocí P-P plotu ověřte, zda se tato data řídí normálním rozložením.

### Výpočet pomocí systému STATISTICA:

Otevřeme nový datový soubor o jedné proměnné a 10 případech. Zjištěné hodnoty zapíšeme do proměnné X.

Grafy – 2D Grafy – Grafy typu P-P – Proměnná X – OK - odškrtneme Neurčovat průměrnou pozici svázaných pozorování - OK.





## Histogram

Umožňuje porovnat tvar hustoty četnosti s tvarem hustoty pravděpodobnosti vybraného teoretického rozložení. (Ve STATISTICE je pojem histogramu širší, skrývá se za ním i sloupkový diagram.)

**Způsob konstrukce:** na vodorovnou osu vynášíme meze třídících intervalů. Nad každým třídícím intervalem sestrojíme obdélník o ploše odpovídající relativní četnosti příslušného třídícího intervalu, tj. výška obdélníku je rovna četnostní hustotě třídícího intervalu (četnostní hustota je relativní četnost třídícího intervalu dělená délkou tohoto intervalu).

**Způsob konstrukce ve STATISTICE:** na vodorovnou osu se vynášejí třídící intervaly (implicitně 10, jejich počet lze změnit, stejně tak i meze třídících intervalů) či varianty znaku a na svislou osu absolutní nebo relativní četnosti třídících intervalů či variant. Do histogramu se zakreslí tvar hustoty (či pravděpodobnostní funkce) vybraného teoretického rozložení.

### Příklad na konstrukci histogramu:

U 70 domácností byly zjišťovány týdenní výdaje na nealkoholické nápoje (v Kč).

Výdaje	35,65	65,95	95,25	125,55	155,85	185,25
Počet dom.	7	16	27	14	4	2

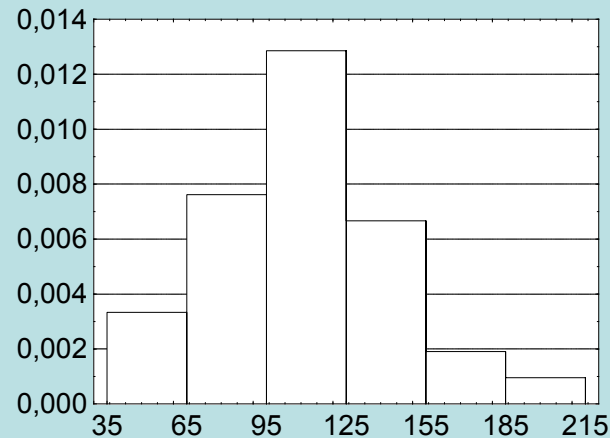
Nakreslete histogram.

### Řešení:

Nejprve sestavíme tabulku rozložení četností:

$u_j, u_{j+1}$	$x_{[j]}$	$d_j$	$n_j$	$p_j$	$N_j$	$F_j$	$f_j$
35,65	50	30	7	$7/70=0,1$	7	$7/70=0,1$	$7/2100=0,0033$
65,95	80	30	16	$16/70=0,23$	23	$23/70=0,33$	$16/2100=0,0076$
95,25	110	30	27	$27/70=0,38$	50	$50/70=0,71$	$27/2100=0,0129$
125,55	140	30	14	$14/70=0,2$	64	$64/70=0,91$	$14/2100=0,0067$
155,85	170	30	4	$4/70=0,06$	68	$68/70=0,97$	$4/2100=0,0019$
185,25	200	30	2	$2/70=0,03$	70	$70/70=1$	$2/2100=0,0010$

S pomocí této tabulky sestojíme histogram:

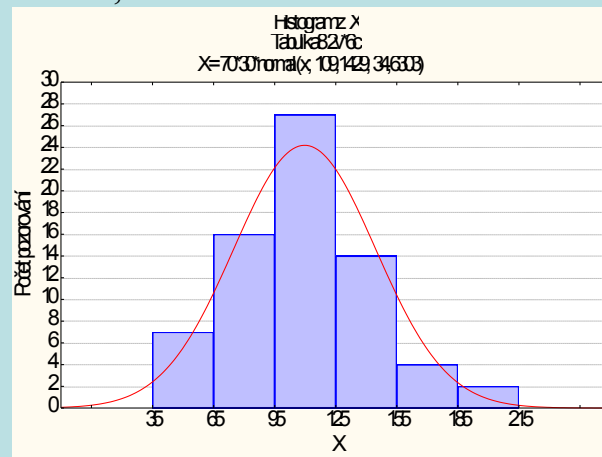


## Výpočet pomocí systému STATISTICA:

Otevřeme nový datový soubor o dvou proměnných a 6 případech. První proměnnou nazveme X, druhou četnost. Do proměnné X napíšeme středy třídících intervalů, do proměnné četnost odpovídající absolutní četnosti:

	1	2
	X	četnost
1	5	7
2	8	10
3	11	20
4	14	14
5	17	4
6	20	2

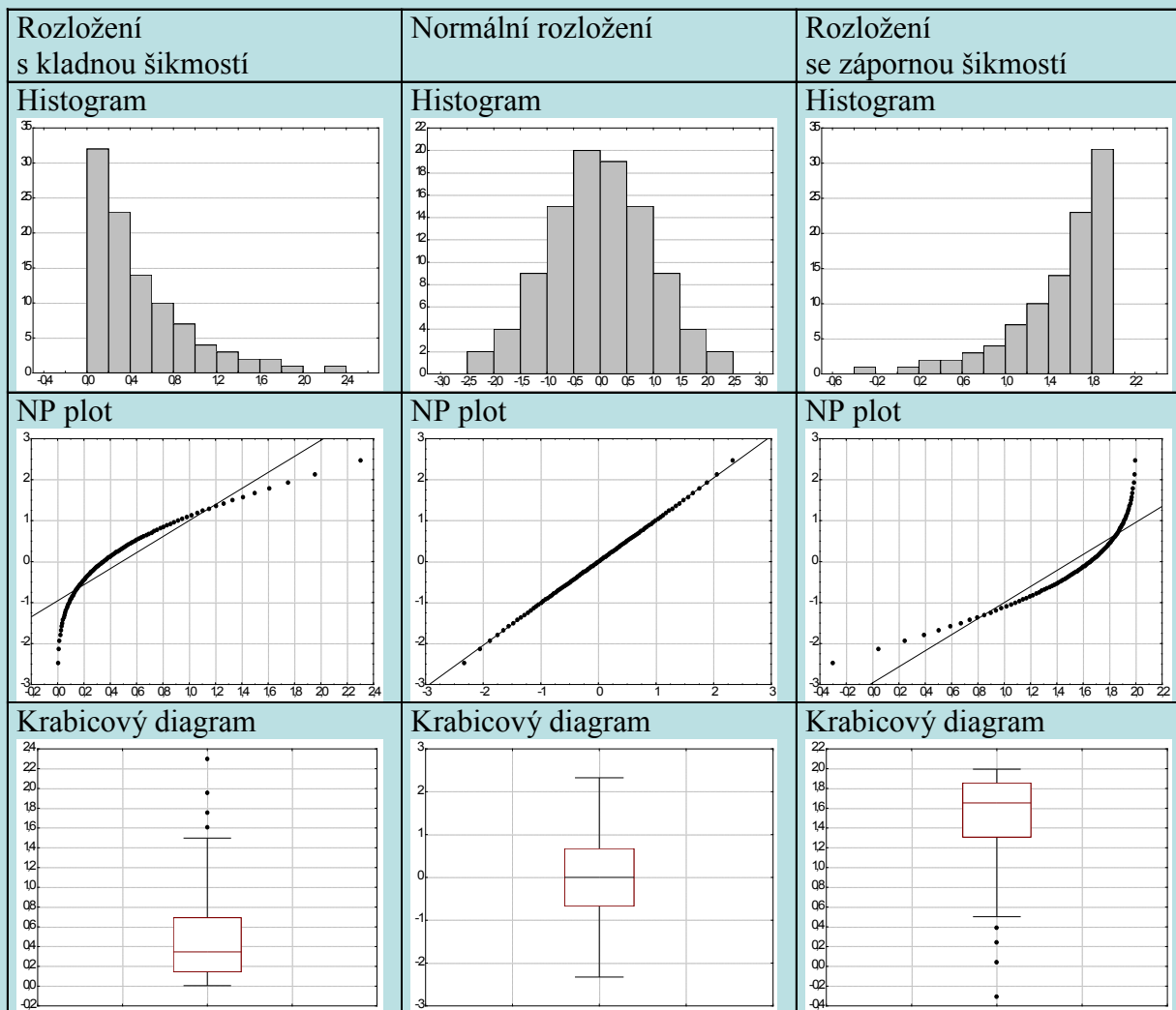
Grafy – Histogramy – zadáme proměnnou vah četnost – Proměnná X - zaškrtneme Hranice – Určit hranice – zaškrtneme Zadejte hraniční rozmezí: Minimum 35, Krok 30, Maximum 215 – OK – OK. Dostaneme graf:



Na rozdíl od histogramu konstruovaného ručně jsou na svislé ose absolutní četnosti, nikoliv četnostní hustoty. V porovnání s grafem hustoty normálního rozložení je vidět, že naše rozložení četností je lehce kladně zešikmené. Naše data tedy nepocházejí z normálního rozložení.

## Vzhled diagnostických grafů pro rozložení s různou šikmostí

Pro ilustraci se podívejme, jak se různá šikmost rozložení projeví na histogramu, N-P plotu a na krabicovém diagramu.



## Průzkumová analýza vícerozměrných dat

### Osnova:

- vícerozměrný datový soubor
- vizualizace vícerozměrných dat
- snížení dimenze dat metodou hlavních komponent
- shluková analýza

**Vícerozměrná data:** vyskytují se v situacích, kdy u každého z  $n$  objektů zjišťujeme hodnoty  $p$  znaků  $X_1, \dots, X_p$ .  
**p-rozměrný datový soubor:** matice  $n \times p$ :

$$\begin{pmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix}$$

Řádky charakterizují objekty, sloupce znaky.

Např. máme  $n$  sportovců, u každého sledujeme tyto znaky: pohlaví (0 – žena, 1 – muž), tělesná výška (v cm), tělesná hmotnost (v kg), nejlepší výkon ve skoku do dálky (v cm), nejlepší výkon ve skoku do výšky (v cm), nejlepší výkon v běhu na 100 m (v s).

Úkoly průzkumové analýzy vícerozměrných dat:

- odhalit vektory pozorování nebo jejich složky, které se jeví jako vybočující
- postihnout závislosti mezi sloupci datového souboru
- identifikovat shluky v datech, které svědčí o nehomogenitě daného výběru
- posoudit vícerozměrnou normalitu dat.

Omezíme se na dva problémy, a to na vizualizaci dat pomocí hlavních komponent a na shlukovou analýzu dat.

## Vizualizace vícerozměrných dat

Je-li  $p = 2$  nebo  $p = 3$ , můžeme hodnoty znaků chápat jako souřadnice v dvou či třírozměrném prostoru a získáme tak dvourozměrný či třírozměrný tečkový diagram. Ze vzhledu těchto tečkových diagramů lze poznat, zda se v datech vyskytují odlehlá pozorování, zda mezi znaky existuje nějaká závislost nebo zda se objekty sdružují do skupin.

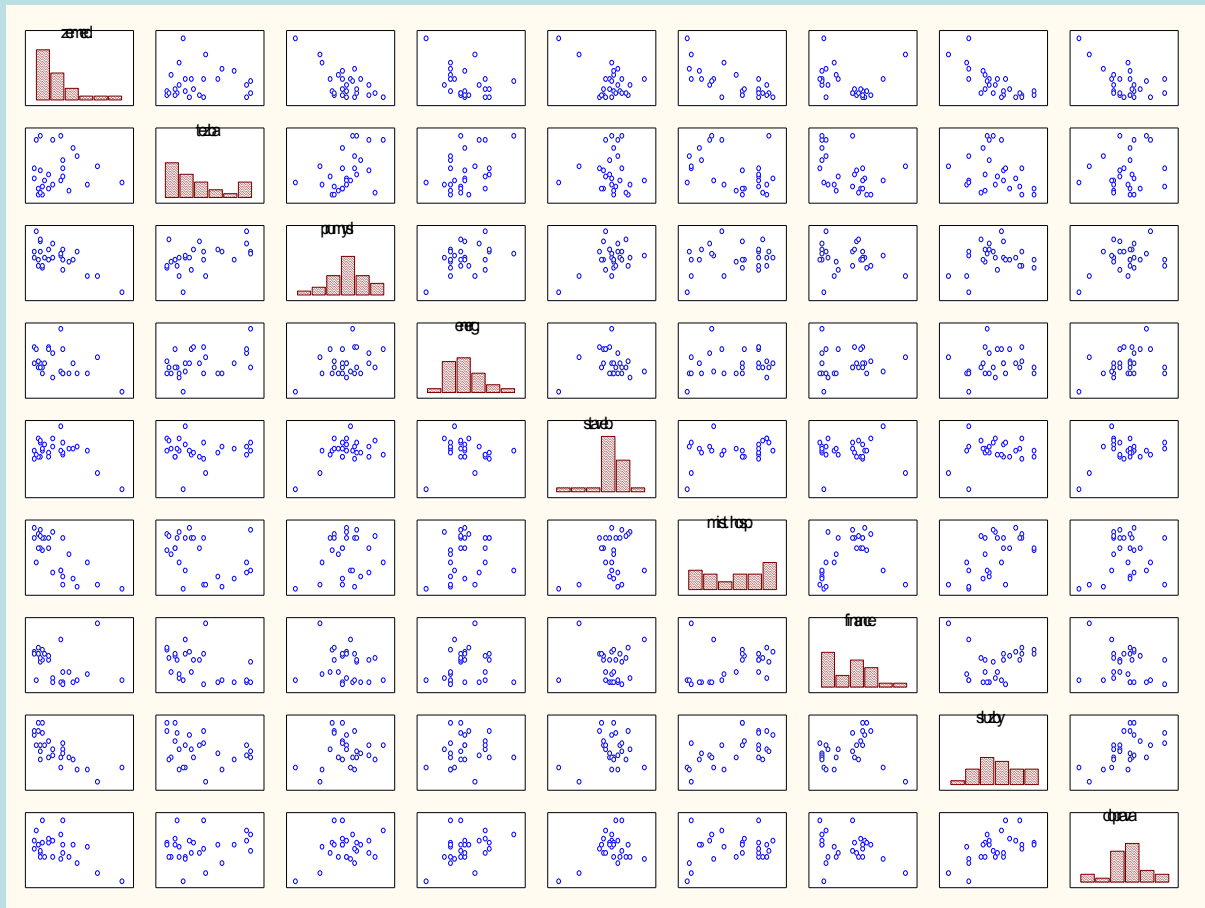
**Příklad:** Máme k dispozici datový soubor z roku 1979 o 26 evropských zemích, který obsahuje údaje o procentuálním zastoupení ekonomicky činného obyvatelstva v různých odvětvích národního hospodářství: zemědělství, těžba, průmyslová výroba, energetika, stavebnictví, místní hospodářství, finanční sektor, služby, doprava a komunikace.

	1	2	3	4	5	6	7	8	9
	zeme	tezba	prum	ener	stave	míst. hc	financ	sluzb	dopra
Belgie	3,	0,	27	0,	8,	19	6,	26	7,
Dánsko	9,	0,	21	0,	8,	14	6,	32	7,
Francie	10	0,	27	0,	8,	16	6,	22	5,
Západní Německo	6,	1,	35	0,	7,	14	5,	22	6,
Irsko	23	1,	20	1,	7,	16	2,	20	6,
Itálie	15	0,	27	0,	10	18	1,	20	5,
Lucembursko	7,	3,	30	0,	9,	18	4,	19	6,
Nizozemsko	6,	0,	22	1,	9,	11	6,	28	6,
Velká Británie	2,	1,	30	1,	6,	16	5,	28	6,
Rakousko	12	1,	31	1,	8,	16	4,	16	7,
Finsko	1,	0,	25	1,	7,	14	5,	24	7,
Recko	41	0,	17	0,	8,	11	2,	11	6,
Norsko	9	0,	22	0,	8,	16	4,	27	9,
Portugalsko	27	0,	24	0,	8,	13	2,	16	5,
Španělsko	22	0,	28	0,	11	9,	8,	11	5,
Svédsko	6,	0,	25	0,	7,	14	6,	32	6,
Svýcarsko	7,	0,	37	0,	9,	17	5,	15	5,
Turecko	66	0,	7,	0,	2,	5,	1,	11	3,
Bulharsko	23	1,	32	0,	7,	8	0,	18	6,
Československo	16	2,	35	1,	8,	9,	0,	17	7,
Východní Německo	4,	2,	41	1,	7,	11	1,	22	8,
Maďarsko	21	3,	29	1,	8,	9,	0,	17	8,
Polsko	31	2,	25	0,	8,	7,	0,	16	6,
Rumunsko	34	2,	30	0,	8,	5,	1,	11	6,
Sovětský svaz	23	1,	25	0,	9,	6,	0,	23	9,
Jugoslávie	48	1,	16	1,	4,	6,	11	5,	4

Vytvořte dvourozměrné tečkové diagramy pro všechny dvojice proměnných.

## Řešení pomocí systému STATISTICA:

Grafy – Maticové grafy – Proměnné – Vybrat vše – OK.



Na hlavní diagonále maticového grafu jsou histogramy jednotlivých proměnných, mimo hlavní diagonálu jsou dvourozměrné tečkové diagramy odpovídajících dvojic proměnných. Vidíme např., že podíl obyvatel zaměstnaných v zemědělství záporně koreluje s podílem obyvatel zaměstnaných v průmyslu, službách či dopravě.



Je-li  $p > 3$ , použijeme k vizualizaci dat **metodu hlavních komponent (principal component analysis)**, která umožňuje vyjádřit informace o variabilitě obsažené v datovém souboru pomocí několika málo nových znaků  $Y_1, \dots, Y_m$  získaných jako lineární kombinace znaků původních  $X_1, \dots, X_p$ ,  $m < p$  :

$$Y_1 = v_{11}X_1 + \dots + v_{1p}X_p,$$

$$Y_2 = v_{21}X_1 + \dots + v_{2p}X_p.$$

.

.

.

$$Y_m = v_{m1}X_1 + \dots + v_{mp}X_p.$$

Tyto nové znaky, kterým se říká hlavní komponenty, jsou

- nekorelované,
- uspořádané podle svého klesajícího rozptylu.

Většina informace o variabilitě původních dat je tedy soustředěna v první hlavní komponentě a nejméně informace je obsaženo v poslední hlavní komponentě. Ukazuje se, že pouze několik prvních hlavních komponent má dostatečně velký rozptyl. Ostatní pak můžeme zanedbat, čímž docílíme snížení dimenze dat. V datovém souboru však musí existovat mezi znaky dostatečně silná korelace, aby bylo možno tuto redukci provést.

Analýza hlavních komponent může být chápána jako transformace z původního do nového souřadnicového systému, jehož osy jsou tvořeny hlavními komponentami. Osy procházejí směry maximálního rozptylu, protože podmínka nezávislosti komponent vede ke kolmosti os.

Data pak znázorníme v prostoru prvních dvou či tří hlavních komponent.

Metodu hlavních komponent (Principal Component Analysis – PCA) popsal v r. 1901 Karl Pearson a ve 30. letech 20. století ji dále rozvinul Harold Hotelling.



Harold Hotelling (1895 – 1973), americký matematik a statistik

## Podstata metody hlavních komponent

Uvažme datový soubor, který vznikl tak, že 6 žáků absolvovalo 4 testy, které měří následující veličiny:

$X_1$  – přírodovědné znalosti,

$X_2$  – literární vědomosti,

$X_3$  – schopnost koncentrace,

$X_4$  – logické myšlení.

Testy se hodnotí na škále od 1 do 10 (1 = špatný výsledek, 10 = výborný výsledek)

	$X_1$	$X_2$	$X_3$	$X_4$
1	7	9	1	8
2	9	8	8	10
3	4	3	1	2
4	2	3	2	2
5	3	1	2	4
6	1	1	1	4

## Označení

$\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$  – vektor pozorování i-tého objektu,  $i = 1, 2, \dots, n$

Např. pro  $i = 3$  máme  $\mathbf{x}_3 = (4 \ 3 \ 1 \ 2)^T$

$m_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$  – průměr j-tého znaku,  $j = 1, 2, \dots, p$ . Např. pro  $j = 1$  máme  $m_1 = \frac{1}{7} (7 + 9 + 4 + 2 + 5 + 3 + 1) = 4,3$

$s_j^2 = \frac{1}{n} \sum_{i=1}^n (x_{ij} - m_j)^2$  – rozptyl j-tého znaku,  $j = 1, 2, \dots, p$ . Např. pro  $j = 1$  máme  $s_1^2 = \frac{1}{7} (7 - 4,3)^2 + \dots + (1 - 4,3)^2 = 9,4$

Datový soubor s průměry, směrodatnými odchylkami a rozptyly:

	1	2	3	4
	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	$\bar{x}_4$
1	7	9	1	8
2	9	8	8	10
3	4	5	1	2
4	2	5	2	2
5	5	1	2	4
6	1	1	1	4
průměr	4,3	4,1	4,0	5,0
s.o.	3,0	3,4	3,9	3,2
rozptyl	9,4	12,1	15,1	10,1

$Z_{ij} = \frac{x_{ij} - m_j}{s_j}$  – (i,j)-tá standardizovaná hodnota,  $i = 1, 2, \dots, n, j = 1, 2, \dots, p$

Např. pro  $i = 1, j = 1$  máme  $Z_{11} = \frac{7 - 4,3}{3,0} = 0,9$

Datový soubor standardizovaných hodnot

	1	2	3	4
	X1	X2	X3	X4
1	0,866	1,385	1,519	0,912
2	1,516	1,098	1,012	1,521
3	-0,10	-0,33	-0,75	-0,91
4	-0,75	-0,33	-0,50	-0,91
5	-0,43	-0,90	-0,50	-0,30
6	-1,08	-0,90	-0,75	-0,30

$\mathbf{z}_i = (z_{i1}, \dots, z_{ip})^T$  – vektor standardizovaných pozorování i-tého objektu,  $i = 1, 2, \dots, n$

$\mathbf{m} = (m_1, \dots, m_p)^T$  – vektor průměrů

Promě	Kovariance (pca)			
	X1	X2	X3	X4
X1	9,466	9,733	10,60	8,800
X2	9,733	12,16	13,20	9,400
X3	10,60	13,20	15,60	11,60
X4	8,800	9,400	11,60	10,80

$\mathbf{S} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T$  – výběrová varianční matice. V našem případě:

Promě	Korelace (pca)			
	X1	X2	X3	X4
X1	1,000	0,906	0,872	0,870
X2	0,906	1,000	0,958	0,820
X3	0,872	0,958	1,000	0,893
X4	0,870	0,820	0,893	1,000

$\mathbf{R} = \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \mathbf{z}_i^T$  – výběrová korelační matice. V našem případě:

(**S** a **R** jsou čtvercové symetrické matice řádu p.)

## Základní pojmy

**A** - čtvercová matice řádu  $p$ .

**Vlastní číslo matice A** – takové číslo  $\lambda$ , které pro libovolný nenulový vektor  $\mathbf{v}$  typu  $p \times 1$  splňuje rovnici  $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ .

**Vlastní vektor matice A** – vektor  $\mathbf{v}$ .

**Charakteristický polynom matice A** - determinant  $|\mathbf{A} - \lambda\mathbf{I}|$ .

**Stopa matice A** - součet jejích diagonálních prvků (značí se  $\text{Tr}(\mathbf{A})$ ).

## Výpočet vlastních čísel matice A

Rovnici  $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$  upravíme na tvar  $(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}$ . Tato soustava  $p$  rovnic má netriviální řešení, právě když charakteristický polynom matice **A** je roven 0. Dostaneme rovnici  $p$ -tého stupně. Jejím řešením jsou vlastní čísla  $\lambda_1, \dots, \lambda_p$ .

## Vlastnosti vlastních čísel

Jejich součet je roven stopě matice **A**:  $\lambda_1 + \dots + \lambda_p = \text{Tr}(\mathbf{A})$ ,

jejich součin je roven determinantu matice **A**:  $\lambda_1 \dots \lambda_p = \det(\mathbf{A})$ ,

jsou seřazena sestupně:  $\lambda_1 \geq \dots \geq \lambda_p$ .

## Vlastnosti vlastních vektorů

Mají jednotkovou délku:  $\mathbf{v}_i^T \mathbf{v}_i = 1$ ,  $i = 1, \dots, p$ ,

jsou vzájemně ortogonální:  $\mathbf{v}_i^T \mathbf{v}_j = 0$  pro všechna  $i \neq j$

## Získání hlavních komponent

Nechť výběrová varianční matice  $\mathbf{S}$  má vlastní čísla  $l_1, \dots, l_p$  a vlastní vektory  $\mathbf{v}_1, \dots, \mathbf{v}_p$ , přičemž  $\mathbf{v}_j^T \mathbf{v}_j = 1, j = 1, \dots, p$  a  $\mathbf{v}_j^T \mathbf{v}_k = 0$  pro  $j \neq k$ .

Znamená to, že vektory  $\mathbf{v}_1, \dots, \mathbf{v}_p$  jsou ortonormální.

Bez újmy na obecnosti předpokládáme, že  $l_1 > l_2 > \dots > l_p$ .

**1. hlavní komponenta** vznikne jako lineární kombinace znaků  $X_1, \dots, X_p$ , kde koeficienty této lineární kombinace jsou souřadnice vlastního vektoru  $\mathbf{v}_1$ , tedy

$$Y_1 = v_{11}X_1 + \dots + v_{1p}X_p.$$

Její rozptyl je  $l_1$ .

Dosadíme-li za  $X_1, \dots, X_p$  vektory pozorování  $\mathbf{x}_i, i = 1, \dots, n$ , dostaneme **vektor souřadnic**  $\mathbf{y}_1 = (y_{11}, \dots, y_{1n})^T$ , kde  $y_{1i} = \mathbf{v}_1^T \mathbf{x}_i$ .

**2. hlavní komponenta** vznikne jako lineární kombinace znaků  $X_1, \dots, X_p$ , kde koeficienty této lineární kombinace jsou souřadnice vlastního vektoru  $\mathbf{v}_2$ , tedy

$$Y_2 = v_{21}X_1 + \dots + v_{2p}X_p.$$

Její rozptyl je  $l_2$ .

Přitom  $\mathbf{v}_1^T \mathbf{v}_2 = 0$ , tj. 1. a 2. hlavní komponenta jsou lineárně nezávislé.

Dosadíme-li za  $X_1, \dots, X_p$  vektory pozorování  $\mathbf{x}_i, i = 1, \dots, n$ , dostaneme **vektor souřadnic**  $\mathbf{y}_2 = (y_{21}, \dots, y_{2n})^T$ , kde  $y_{2i} = \mathbf{v}_2^T \mathbf{x}_i$ .

.....  
**j-tá hlavní komponenta** vznikne jako lineární kombinace znaků  $X_1, \dots, X_p$ , kde koeficienty této lineární kombinace jsou souřadnice vlastního vektoru  $\mathbf{v}_j$ , tedy

$$Y_j = v_{j1}X_1 + \dots + v_{jp}X_p.$$

Její rozptyl je  $l_j$ . Přitom  $\mathbf{v}_j^T \mathbf{v}_k = 0, j = 1, \dots, k-1$ , tj. j-tá hlavní komponenta je lineárně nezávislá se všemi ostatními hlavními komponentami.

Dosadíme-li za  $X_1, \dots, X_p$  vektory pozorování  $\mathbf{x}_i, i = 1, \dots, n$ , dostaneme **vektor souřadnic**  $\mathbf{y}_j = (y_{j1}, \dots, y_{jn})^T$ , kde  $y_{ji} = \mathbf{v}_j^T \mathbf{x}_i$ .

Lze dokázat, že celková variabilita obsažená v datech je rovna stopě matice  $\mathbf{S}$ , tj. součtu vlastních čísel  $l_1 + \dots + l_p$ .

1. hlavní komponenta tedy vyčerpává  $\frac{l_1}{l_1 + \dots + l_p} \cdot 100\%$  celkové variability.

Pokud je číslo  $\frac{l_1}{l_1 + \dots + l_p}$  dostatečně blízké 1, znamená to, že 1. hlavní komponenta dobře nahrazuje celý datový soubor. Je-li toto číslo podstatně menší než 1, musíme vzít tolik hlavních komponent, aby jejich součet dělený stopou matice  $\mathbf{S}$  byl dostatečně blízký 1.

(V mnoha aplikacích se stává, že i při velkém počtu znaků stačí poměrně malý počet hlavních komponent.)

Znázorníme-li rozmístění objektů na ploše prvních dvou hlavních komponent, můžeme poznat, které objekty se řadí do skupin neboli shluků.

(Před provedením metody hlavních komponent je třeba se rozhodnout, zda budeme pracovat s původními hodnotami znaků nebo standardizovanými hodnotami.)

**Důležité upozornění:** Proměnné  $X_1, \dots, X_p$  musí být mezi sebou dostatečně korelované, jinak metoda hlavních komponent nedá dobré výsledky.

**Koeficient korelace**  $i$ -tého znaku  $X_i$  s  $k$ -tou hlavní komponentou  $Y_k$  lze vyjádřit jako  $R_{X_i, Y_k} = \frac{v_{ik} \sqrt{l_k}}{s_i}$ .

**Reprodukce výchozí kovarianční matice:** platí vzorec  $\mathbf{S} = \sum_{i=1}^p l_i \mathbf{v}_i \mathbf{v}_i^T$  (tzv. spektrální rozklad matice  $\mathbf{S}$ ).

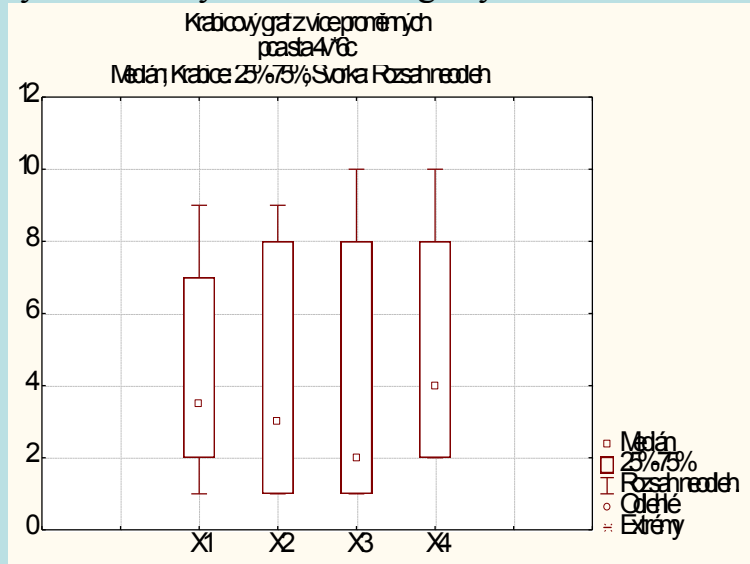
Rozhodneme-li se uvažovat právě  $m$  hlavních komponent ( $m \leq p$ ), pak pomocí tohoto vztahu můžeme posoudit, jak těchto  $m$  hlavních komponent reprodukuje rozptyly a kovariance původních proměnných. Lze posoudit i reziduální matici, tj. matici, kterou získáme jako rozdíl výchozí kovarianční matice a reprodukované kovarianční matice.

## Doporučený postup při analýze hlavních komponent

- a) Provedeme tabulkové a grafické zpracování datového souboru, abychom se blíže seznámili s daty.
- b) Sestavíme korelační matici a prověříme, zda jsou korelace natolik silné, aby mělo smysl provádět analýzu hlavních komponent.
- c) Rozhodneme, kolika hlavními komponentami lze popsat datový soubor bez podstatné ztráty informace. Označme tento vhodný počet jako  $m$ . Při stanovení  $m$  můžeme použít tato pomocná kritéria:
  - **Kaiserovo kritérium** - za  $m$  volíme počet těch vlastních čísel matice  $\mathbf{R}$ , která jsou větší než 1.
  - **Sutinový test** (scree test) – grafická metoda, která spočívá v subjektivním posouzení vzhledu sutinového grafu (scree plot), tj. grafu znázorňujícího velikosti sestupně uspořádaných vlastních čísel matice  $\mathbf{R}$ . Objeví-li se v grafu určité zploštění, pak za  $m$  vezmeme to pořadové číslo, kde se zploštění projevilo.
  - **Kritérium založené na kumulativním procentu vysvětleného rozptylu**. Požadujeme, aby vybrané hlavní komponenty vysvětlily aspoň 70% celkového rozptylu.
  - **Kritérium založené na reziduální korelační či kovarianční matici**. Požadujeme, aby prvky reziduální matice byly co možná nejmenší.
- d) Pokusíme se o interpretaci prvních  $m$  hlavních komponent. Zkoumáme přitom, jak jsou jednotlivé vybrané hlavní komponenty utvořeny z původních znaků a jak s nimi korelují.
- e) Vypočítáme vektory souřadnic a následně sestrojíme dvourozměrné tečkové diagramy.



Pro náš datový soubor obsahující výsledky 6 žáků ve 4 testech nejprve znázorníme data pomocí krabicových diagramů: Grafy – 2D Grafy – Krabicové grafy – zvolíme Vícenásobný – Proměnné - Závisle proměnné X1-X4 – OK – OK



Nyní vypočteme korelační matici: Statistiky – Vícerozměrné průzkumné techniky – Hlavní komponenty & klasifikační analýza – Proměnné X1 až X4, OK – OK – Popisné statistiky – Korelační matice

Promě	Korelace (pca.sta)			
	X1	X2	X3	X4
X1	1,000	0,906	0,872	0,870
X2	0,906	1,000	0,958	0,820
X3	0,872	0,958	1,000	0,893
X4	0,870	0,820	0,893	1,000

Dále vypočteme vlastní čísla a procento vysvětleného rozptylu: na záložce Základní výsledky vybereme Vlastní čísla.

Vlastní čísla korelační matice a sd				
Pouze aktiv. proměnné				
Pořadí	Vl. čís	% ce rozpty	Kumula vl. čís	Kumula %
1	3,661	91,53	3,661	91,53
2	0,188	4,71	3,850	96,24
3	0,134	3,35	3,984	99,60
4	0,015	0,39	4,000	100,00

Vidíme, že 1. vlastní číslo  $l_1 = 3,66$ , tedy 1. hlavní komponenta vyčerpává 91,5% variability dat,

2. vlastní číslo  $l_2 = 0,19$ , 2. hlavní komponenta vyčerpává 4,7% variability dat atd.

Podle Kaiserova kritéria by stačilo uvažovat pouze 1. hlavní komponentu, protože pouze první vlastní číslo je větší než 1.

Kvůli znázornění objektů však budeme uvažovat první dvě hlavní komponenty.

Dále vypočítáme vlastní vektory: na záložce Proměnné vybereme Vlastní vektory

Vlastní vektory korelační m				
Pouze aktiv. proměnné				
Promě	Fakto	Fakto	Fakto	Fakto
X1	-0,498	-0,000	0,817	-0,289
X2	-0,503	0,582	-0,082	0,632
X3	-0,508	0,185	-0,539	-0,645
X4	-0,488	-0,791	-0,187	0,314

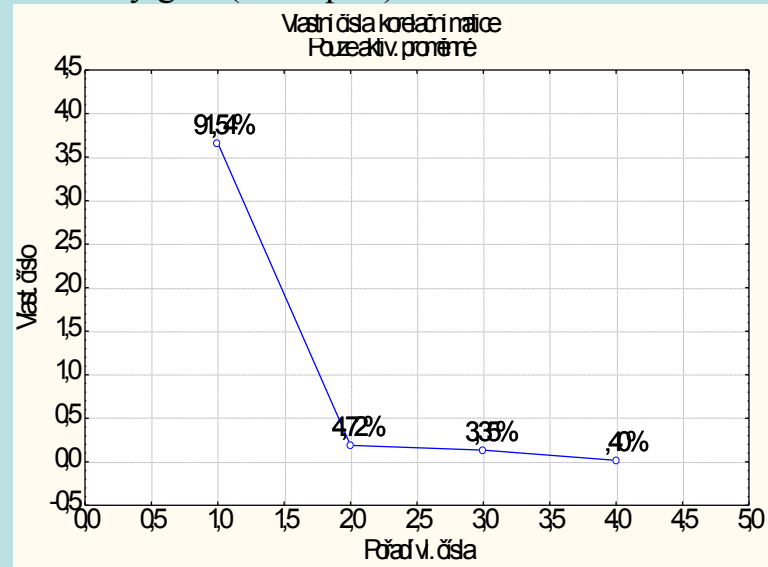
1. hlavní komponenta:

$$Y_1 = -0,49X_1 - 0,5X_2 - 0,51X_3 - 0,49X_4,$$

2. hlavní komponenta:

$$Y_2 = -0,0005X_1 + 0,58X_2 + 0,19X_3 - 0,79X_4 \text{ atd.}$$

Sutinový graf (scree plot):



V sutinovém grafu nastává výrazné zploštění po 1. vlastním čísle.

Výpočet koeficientů korelace 1. a 2. hlavní komponenty a původních čtyř proměnných: na záložce Proměnné vybereme Korelace faktorů & proměnných

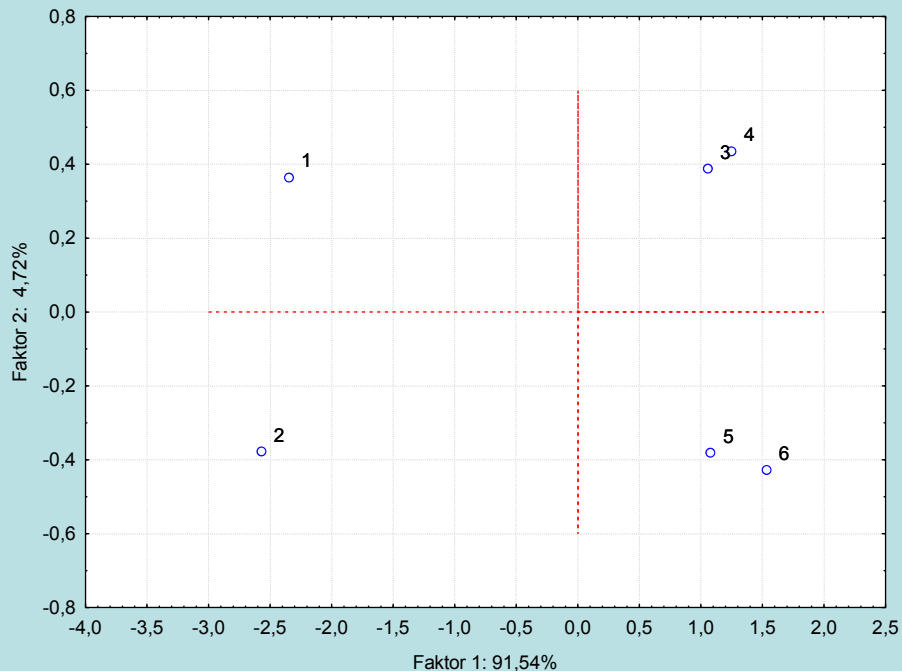
Promě	Faktor 1	Faktor 2
X1	-0,953	-0,000
X2	-0,963	0,252
X3	-0,973	0,080
X4	-0,935	-0,343

Vidíme, že 1. hlavní komponenta vysoce záporně koreluje se všemi proměnnými. 2. hlavní komponenta slabě kladně koreluje s druhou proměnnou a středně silně záporně koreluje s třetí proměnnou.

Podívejme se rovněž na vektory souřadnic (v systému STATISTICA se jim říká faktorové souřadnice případů): na záložce Případy vybereme Faktorové souřadnice případů.

Příp.	Faktor 1	Faktor 2
1	-2,341	0,364
2	-2,561	-0,378
3	1,051	0,387
4	1,250	0,434
5	1,079	-0,381
6	1,532	-0,427

Znázornění objektů (žáků) na ploše prvních dvou hlavních komponent:



## Shluková analýza

### Cíl shlukové analýzy

Cílem shlukové analýzy je roztržidění  $n$  objektů, z nichž každý je popsán  $p$  znaky, do několika pokud možno stejnorodých (homogenních) skupin (shluků, clusterů). Požadujeme, aby objekty uvnitř shluků si byly podobné co nejvíce, zatímco objekty z různých shluků co nejméně. Přesný počet shluků většinou není přesně znám.

Shluková analýza nachází uplatnění v celé řadě oborů, např. v biologii. U  $n$  populací změříme  $p$  biometrických charakteristik a zjišťujeme, zda určité skupiny populací tvoří shluky.

Shluková analýza je ovšem průzkumovou metodou a měla by sloužit jako určité vodítko při dalším zpracování dat.

### Podobnost objektů

Podobnost (či rozdílnost) objektů posuzujeme pomocí různých měr vzdálenosti. Pro znaky intervalového či poměrového typu nejčastěji používáme **euklidovskou vzdálenost**.

Nechť  $k$ -tý objekt je popsán vektorem pozorování  $\mathbf{x}_k = (x_{k1}, \dots, x_{kp})^T$  a  $l$ -tý objekt vektorem  $\mathbf{x}_l = (x_{l1}, \dots, x_{lp})^T$ .

Euklidovská vzdálenost  $k$ -tého a  $l$ -tého objektu:

$$d_{kl} = \sqrt{\sum_{j=1}^p (x_{kj} - x_{lj})^2}.$$

Vzdálenosti vypočtené pro všechny dvojice objektů se uspořádají do **matice vzdáleností**. Je zřejmé, že je to čtvercová symetrická matice, která má na hlavní diagonále nuly.

### Matice euklidovských vzdáleností pro datový soubor s údaji o 6 žácích:

Statistiky – Vícerozměrné průzkumné techniky – Shluková analýza – Spojování (hierarchické shlukování) – OK – Proměnné  $X_1 - X_4$  – OK – na záložce Details vybereme Shlukovat Případy (řádky) – OK – na záložce Details vybereme Matice vzdáleností.

Příp.	Euklid. vzdálenosti					
P 1	0	3	12	12	12	14
P 2	3	0	12	13	12	14
P 3	12	12	0	2	3	4
P 4	12	13	2	0	3	3
P 5	12	12	3	3	0	2
P 6	14	14	4	3	2	0

## Hierarchické shlukování

Při aplikacích shlukové analýzy se nejčastěji používá **aglomerativní hierarchická procedura**. Její princip spočívá v postupném slučování objektů, a to nejprve nejbližších a v dalších krocích pak stále vzdálenějších.

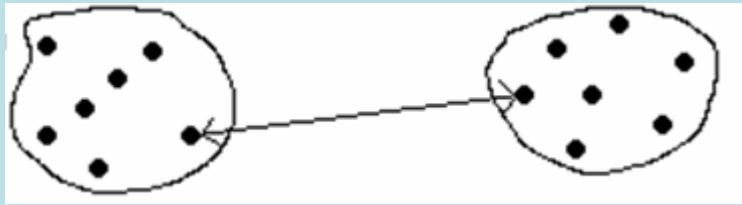
### Algoritmus:

1. krok: Každý objekt považujeme za samostatný shluk.
2. krok: Najdeme dva shluky, jejichž vzdálenost je minimální.
3. krok: Tyto dva shluky spojíme v nový, větší shluk a přepočítáme matici vzdáleností. Její řád se sníží o 1. Vrátime se na 2. krok.

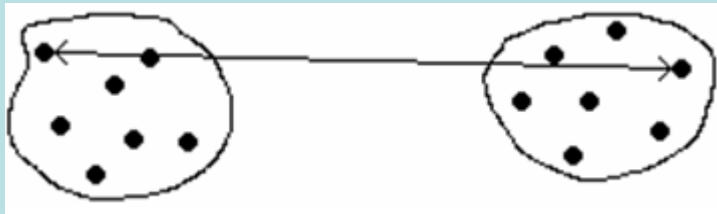
Funkce algoritmu končí, až jsou všechny objekty spojeny do jediného shluku.

Vzdálenost mezi shluky se počítá různými způsoby. Uvedeme tři z nich.

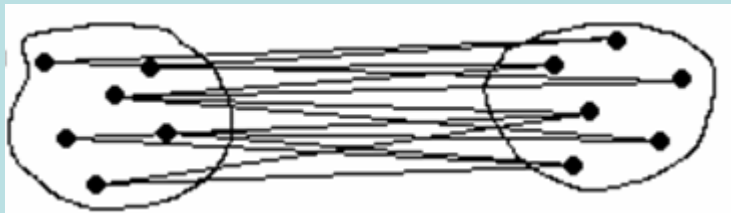
a) **Metoda nejbližšího souseda:** Vzdálenost mezi dvěma shluky je minimem ze všech vzdáleností mezi jejich objekty.



b) **Metoda nejvzdálenějšího souseda:** Vzdálenost mezi dvěma shluky je maximem ze všech vzdáleností mezi jejich objekty.

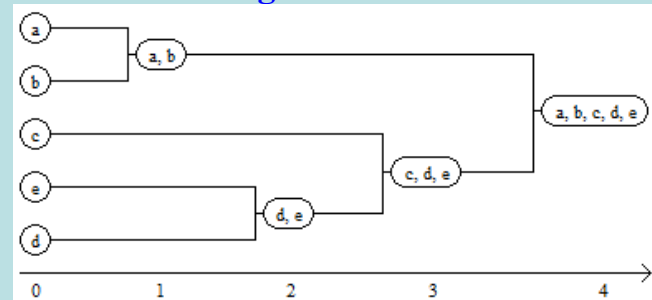


c) **Metoda průměrné vazby:** Vzdálenost mezi dvěma shluky je průměrem ze všech vzdáleností mezi jejich objekty.



Výsledky aglomerativní hierarchické procedury se zpravidla znázorňují pomocí **dendrogramu**. Je to graficky znázorněná posloupnost dvojic  $\{v_1, S^{(1)}\}, \dots, \{v_n, S^{(n)}\}$ , kde  $v_i$  je neklesající posloupnost úrovní spojování a  $S^{(i)}$  je rozřídění objektů odpovídající úrovni  $v_i$ ,  $i = 1, \dots, n$ .

### Příklad dendrogramu:



V levém sloupci jsou jednotlivé objekty, další sloupce reprezentují shluky, do nichž byly objekty zařazeny a délky čar představují vzdálenosti mezi shluky.

### Poznámka:

Hierarchická shluková analýza může být použita nejen na shlukování objektů, ale též na shlukování znaků.

**Dendrogram podobnosti objektů** je standardní výstup hierarchických shlukovacích metod, z něhož je zjevná struktura objektů ve shlucích.

**Dendrogram podobnosti znaků** odhaluje nejčastěji dvojice či trojice (všeobecně m-tice) znaků, které si jsou velmi podobné a silně spolu korelují. Znaky, které jsou ve společném shluku, si jsou značně podobné a jsou tudíž vzájemně nahraditelné. To má značný význam při plánování experimentu - některé vlastnosti či znaky není zapotřebí vůbec zjišťovat či měřit, protože jsou snadno nahraditelné jinými znaky a nemají velkou vypovídací hodnotu.



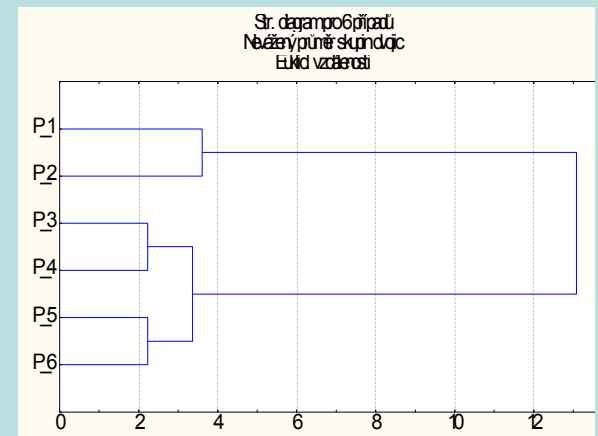
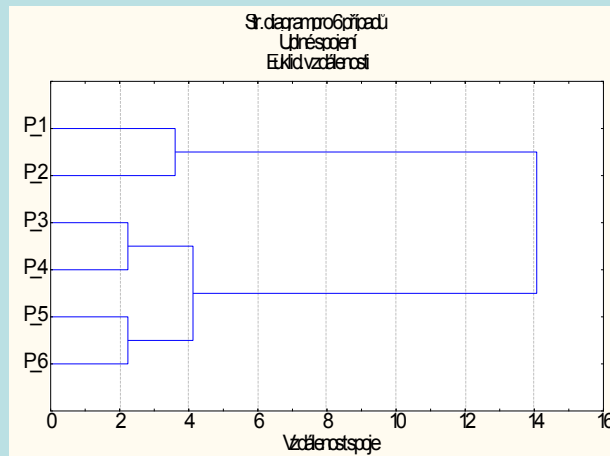
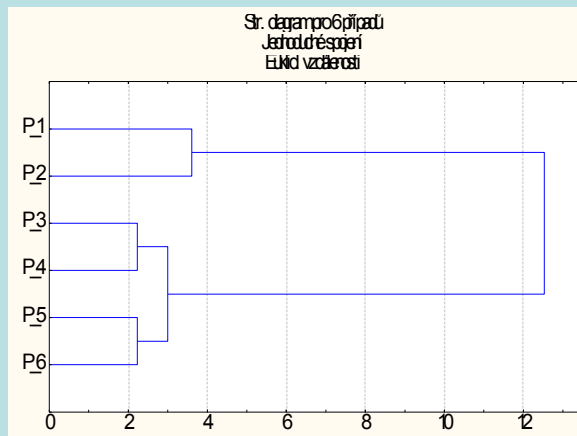
## Vytvoření dendrogramu v systému STATSTICA:

- pro metodu nejbližšího souseda:

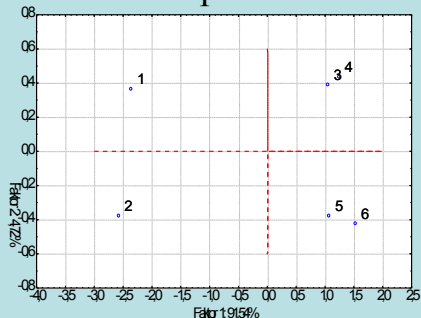
Statistiky – Vícerozměrné průzkumné techniky – Shluková analýza – Spojování (hierarchické shlukování) – OK – Proměnné X1 – X4 – OK – na záložce Details vybereme Shlukovat Případy (řádky), pravidlo slučování ponecháme Jednoduché spojení, míru vzdálenosti ponecháme Euklidovské vzd. – OK – Horizontální graf hierarch. stromu

- pro metodu nejvzdálenějšího souseda: na záložce Details vybereme pravidlo slučování Úplné spojení,

- pro metodu úplné vazby: Na záložce Details vybereme pravidlo slučování Nevážený průměr skupin dvojic.



Vidíme, že výsledky všech tří metod jsou velmi podobné a odpovídají rozmístění objektů (žáků) na ploše prvních dvou hlavních komponent.



**Příklad:** Uvažme datový soubor s údaji o 26 evropských státech. Tento datový soubor budeme analyzovat metodou hlavních komponent a následně provedeme shlukovou analýzu.

### Provedení PCA

Nejprve pomocí korelační matice posoudíme, zda má smysl aplikovat PCA.

Statistiky – Vícerozměrné průzkumné techniky – Hlavní komponenty & klasifikační analýza – Proměnné X1 až X19, OK – OK – Popisné statistiky – Korelační matice.

Promě	Korelace (staty1979.sta)								
	X1	X2	X3	X4	X5	X6	X7	X8	X9
X1	1,0	0,0	-0,6	-0,4	-0,4	-0,1	-0,4	-0,1	-0,4
X2	0,0	1,0	0,4	0,4	-0,0	-0,4	-0,4	-0,4	0,1
X3	-0,6	0,4	1,0	0,3	0,4	0,2	-0,1	0,1	0,3
X4	-0,4	0,4	0,3	1,0	0,0	0,2	0,1	0,1	0,3
X5	-0,4	-0,0	0,4	0,0	1,0	0,3	0,0	0,1	0,3
X6	-0,1	-0,4	0,2	0,2	0,3	1,0	0,3	0,5	0,1
X7	-0,4	-0,4	-0,1	0,1	0,0	0,3	1,0	0,1	-0,4
X8	-0,1	-0,4	0,1	0,1	0,1	0,5	0,1	1,0	0,5
X9	-0,4	0,1	0,3	0,3	0,3	0,1	-0,4	0,5	1,0

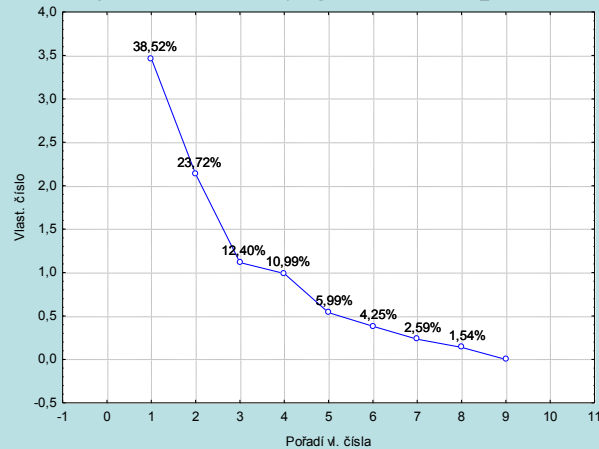
Některé korelační koeficienty jsou v absolutní hodnotě dostatečně velké a zřejmě tedy bude mít smysl provést analýzu hlavních komponent.

Nyní získáme vlastní čísla výběrové korelační matice a procento vysvětleného rozptylu: na záložce Základní výsledky vybereme Vlastní čísla.

Pořadí	Vl. čís	% cel rozpty	Kumula vl. čísl	Kumula %
1	3,466	38,51	3,466	38,51
2	2,135	23,72	5,601	62,23
3	1,115	12,39	6,717	74,63
4	0,989	10,99	7,706	85,62
5	0,539	5,99	8,245	91,61
6	0,382	4,24	8,627	95,86
7	0,233	2,59	8,861	98,45
8	0,138	1,54	9,000	100,00

První hlavní komponenta tedy vysvětluje 38,52% variability obsažené v devíti sledovaných proměnných, druhá 23,72%, třetí 12,40% atd. Celkové procento variability vysvětlené prvními třemi hlavními komponentami je 74,63%.

Sestrojíme sutinový graf (scree plot): na záložce Základní výsledky vybereme Sutinový graf.

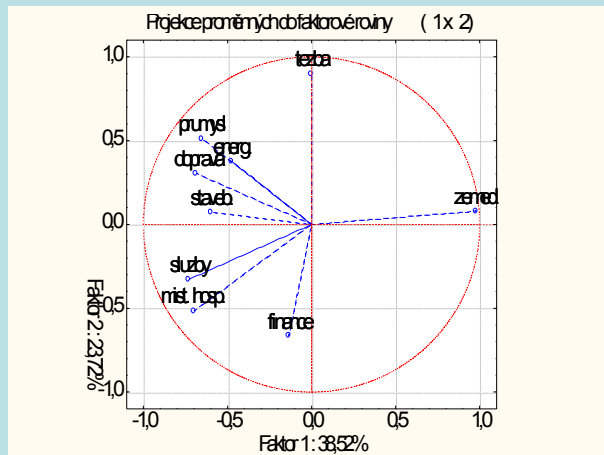


Počet m hlavních komponent zvolíme tři. V nabídce Výsledky hlavních komponent snížíme počet faktorů na 3.

Vypočteme korelační koeficienty prvních tří hlavních komponent a původních devíti proměnných: na záložce Proměnné vybereme Korelace faktorů & proměnných.

Proměň	Faktor1	Faktor2	Faktor3
X1	0,978	0,081	-0,049
X2	-0,000	0,901	0,216
X3	-0,652	0,513	0,112
X4	-0,474	0,378	0,649
X5	-0,595	0,073	-0,304
X6	-0,698	-0,513	0,119
X7	-0,136	-0,663	0,589
X8	-0,727	-0,327	-0,251
X9	-0,684	0,304	-0,337

Graficky lze znázornit souvislost mezi novými proměnnými (např. 1. a 2. HK) a původními proměnnými X1, ..., X9 takto: na záložce Proměnné vybereme 2D graf fakt. souřadnic prom. - Osa x: Faktor I, Osa y: Faktor 2 - OK. Na ose x budou souřadnice vstupních proměnných vzhledem k první hlavní komponentě, na ose Y vzhledem ke druhé komponentě.



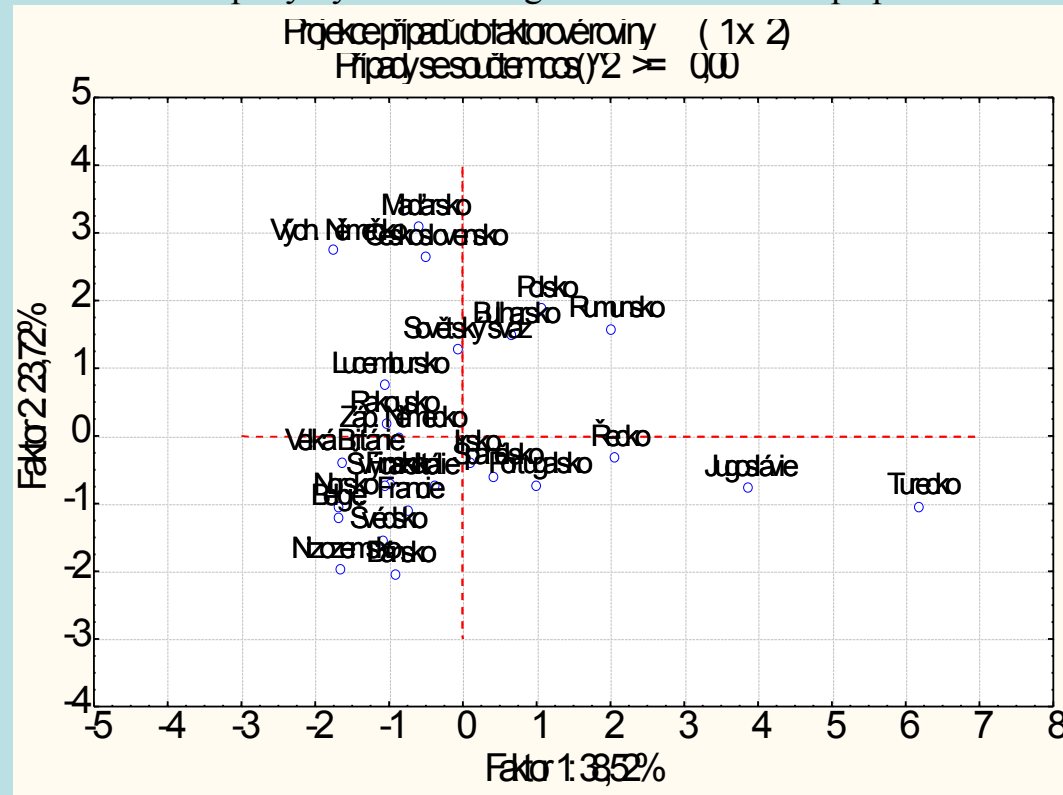
1. HK vysoce kladně koreluje s proměnnou X1, tj se zemědělstvím a negativně s proměnnou X8 – služby. Jelikož je podíl lidí v zemědělství a ve službách obecně považován za určité měřítko vyspělosti země, můžeme první komponentu interpretovat jako míru zaostalosti/vyspělosti.
2. HK výrazně pozitivně koreluje s těžebním průmyslem, energetikou a zpracovatelským průmyslem. Negativně koreluje se službami a finanční sférou. Budeme ji proto interpretovat jako míru toho, nakolik se země orientuje na průmyslovou výrobu. (Ne vždy mají komponenty takto jasnou interpretaci. Jsou jen jistou matematickou transformací vstupních proměnných, která může a nemusí odrážet nějakou reálnou vlastnost objektů!).

Podívejme se rovněž na vektory souřadnic (v systému STATISTICA se jim říká faktorové souřadnice případů): na záložce Případy vybereme Faktorové souřadnice případů.

Případ	Fakto	Fakto	Fakto
Belgie	-1,68	-1,20	0,16
Dánsko	-0,90	-2,05	-0,85
Francie	-0,74	-1,11	0,38
Zapadnecko	-0,85	-0,03	0,56
Irsko	0,11	-0,40	0,53
Italie	-0,36	-0,74	-1,29
Lucemburs	-1,04	0,74	0,46
Nizozemsko	-1,65	-1,98	-0,08
Velká Britá	-1,61	-0,39	1,35
Rakousko	-1,01	0,16	1,16
Finsko	-0,97	-0,73	0,54
Recko	2,07	-0,33	-0,92
Norsko	-1,66	-1,05	-1,14
Portugalsko	0,99	-0,74	-0,75
Španělsko	0,43	-0,60	0,31
Svédsko	-1,07	-1,55	-0,22
Svýcarsko	-1,04	-0,74	0,28
Turecko	6,19	-1,04	-0,64
Bulharsko	0,67	1,48	-1,03
Českosloven	-0,48	2,63	0,07
Východneck	-1,73	2,73	0,26
Maďarsko	-0,57	3,07	1,09
Polsko	1,08	1,87	-0,54
Rumunsko	2,01	1,57	-0,48
Sovětský s	-0,04	1,26	-2,30
Jugoslávie	3,87	-0,78	3,07

1. HK vysoce kladně koreluje s proměnnou  $X_1$  (zemědělství) a záporně se všemi ostatními proměnnými. Tato hlavní komponenta tedy rozlišuje země na zemědělské a průmyslové. Povšimněte si, že souřadnice této hlavní komponenty jsou nejvyšší u Turecka (6,2) a Jugoslávie (3,9).
2. HK vysoce kladně koreluje s proměnnou  $X_2$  (těžba) a podstatně slaběji s proměnnou  $X_3$  (průmyslová výroba). Vysoké hodnoty souřadnic této hlavní komponenty najdeme u Maďarska, Východního Německa a Československa.
3. HK středně silně koreluje s proměnnou  $X_4$  (energetika) a  $X_7$  (finanční sektor). Nejvyšší hodnotu najdeme u Jugoslávie.

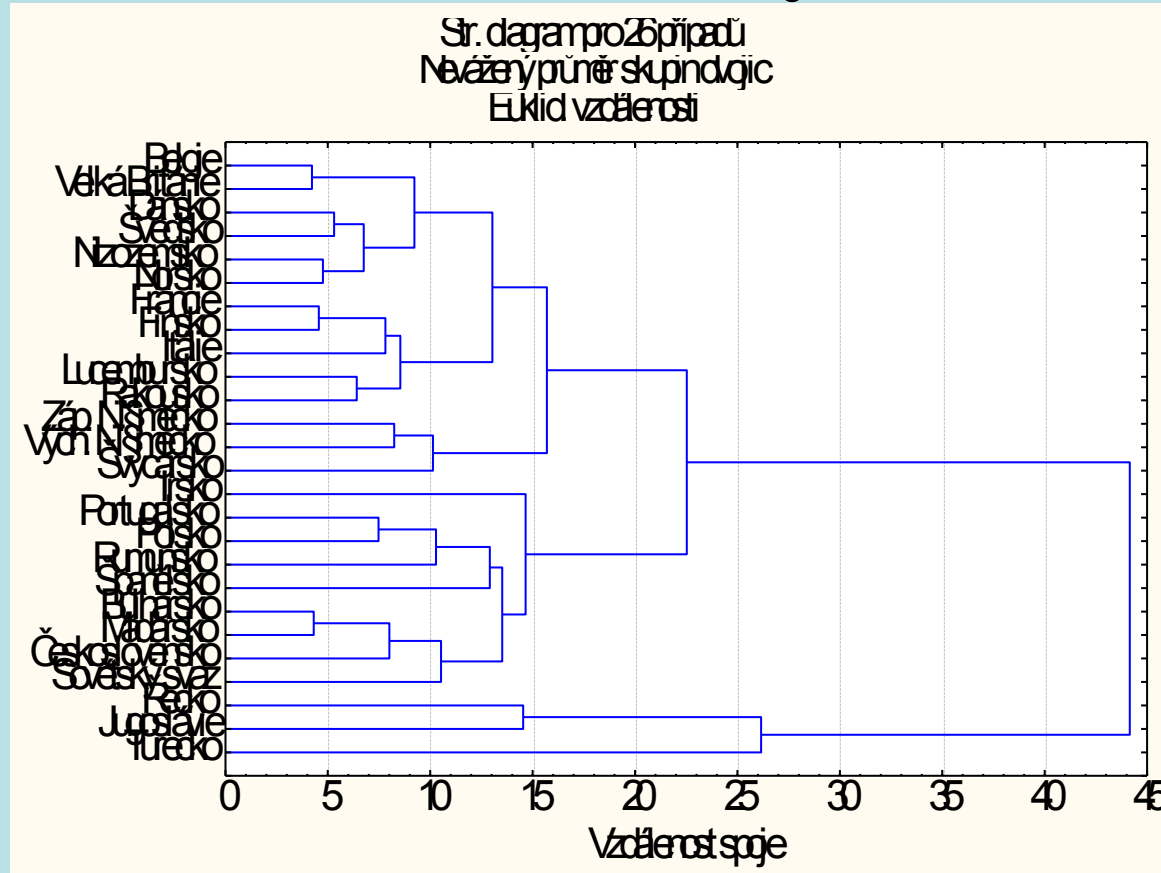
Nyní znázorníme rozmístění zemí na ploše prvních dvou hlavních komponent:  
 Na záložce Případy vybereme 2D graf fakt. Souřadnic příp.



Státy napravo jsou státy s vysokým podílem zemědělství. Vyniká zde zejména Turecko a Jugoslávie. Všechny státy obvykle považované za ekonomicky vyspělé jsou naopak na levé straně. Jsou to státy, kde je nižší podíl osob zaměstnaných v zemědělství, zato vyšší podíl osob pracujících ve službách. Je zde také hezky vidět zaměření zemí tehdejšího socialistického bloku na průmyslovou výrobu - horní část grafu. A naopak severské státy a státy Beneluxu orientované na finanční a další služby v dolní části.

## Provedení shlukové analýzy

Statistiky – Vícerozměrné průzkumné techniky – Shluková analýza - Spojování (hierarchické shlukování) – OK - Proměnné X1 až X4, OK, Detaily - Shlukovat případy (řádky) – Pravidlo slučování: Nevážený průměr skupin dvojic – Míry vzdálenosti: Euklidovské vzdálenosti - OK – Horizontální graf hierarch. stromu.



Ukazuje se, že země se dělí do tří skupin: první skupinu tvoří rozvinuté demokratické země společně s NDR, druhou skupinu socialistické země s Irskem, Portugalskem a Španělskem a třetí Řecko s Jugoslávií. Turecko se chová jako singulární entita.

## Základní pojmy matematické statistiky I

### Motivace:

Matematická statistika je věda, která analyzuje a interpretuje data především za účelem získání předpovědi a zlepšení rozhodování v různých oborech lidské činnosti. Přitom se řídí principem statistické indukce, tj. na základě znalostí o náhodném výběru z určitého rozložení pravděpodobností se snaží učinit závěry o vlastnostech tohoto rozložení.

Ústředním pojmem matematické statistiky je tedy pojem náhodného výběru.

### Osnova:

- náhodný výběr z jednorozměrného a vícerozměrného rozložení
- statistika jako funkce náhodného výběru
- bodové a intervalové odhady parametrů a parametrických funkcí



### Definice náhodného výběru:

- a) Necht'  $X_1, \dots, X_n$  jsou stochasticky nezávislé náhodné veličiny, které mají všechny stejné rozložení  $L(\underline{Q})$ . Řekneme, že  $X_1, \dots, X_n$  je **náhodný výběr rozsahu  $n$  z rozložení  $L(\underline{Q})$** . (Číselné realizace  $x_1, \dots, x_n$  náhodného výběru  $X_1, \dots, X_n$  uspořádané do sloupcového vektoru odpovídají datovému souboru zavedenému v popisné statistice.)
- b) Necht'  $(X_1, Y_1), \dots, (X_n, Y_n)$  jsou stochasticky nezávislé dvourozměrné náhodné vektory, které mají všechny stejné dvourozměrné rozložení  $L_2(\underline{Q})$ . Řekneme, že  $(X_1, Y_1), \dots, (X_n, Y_n)$  je **dvourozměrný náhodný výběr rozsahu  $n$  z dvourozměrného rozložení  $L_2(\underline{Q})$** . (Číselné realizace  $(x_1, y_1), \dots, (x_n, y_n)$  náhodného výběru  $(X_1, Y_1), \dots, (X_n, Y_n)$  uspořádané do matice typu  $2 \times n$  odpovídají dvourozměrnému datovému souboru zavedenému v popisné statistice.)
- c) Analogicky lze definovat  $p$ -rozměrný **náhodný výběr rozsahu  $n$  z  $p$ -rozměrného rozložení  $L_p(\underline{Q})$** .

### Definice statistiky:

Libovolná funkce  $T = T(X_1, \dots, X_n)$  náhodného výběru  $X_1, \dots, X_n$  (resp.  $T = T(X_1, Y_1, \dots, X_n, Y_n)$  náhodného výběru  $(X_1, Y_1), \dots, (X_n, Y_n)$ ) se nazývá (výběrová) **statistika**.

### Definice důležitých statistik:

a) Necht'  $X_1, \dots, X_n$  je náhodný výběr,  $n \geq 2$ .

Onačme  $M = \frac{1}{n} \sum_{i=1}^n X_i$  ... výběrový průměr,  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - M)^2$  ... výběrový rozptyl,  $S = \sqrt{S^2}$  ... výběrová směrodatná odchylka

Pro libovolné, ale pevně dané reálné číslo  $x$  je statistikou též hodnota **výběrové distribuční funkce**  $F_n(x) = \frac{1}{n} \text{card}\{X_i \leq x\}$

b) Necht' je dáno  $r \geq 2$  stochasticky nezávislých náhodných výběrů o rozsazích  $n_1 \geq 2, \dots, n_r \geq 2$ .

Celkový rozsah je  $n = \sum_{j=1}^r n_j$ .

Označme  $M_1, \dots, M_r$  výběrové průměry a  $S_1^2, \dots, S_r^2$  výběrové rozptyly jednotlivých výběrů. Necht'  $c_1, \dots, c_r$  jsou reálné konstanty, aspoň jedna nenulová.

$\sum_{j=1}^r c_j M_j$  ... lineární kombinace výběrových průměrů,  $S^2 = \frac{\sum_{j=1}^r c_j^2 S_j^2}{\sum_{j=1}^r c_j^2}$  ... vážený průměr výběrových rozptylů.

c) Necht'  $(X_1, Y_1), \dots, (X_n, Y_n)$  je náhodný výběr z dvourozměrného rozložení o rozsahu  $n$ .

Označme  $M_1 = \frac{1}{n} \sum_{i=1}^n X_i$ ,  $M_2 = \frac{1}{n} \sum_{i=1}^n Y_i$  výběrové průměry,  $S_1^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - M_1)^2$ ,  $S_2^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - M_2)^2$  výběrové rozptyly.

$S_{12} = \frac{1}{n-1} \sum_{i=1}^n (X_i - M_1)(Y_i - M_2)$  ... výběrová kovariance,  $R_{12} = \begin{cases} \frac{S_{12}}{S_1 S_2} \text{ pokud } S_1 S_2 \neq 0 \\ 0 \text{ jinak} \end{cases}$  ... výběrový koeficient korelace.

Pro libovolnou, ale pevně zvolenou dvojici reálných čísel  $x, y$  je statistikou též hodnota **výběrové simultánní distribuční funkce**  $F_n(x, y) = \frac{1}{n} \text{card}\{X_i \leq x \wedge Y_i \leq y\}$

**Upozornění:** Číselné realizace statistik  $M, S^2, S, S_{12}, R_{12}$  odpovídají číselným charakteristikám  $m, s^2, s, s_{12}, r_{12}$  zavedeným v popisné statistice, ale u rozptylu, směrodatné odchylky, kovariance a koeficientu korelace je multiplikační konstanta  $\frac{1}{n}$ , nikoliv  $\frac{1}{n-1}$ , jak tomu bylo v popisné statistice. Jak uvidíme později, uvedené číselné realizace mohou být považovány za odhady číselných realizací náhodných veličin zavedených v počtu pravděpodobnosti.

Charakteristika vlastnosti	Počet pravděpodobnosti	Matematická statistika	Popisná statistika
poloha	$E(X) = \mu$	$M$	$m$
variabilita	$D(X) = \sigma^2$	$S^2$	$\frac{n}{n-1} s^2$
variabilita	$\sqrt{D(X)} = \sigma$	$S$	$\sqrt{\frac{n}{n-1}} s$
společná variabilita	$C(X_1, X_2) = \sigma_{12}$	$S_{12}$	$\frac{n}{n-1} s_{12}$
těsnost vztahu	$R(X_1, X_2) = \rho$	$R_{12}$	$r_{12}$
rozložení	$\Phi(x)$	$F_n(x)$	$F(x)$

**Příklad** (výpočet realizací výběrového průměru, výběrového rozptylu a hodnot výběrové distribuční funkce):

Desetkrát nezávisle na sobě byla změřena jistá konstanta  $\mu$ . Výsledky měření byly: 2, 1,8, 2,1, 2,4, 1,9, 2,1, 2, 1,8, 2,3, 2,2. Tyto výsledky považujeme za číselné realizace náhodného výběru  $X_1, \dots, X_{10}$ . Vypočítejte realizaci  $m$  výběrového průměru  $M$ , realizaci  $s^2$  výběrového rozptylu  $S^2$ , realizaci  $s$  výběrové směrodatné odchylky  $S$  a hodnoty výběrové distribuční funkce  $F_{10}(x)$ .

**Řešení:**

$$m = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{10} (2 + 1,8 + 2,1 + 2,4 + 1,9 + 2,1 + 2 + 1,8 + 2,3 + 2,2) = 2,06$$

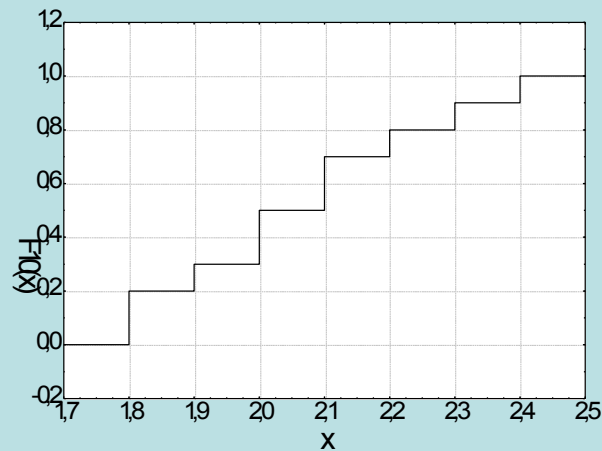
$$s^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \left( \frac{1}{n} \sum_{i=1}^n x_i \right)^2 = \frac{1}{10} (2^2 + 1,8^2 + 2,1^2 + 2,4^2 + 1,9^2 + 2,1^2 + 2^2 + 1,8^2 + 2,3^2 + 2,2^2) - 2,06^2 = 0,40$$

$$s = \sqrt{0,40} = 0,2011$$

Pro usnadnění výpočtu hodnot výběrové distribuční funkce  $F_{10}(x)$  uspořádáme měření podle velikosti:

1,8, 1,8, 1,9, 2, 2, 2,1, 2,1, 2,2, 2,3, 2,4.

- $x < 1,8$ :  $F_{10}(x) = 0$
- $1,8 \leq x < 1,9$ :  $F_{10}(x) = \frac{2}{10} = 0,2$
- $1,9 \leq x < 2,0$ :  $F_{10}(x) = \frac{3}{10} = 0,3$
- $2,0 \leq x < 2,1$ :  $F_{10}(x) = \frac{5}{10} = 0,5$
- $2,1 \leq x < 2,2$ :  $F_{10}(x) = \frac{7}{10} = 0,7$
- $2,2 \leq x < 2,3$ :  $F_{10}(x) = \frac{8}{10} = 0,8$
- $2,3 \leq x < 2,4$ :  $F_{10}(x) = \frac{9}{10} = 0,9$
- $x \geq 2,4$ :  $F_{10}(x) = 1$



### Příklad (výpočet realizace výběrového koeficientu korelace):

U 11 náhodně vybraných aut jisté značky bylo zjišťováno jejich stáří (náhodná veličina X – v letech) a cena (náhodná veličina Y – v tisících Kč). Výsledky:

(5, 85), (4, 103), (6, 70), (5, 82), (5, 89), (5, 98), (6, 66), (6, 95), (2, 169), (7, 70), (7, 48).

Vypočtěte a interpretujte číselnou realizaci  $r_{12}$  výběrového koeficientu korelace  $R_{12}$ .

### Řešení:

$$m_1 = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{11} (5 + 4 + 6 + 5 + 5 + 5 + 6 + 6 + 2 + 7 + 7) = 5,28$$

$$m_2 = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{11} (85 + 103 + 70 + 82 + 89 + 98 + 66 + 95 + 169 + 70 + 48) = 97,85$$

$$s_1^2 = \frac{1}{n} \sum_{i=1}^n (x_i - m_1)^2 = \frac{1}{11} (5^2 + 4^2 + 6^2 + 5^2 + 5^2 + 5^2 + 6^2 + 6^2 + 2^2 + 7^2 + 7^2) = 2,02$$

$$s_2^2 = \frac{1}{n} \sum_{i=1}^n (y_i - m_2)^2 = \frac{1}{11} (85^2 + 103^2 + 70^2 + 82^2 + 89^2 + 98^2 + 66^2 + 95^2 + 169^2 + 70^2 + 48^2) = 7,85$$

$$s_{12} = \frac{1}{n} \sum_{i=1}^n (x_i - m_1)(y_i - m_2) = \frac{1}{11} (5 \cdot 35 + 4 \cdot 6 + 6 \cdot 10 + 5 \cdot 10 + 5 \cdot 10 + 5 \cdot 10 + 6 \cdot 10 + 6 \cdot 10 + 2 \cdot 10 + 7 \cdot 10 + 7 \cdot 10) = 1,85$$

$$r_{12} = \frac{s_{12}}{s_1 \cdot s_2} = \frac{1,85}{\sqrt{2,02} \cdot \sqrt{7,85}} = -0,2$$

Mezi náhodnými veličinami X a Y existuje silná nepřímá lineární závislost. Čím starší auto, tím nižší cena.

## Bodové a intervalové odhady parametrů a parametrických funkcí

Vycházíme z náhodného výběru  $X_1, \dots, X_n$  z rozložení  $L(\theta)$ , které závisí na parametru  $\theta$ . Množinu všech přípustných hodnot tohoto parametru označíme  $\Xi$ . Tato množina se nazývá **parametrický prostor**.

Např. je-li  $X_1, \dots, X_n$  náhodný výběr z rozložení  $N(\mu, \sigma^2)$ , pak  $\theta = (\mu, \sigma^2)$  a v tomto případě parametrický prostor  $\Xi = \mathbb{R} \times \mathbb{R}_{>0}$ .

Parametr  $\theta$  neznáme a chceme ho odhadnout pomocí daného náhodného výběru (případně chceme odhadnout nějakou **parametrickou funkci**  $h(\theta)$ ).

**Bodovým odhadem** parametrické funkce  $h(\theta)$  je statistika  $T_n = T(X_1, \dots, X_n)$ , která nabývá hodnot blízkých  $h(\theta)$ , ať je hodnota parametru  $\theta$  jakákoliv. Existují různé metody, jak konstruovat bodové odhady (např. metoda momentů či metoda maximální věrohodnosti, ale těmi se zde zabývat nebudeme) a také různé typy bodových odhadů. Omezíme se na odhady nestranné, asymptoticky nestranné a konzistentní.

**Intervalovým odhadem** parametrické funkce  $h(\theta)$  rozumíme interval  $(D, H)$ , jehož meze jsou statistiky  $D = D(X_1, \dots, X_n)$ ,  $H = H(X_1, \dots, X_n)$  a který s dostatečně velkou pravděpodobností pokrývá  $h(\theta)$ , ať je hodnota parametru  $\theta$  jakákoliv.

## Typy bodových odhadů

Nechť  $X_1, \dots, X_n$  je náhodný výběr z rozložení  $L(\underline{Q})$ ,  $h(\underline{Q})$  je parametrická funkce,  $T, T_1, T_2, \dots$  jsou statistiky.

a) Řekneme, že statistika  $T$  je **nestranným odhadem** parametrické funkce  $h(\underline{Q})$ , jestliže

$$\forall \underline{Q} \quad E(T) = h(\underline{Q}).$$

(Význam nestrannosti spočívá v tom, že odhad  $T$  nesmí parametrickou funkci  $h(\underline{Q})$  systematicky nadhodnocovat ani podhodnocovat. Není-li tato podmínka splněna, jde o vychýlený odhad.)

b) Jsou-li  $T_1, T_2$  nestranné odhady téže parametrické funkce  $h(\underline{Q})$ , pak řekneme, že  $T_1$  je **lepší odhad** než  $T_2$ , jestliže

$$\forall \underline{Q} \quad D(T_1) < D(T_2).$$

c) Posloupnost  $\{T_n\}_{n=1}^{\infty}$  se nazývá **posloupnost asymptoticky nestranných odhadů** parametrické funkce  $h(\underline{Q})$ , jestliže

$$\forall \epsilon > 0 \quad \lim_{n \rightarrow \infty} P(|T_n - h(\underline{Q})| < \epsilon) = 1$$

(Význam asymptotické nestrannosti spočívá v tom, že s rostoucím rozsahem výběru klesá vychýlení odhadu.)

d) Posloupnost  $\{T_n\}_{n=1}^{\infty}$  se nazývá **posloupnost konzistentních odhadů** parametrické funkce  $h(\underline{Q})$ , jestliže

$$\forall \epsilon > 0 \quad \lim_{n \rightarrow \infty} P(|T_n - h(\underline{Q})| < \epsilon) = 1$$

(Význam konzistence spočívá v tom, že s rostoucím rozsahem výběru klesá pravděpodobnost, že odhad se bude realizovat „daleko“ od parametrické funkce  $h(\underline{Q})$ .)

Lze dokázat, že z nestrannosti odhadu vyplývá jeho asymptotická nestrannost a z asymptotické nestrannosti vyplývá konzistence, pokud posloupnost rozptylů odhadu konverguje k nule.

## Vlastnosti důležitých statistik

a) **Případ jednoho náhodného výběru:** Necht'  $X_1, \dots, X_n$  je náhodný výběr z rozložení se střední hodnotou  $\mu$ , rozptylem  $\sigma^2$  a distribuční funkcí  $\Phi(x)$ . Necht'  $n \geq 2$ . Označme  $M_n$  výběrový průměr,  $S_n^2$  výběrový rozptyl a pro libovolné, ale pevně dané  $X_{\sim}$  označme  $F_n(x)$  hodnotu výběrové distribuční funkce. Pak pro libovolné hodnoty parametrů  $\mu$ ,  $\sigma^2$  a libovolné, ale pevně dané reálné číslo  $x$  platí:

$$E(M_n) = \mu,$$

$$D(M_n) = \frac{\sigma^2}{n},$$

$$E(S_n^2) = \sigma^2,$$

$$D(S_n^2) = \frac{\gamma_4 - 3\sigma^4}{n}, \text{ kde } \gamma_4 \text{ je 4. centrální moment,}$$

$$E(F_n(x)) = \Phi(x),$$

$$D(F_n(x)) = \frac{\Phi(x)(1-\Phi(x))}{n}.$$

Znamená to, že  $M_n$  je nestranným odhadem  $\mu$ ,  $S_n^2$  je nestranným odhadem  $\sigma^2$ , pro libovolné, ale pevně dané  $X_{\sim}$  je výběrová distribuční funkce  $F_n(x)$  nestranným odhadem  $\Phi(x)$ .

Posloupnost  $\{M_{n_1}\}$  je posloupnost konzistentních odhadů  $\mu$ ,

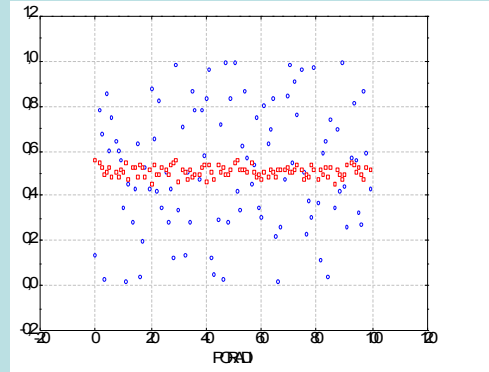
$\{S_{n_1}^2\}$  je posloupnost konzistentních odhadů  $\sigma^2$ ,

pro libovolné, ale pevně dané  $X_{\sim}$  je  $\{F_{n_1}(x)\}$  posloupnost konzistentních odhadů  $\Phi(x)$ .



## Ilustrace:

Vlastnosti výběrového průměru a výběrového rozptylu budeme ilustrovat na náhodném výběru rozsahu 100 z rozložení  $R_s(0,1)$ . V tomto případě  $E(X_i) = 1/2$ ,  $D(X_i) = 1/12$ ,  $i = 1, \dots, 100$ . Pomocí systému STATISTICA vygenerujeme pro každou z náhodných veličin  $X_1, \dots, X_{100}$  100 realizací a uložíme je do proměnných  $v_1, \dots, v_{100}$ . Dále vypočítáme průměr a rozptyl těchto realizací, uložíme je do proměnných PRUMER a ROZPTYL. Graficky znázorníme hodnoty některé z proměnných  $v_1, \dots, v_{100}$  (např.  $v_1$ ) a hodnoty proměnné PRUMER:



Vidíme, že hodnoty proměnné  $v_1$  kolísají od 0 do 1, zatímco hodnoty proměnné PRUMER se nacházejí v úzkém pásu kolem 1/2.

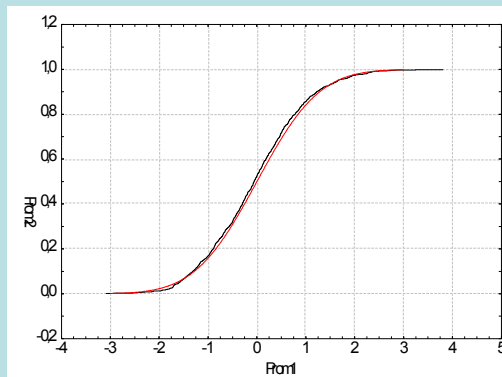
Dále vypočteme průměr a rozptyl např. proměnné  $v_1$  a proměnné PRUMER a dále vypočteme průměr proměnné ROZPTYL.

Promě	Popisné statistik	
	Prům	Rozpt
Prom1	0,536	0,078
PRUM	0,503	0,000

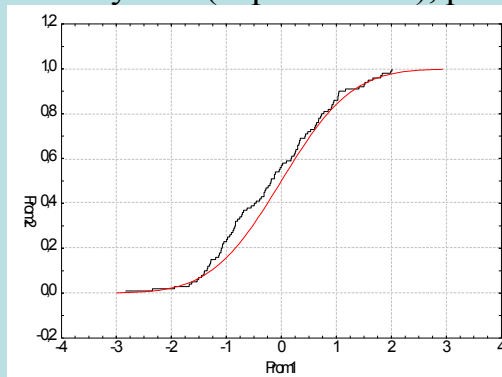
Promě	Popisné statistik	
	Prům	Rozpt
ROZP	0,083	

Průměr proměnné  $v_1$  by měl být blízký 0,5, rozptyl  $1/12 = 0,083$ . Průměr proměnné PRUMER by se měl blížit 0,5, zatímco rozptyl by měl být  $n = 100$  x menší než  $1/12$ , tj.  $0,00083$ . Dále průměr proměnné ROZPTYL by se měl blížit  $1/12 = 0,083$ .

Nestrannost výběrové distribuční funkce budeme ilustrovat na náhodném výběru rozsahu 1000 z rozložení  $N(0,1)$ . Získáme výběrovou distribuční funkci tohoto výběru a její graf porovnáme s grafem distribuční funkce náhodné veličiny se standardizovaným normálním rozložením. Graf výběrové distribuční funkce má černou barvu, graf distribuční funkce standardizovaného normálního rozložení má červenou barvu.



Průběh výběrové distribuční funkce  $F_{1000}(x)$  je velmi podobný průběhu distribuční funkce  $\Phi(x)$ . Pokud bychom postup zopakovali s podstatně menším rozsahem náhodného výběru (např.  $n = 100$ ), průběh obou funkcí by se lišil výrazněji:



b) **Případ  $r \geq 2$  stochasticky nezávislých náhodných výběrů:** Necht'  $X_{11}, \dots, X_{n_1}, \dots, X_{r1}, \dots, X_{n_r}$  je  $r$  stochasticky nezávislých náhodných výběrů o rozsazích  $n_1 \geq 2, \dots, n_r \geq 2$  z rozložení se středními hodnotami  $\mu_1, \dots, \mu_r$  a rozptylem  $\sigma^2$ . Celkový rozsah je  $n = \sum_{j=1}^r n_j$ . Necht'  $c_1, \dots, c_r$  jsou reálné konstanty, aspoň jedna nenulová. Pak pro libovolné hodnoty parametrů  $\mu_1, \dots, \mu_r$  a  $\sigma^2$

platí:

$$E\left(\sum_{j=1}^r c_j M_j\right) = \sum_{j=1}^r c_j \mu_j,$$

$$E(S_*^2) = \sigma^2.$$

Znamená to, že lineární kombinace výběrových průměrů  $\sum_{j=1}^r c_j M_j$  je nestranným odhadem lineární kombinace středních hod-

not  $\sum_{j=1}^r c_j \mu_j$  a vážený průměr výběrových rozptylů  $S_*^2 = \frac{\sum_{j=1}^r c_j^2 S_j^2}{\sum_{j=1}^r c_j^2}$  je nestranným odhadem rozptylu  $\sigma^2$ .

c) **Případ jednoho náhodného výběru z dvourozměrného rozložení:** Necht'  $(X_1, Y_1), \dots, (X_n, Y_n)$  je náhodný výběr z dvourozměrného rozložení s kovariancí  $\sigma_{12}$  a koeficientem korelace  $\rho$ . Pak pro libovolné hodnoty parametrů  $\sigma_{12}$  a  $\rho$  platí:

$$E(S_{12}) = \sigma_{12},$$

$$E(R_{12}) \approx \rho \quad (\text{shoda je vyhovující pro } n \geq 30).$$

Znamená to, že výběrová kovariance  $S_{12}$  je nestranným odhadem kovariance  $\sigma_{12}$ , avšak výběrový koeficient korelace  $R_{12}$  je vychýleným odhadem koeficientu korelace  $\rho$ .

## Pojem intervalu spolehlivosti

Nechť  $X_1, \dots, X_n$  je náhodný výběr z rozložení  $L(\theta)$ ,  
 $h(\theta)$  je parametrická funkce,  
 $\alpha \in (0,1)$ ,  
 $D = D(X_1, \dots, X_n)$ ,  $H = H(X_1, \dots, X_n)$  jsou statistiky.

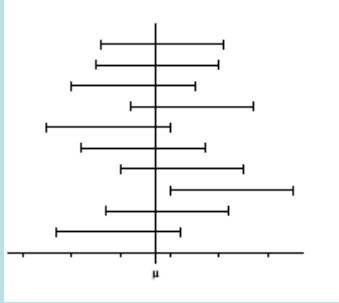
- a) Interval  $(D, H)$  se nazývá **100(1- $\alpha$ )% (oboustranný) interval spolehlivosti** pro parametrickou funkci  $h(\theta)$ ,  
jestliže:  $\forall \theta \quad P(D < h(\theta) < H) \geq 1-\alpha$ .
- b) Interval  $(D, \infty)$  se nazývá **100(1- $\alpha$ )% levostranný interval spolehlivosti** pro parametrickou funkci  $h(\theta)$ ,  
jestliže:  $\forall \theta \quad P(D < h(\theta)) \geq 1-\alpha$ .
- c) Interval  $(-\infty, H)$  se nazývá **100(1- $\alpha$ )% pravostranný interval spolehlivosti** pro parametrickou funkci  $h(\theta)$ ,  
jestliže:  $\forall \theta \quad P(h(\theta) < H) \geq 1-\alpha$ .

Číslo  $\alpha$  se nazývá **riziko** (zpravidla  $\alpha = 0,05$ , méně často 0,1 či 0,01), číslo  $1 - \alpha$  se nazývá **spolehlivost**.

## Postup při konstrukci intervalu spolehlivosti

- Vyjdeme ze statistiky  $V$ , která je nestranným bodovým odhadem parametrické funkce  $h(\theta)$ .
- Najdeme tzv. pivotovou statistiku  $W$ , která vznikne transformací statistiky  $V$ , je monotónní funkcí  $h(\theta)$  a přitom její rozložení je známé a na  $h(\theta)$  nezávisí. Pomocí známého rozložení pivotové statistiky  $W$  najdeme kvantily  $w_{\alpha/2}$ ,  $w_{1-\alpha/2}$ , takže platí:  $\forall \theta: P(w_{\alpha/2} < W < w_{1-\alpha/2}) \geq 1 - \alpha$ .
- Nerovnost  $w_{\alpha/2} < W < w_{1-\alpha/2}$  převedeme ekvivalentními úpravami na nerovnost  $D < h(\theta) < H$ .
- Statistiky  $D$ ,  $H$  nahradíme jejich číselnými realizacemi  $d$ ,  $h$  a získáme tak  $100(1-\alpha)\%$  empirický interval spolehlivosti, o němž prohlásíme, že pokrývá  $h(\theta)$  s pravděpodobností aspoň  $1 - \alpha$ . (Tvrzení, že  $(d, h)$  pokrývá  $h(\theta)$  s pravděpodobností aspoň  $1 - \alpha$  je třeba chápat takto: jestliže mnohonásobně nezávisle získáme realizace  $x_1, \dots, x_n$  náhodného výběru  $X_1, \dots, X_n$  z rozložení  $L(\theta)$  a pomocí každé této realizace sestojíme  $100(1-\alpha)\%$  empirický interval spolehlivosti pro  $h(\theta)$ , pak podíl počtu těch intervalů, které pokrývají  $h(\theta)$  k počtu všech sestojených intervalů bude přibližně  $1 - \alpha$ .)

**Ilustrace:** Jestliže 100x nezávisle na sobě uskutečníme náhodný výběr z rozložení se střední hodnotou  $\mu$  a pokaždé sestojíme 95% empirický interval spolehlivosti pro  $\mu$ , pak přibližně v 95-ti případech bude ležet parametr  $\mu$  v intervalech spolehlivosti a asi v 5-ti případech interval spolehlivosti  $\mu$  nepokryje.



Volba oboustranného, jednostranného, nebo jednostranného intervalu závisí na konkrétní situaci. Např. oboustranný interval spolehlivosti použije konstruktér, kterého zajímá dolní i horní hranice pro skutečnou délku  $\mu$  nějaké součástky. Jednostranný interval spolehlivosti použije výkupčí drahých kovů, který potřebuje znát dolní mez pro skutečný obsah zlata  $\mu$  v kupovaném slitku. Jednostranný interval spolehlivosti použije chemik, který potřebuje znát horní mez pro obsah nečistot  $\mu$  v analyzovaném vzorku.

**Příklad:** Necht'  $X_1, \dots, X_n$  je náhodný výběr z rozložení  $N(\mu, \sigma^2)$ , kde  $n \geq 2$  a rozptyl  $\sigma^2$  známe. Sestrojte  $100(1-\alpha)\%$  interval spolehlivosti pro neznámou střední hodnotu  $\mu$ .

**Řešení:** V tomto případě parametrická funkce  $h(\theta) = \mu$ . Nestranným odhadem střední hodnoty je výběrový průměr  $M = \frac{1}{n} \sum_{i=1}^n X_i$ . Protože  $M$  je lineární kombinací normálně rozložených náhodných veličin, bude mít také normální rozložení se

střední hodnotou  $E(M) = \mu$  a rozptylem  $D(M) = \frac{\sigma^2}{n}$ . Pivotovou statistikou  $W$  bude standardizovaná náhodná veličina

$$U = \frac{M - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1).$$

Kvantil  $w_{\alpha/2} = u_{\alpha/2} = -u_{1-\alpha/2}$ ,  $w_{1-\alpha/2} = u_{1-\alpha/2}$ .

$$\forall \mu: 1 - \alpha \leq P(-u_{1-\alpha/2} < U < u_{1-\alpha/2}) = P\left(-u_{1-\alpha/2} < \frac{M - \mu}{\frac{\sigma}{\sqrt{n}}} < u_{1-\alpha/2}\right) = P\left(M - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2} < M + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}\right).$$

Meze  $100(1-\alpha)\%$  intervalu spolehlivosti pro střední hodnotu  $\mu$  při známém rozptyle  $\sigma^2$  tedy jsou:

$$D = M - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}, H = M + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}.$$

Při konstrukci jednostranných intervalů spolehlivosti se riziko nepůlí, tedy  $100(1-\alpha)\%$  jednostranný interval spolehlivosti pro  $\mu$  je  $\left(M - \frac{\sigma}{\sqrt{n}} u_{1-\alpha}, \infty\right)$  a pravostranný je  $\left(-\infty, M + \frac{\sigma}{\sqrt{n}} u_{1-\alpha}\right)$ .

Dosadíme-li do vzorců pro dolní a horní mez číselnou realizaci  $m$  výběrového průměru  $M$ , dostaneme  $100(1-\alpha)\%$  empirický interval spolehlivosti. Postup si ukážeme na následujícím numerickém příkladu.

**Příklad:** 10 krát nezávisle na sobě byla změřena jistá konstanta  $\mu$ . Výsledky měření byly:

2 1,8 2,1 2,4 1,9 2,1 2 1,8 2,3 2,2.

Výsledky považujeme za číselné realizace náhodného výběru  $X_1, \dots, X_{10}$  z rozložení  $N(\mu, \sigma^2)$ , kde  $\mu$  neznáme a  $\sigma^2 = 0,04$ .

Najděte 95% empirický interval spolehlivosti pro  $\mu$ , a to

a) oboustranný,

b) levostranný,

c) pravostranný.

**Řešení:**

Vypočteme realizaci výběrového průměru:  $m = 2,06$ . Riziko  $\alpha$  je 0,05. V tabulkách najdeme kvantil  $u_{0,975} = 1,96$  pro oboustranný interval spolehlivosti a kvantil  $u_{0,95} = 1,64$  pro jednostranné intervaly spolehlivosti.

$$\text{ad a) } d = m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2} = 2,06 - \frac{0,2}{\sqrt{10}} 1,96 = 1,94$$

$$h = m + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2} = 2,06 + \frac{0,2}{\sqrt{10}} 1,96 = 2,18$$

$1,94 < \mu < 2,18$  s pravděpodobností aspoň 0,95.

$$\text{ad b) } d = m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha} = 2,06 - \frac{0,2}{\sqrt{10}} 1,64 = 1,96$$

$1,96 < \mu$  s pravděpodobností aspoň 0,95.

$$\text{ad c) } h = m + \frac{\sigma}{\sqrt{n}} u_{1-\alpha} = 2,06 + \frac{0,2}{\sqrt{10}} 1,64 = 2,16$$

$\mu < 2,16$  s pravděpodobností aspoň 0,95.

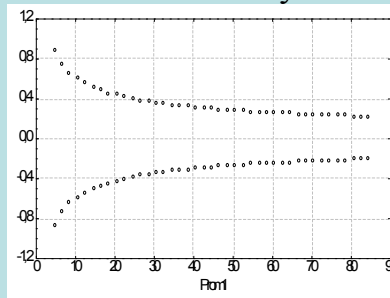
## Šířka intervalu spolehlivosti

Nechť  $(d, h)$  je  $100(1-\alpha)\%$  empirický interval spolehlivosti pro  $h(\hat{Q})$  zkonstruovaný pomocí číselných realizací  $x_1, \dots, x_n$  náhodného výběru  $X_1, \dots, X_n$  z rozložení  $L(\hat{Q})$ .

- Při konstantním riziku klesá šířka  $h-d$  s rostoucím rozsahem náhodného výběru.
- Při konstantním rozsahu náhodného výběru klesá šířka  $h-d$  s rostoucím rizikem.

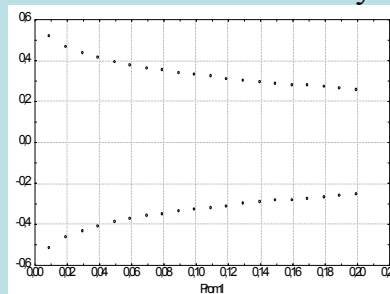
### Ilustrace

ad a) Grafické znázornění závislosti dolních a horních meze 95% empirických intervalů spolehlivosti pro střední hodnotu normálního rozložení při známém rozptylu na rozsahu náhodného výběru:



Šířka intervalu spolehlivosti klesá se zvětšujícím se rozsahem náhodného výběru, zprvu rychle a pak stále pomaleji.

ad b) Grafické znázornění závislosti dolních a horních mezí 100(1- $\alpha$ )% empirických intervalů spolehlivosti pro střední hodnotu normálního rozložení při známém rozptylu a konstantním rozsahu výběru na riziku:



Vidíme, že šířka intervalu spolehlivosti s rostoucím rizikem klesá.



**Příklad:** (stanovení minimálního rozsahu výběru z normálního rozložení)

Nechť  $X_1, \dots, X_n$  je náhodný výběr z  $N(\mu, \sigma^2)$ , kde  $\sigma^2$  známe. Jaký musí být minimální rozsah výběru  $n$ , aby šířka 100(1- $\alpha$ )% empirického intervalu spolehlivosti pro střední hodnotu  $\mu$  nepřesáhla číslo  $\Delta$ ?

**Řešení:** Požadujeme, aby  $\Delta \geq h - d = n \left( \frac{\sigma}{\sqrt{n}} + \frac{\sigma}{\sqrt{n}} \right) = \frac{2\sigma}{\sqrt{n}}$ . Z této podmínky dostaneme, že

$n \geq \frac{4\sigma^2}{\Delta^2}$ . Za rozsah výběru zvolíme nejmenší přirozené číslo vyhovující této podmínce.

**Příklad:** Hloubka moře se měří přístrojem, jehož systematická chyba je nulová a náhodné chyby měření mají normální rozložení se směrodatnou odchylkou  $\sigma = 1$  m. Kolik měření je nutno provést, aby se hloubka stanovila s chybou nejvýše  $\pm 0,25$  m při spolehlivosti 0,95?

**Řešení:** Hledáme rozsah výběru tak, aby šířka 95% intervalu spolehlivosti pro střední hodnotu  $\mu$  nepřesáhla 0,5 m. Přitom  $\sigma$  známe. Z předešlého příkladu vyplývá, že  $n \geq \frac{4\sigma^2}{\Delta^2} = \frac{4 \cdot 1^2}{0,5^2} = 16$ . Nejmenší počet měření je tedy 16.

# Základní pojmy matematické statistiky II

## Osnova:

Základní typy uspořádání pokusů

- jednoduché pozorování
- dvojné pozorování
- mnohonásobné pozorování

Úvod do testování hypotéz

- nulová a alternativní hypotéza
- chyba 1. a 2. druhu
- testování pomocí kritického oboru
- testování pomocí intervalu spolehlivosti
- testování pomocí p-hodnoty

Testování normality

- Kolmogorovův – Smirnovův test a jeho Lilieforsova varianta
- Shapiroův – Wilkův test
- srovnání S-W testu a Lilieforsova testu pomocí simulačních studií

## Základní typy uspořádání pokusů

Metody matematické statistiky často slouží k vyhodnocování výsledků pokusů. Aby mohl být pokus správně vyhodnocen, musí být dobře naplánován. Uvedeme zde nejjednodušší typy uspořádání pokusů.

Předpokládejme například, že sledujeme hmotnostní přírůstky selat téhož plemene při různých výkrmných dietách.

a) **Jednoduché pozorování:** Náhodná veličina  $X$  je pozorována za týchž podmínek. Situace je charakterizována jedním náhodným výběrem  $X_1, \dots, X_n$ .

Náhodně vylosujeme  $n$  selat téhož plemene, podrobíme je jediné výkrmné dietě a zjistíme u každého selete hmotnostní přírůstek. Tím dostaneme realizaci jednoho náhodného výběru.

b) **Dvojné pozorování:** Náhodná veličina  $X$  je pozorována za dvojích různých podmínek. Existují dvě odlišná uspořádání tohoto pokusu.

**Dvouvýběrové porovnávání:** situace je charakterizována dvěma nezávislymi náhodnými výběry  $X_{11}, \dots, X_{1n_1}$  a  $X_{21}, \dots, X_{2n_2}$ .

Náhodně vylosujeme  $n_1$  a  $n_2$  selat téhož plemene, náhodně je rozdělíme na dva soubory o  $n_1$  a  $n_2$  jedincích, první podrobíme výkrmné dietě č. 1 a druhý výkrmné dietě číslo 2. Tak dostaneme realizace dvou nezávislých náhodných výběrů.

**Párové porovnávání:** situace je charakterizována jedním náhodným výběrem  $(X_{11}, X_{12}), \dots, (X_{n1}, X_{n2})$  z dvourozměrného rozložení. Přejdeme k rozdílovému náhodnému výběru  $Z_i = X_{i1} - X_{i2}$ ,  $i = 1, \dots, n$  a tím dostaneme jednoduché pozorování.

Náhodně vylosujeme  $n$  vrhů stejně starých selat téhož plemene, z každého odebereme dva sourozence a náhodně jim přiřadíme první a druhou výkrmnou dietu. Tak dostaneme realizaci jednoho dvourozměrného náhodného výběru, kde první složka odpovídá první dietě a druhá složka druhé dietě.

(Párové porovnávání je efektivnější, protože skutečný rozdíl v účinnosti obou diet je překrýván pouze náhodnými vlivy při samotném krmení a trvání, kdežto vliv různých dědičných vloh, který byl losováním znáhodněn, je u sourozeneckého páru selat částečně vyloučen.)

c) **Mnohonásobné pozorování:** Náhodná veličina  $X$  je pozorována za  $r \geq 3$  různých podmínek. Existují dvě odlišná uspořádání tohoto pokusu.

**Mnohovýběrové porovnávání:** situace je charakterizována  $r$  nezávislými náhodnými výběry  $X_{11}, \dots, X_{1n_1}$  až  $X_{r1}, \dots, X_{rn_r}$ . Náhodně vylosujeme  $n_1, n_2, \dots, n_r$  selat téhož plemene, náhodně je rozdělíme na  $r$  souborů o  $n_1, n_2, \dots, n_r$  jedincích, první podrobíme výkrmné dietě č. 1, druhý výkrmné dietě číslo 2 atd. až  $r$ -tý podrobíme výkrmné dietě číslo  $r$ . Tak dostaneme realizace  $r$  nezávislých náhodných výběrů.

**Blokové porovnávání:** situace je charakterizována jedním náhodným výběrem  $(X_{11}, \dots, X_{1r}), \dots, (X_{n1}, \dots, X_{nr})$  z  $r$ -rozměrného rozložení.

Náhodně vylosujeme  $n$  vrhů stejně starých selat téhož plemene, z každého odebereme  $r$  sourozenců a náhodně jim přiřadíme první až  $r$ -tou výkrmnou dietu. Tak dostaneme realizaci jednoho  $r$ -rozměrného náhodného výběru, kde první složka odpovídá první dietě, druhá složka druhé dietě atd. až  $r$ -tá složka odpovídá  $r$ -té dietě.

## Úvod do testování hypotéz

**Motivace:** Častým úkolem statistika je na základě dat ověřit předpoklady o parametrech nebo typu rozložení, z něhož pochází náhodný výběr. Takovému předpokladu se říká nulová hypotéza. Nulová hypotéza vyjadřuje nějaký teoretický předpoklad, často skeptického rázu a uživatel ji musí stanovit předem, bez přihlídnutí k datovému souboru. Proti nulové hypotéze stavíme alternativní hypotézu, která říká, co platí, když neplatí nulová hypotéza. Alternativní hypotéza je formulována tak, aby mohla platit jenom jedna z těchto dvou hypotéz. Pravdivost alternativní hypotézy by znamenala objevení nějakých nových skutečností, nebo zásadnější změnu v dosavadních představách.

Např. výzkumník by chtěl na základě dat prověřit tezi (nový objev), že pasivní kouření škodí zdraví. Jako nulovou hypotézu tedy položí tvrzení, že pasivní kouření neškodí zdraví a proti nulové hypotéze postaví alternativní, že pasivní kouření škodí zdraví.

Testováním hypotéz se myslí rozhodovací postup, který je založen na daném náhodném výběru a s jehož pomocí rozhodneme o zamítnutí či nezamítnutí nulové hypotézy.

## Nulová a alternativní hypotéza

Nechť  $X_1, \dots, X_n$  je náhodný výběr z rozložení  $L(\theta)$ , kde parametr  $\theta$  neznáme. Nechť  $h(\theta)$  je parametrická funkce a  $c$  daná reálná konstanta.

a) **Oboustranná alternativa:** Tvrzení  $H_0: h(\theta) = c$  se nazývá **jednoduchá nulová hypotéza**. Proti nulové hypotéze postavíme **složenou oboustrannou alternativní hypotézu**  $H_1: h(\theta) \neq c$ .

b) **Levostranná alternativa:** Tvrzení  $H_0: h(\theta) \geq c$  se nazývá **složená pravostranná nulová hypotéza**. Proti jednoduché nebo složené pravostranné nulové hypotéze postavíme **složenou levostrannou alternativní hypotézu**  $H_1: h(\theta) < c$ .

c) **Pravostranná alternativa:** Tvrzení  $H_0: h(\theta) \leq c$  se nazývá **složená levostranná nulová hypotéza**. Proti jednoduché nebo složené levostranné nulové hypotéze postavíme **složenou pravostrannou alternativní hypotézu**  $H_1: h(\theta) > c$ .

**Testováním  $H_0$  proti  $H_1$**  rozumíme rozhodovací postup založený na náhodném výběru  $X_1, \dots, X_n$ , s jehož pomocí zamítneme či nezamítneme platnost nulové hypotézy.

## Chyba 1. a 2. druhu

Při testování  $H_0$  proti  $H_1$  se můžeme dopustit jedné ze dvou chyb: **chyba 1. druhu** spočívá v tom, že  $H_0$  zamítneme, ač ve skutečnosti platí a **chyba 2. druhu** spočívá v tom, že  $H_0$  nezamítneme, ač ve skutečnosti neplatí. Situaci přehledně znázorňuje tabulka:

skutečnost	rozhodnutí	
	$H_0$ nezamítáme	$H_0$ zamítáme
$H_0$ platí	správné rozhodnutí	chyba 1. druhu
$H_0$ neplatí	chyba 2. druhu	správné rozhodnutí

Pravděpodobnost chyby 1. druhu se značí  $\alpha$  a nazývá se **hladina významnosti testu** (většinou bývá  $\alpha = 0,05$ , méně často 0,1 či 0,01). Pravděpodobnost chyby 2. druhu se značí  $\beta$ . Číslo  $1-\beta$  se nazývá **síla testu** a vyjadřuje pravděpodobnost, že bude  $H_0$  zamítnuta za předpokladu, že neplatí. Obvykle se snažíme, aby síla testu byla aspoň 0,8. Obě hodnoty,  $\alpha$  i  $1-\beta$ , závisí na velikosti efektu, který se snažíme detekovat. Čím drobnější efekt, tím musí být větší rozsah náhodného výběru.

skutečnost	rozhodnutí	
	zdravý	nemocný
jsem zdravý	zdravý a neléčený	zdravý a léčený
jsem nemocný	nemocný a neléčený	nemocný a léčený



## Testování pomocí kritického oboru

Najdeme statistiku  $T_0 = T_0(X_1, \dots, X_n)$ , kterou nazveme **testovým kritériem**. Množina všech hodnot, jichž může testové kritérium nabýt, se rozpadá na **obor nezamítnutí nulové hypotézy** (značí se V) a **obor zamítnutí nulové hypotézy** (značí se W a nazývá se též **kritický obor**). Tyto dva obory jsou odděleny kritickými hodnotami (pro danou hladinu významnosti  $\alpha$  je lze najít ve statistických tabulkách).

Jestliže číselná realizace  $t_0$  testového kritéria  $T_0$  padne do kritického oboru W, pak nulovou hypotézu zamítáme na hladině významnosti  $\alpha$  a znamená to skutečné vyvrácení testované hypotézy. Jestliže  $t_0$  padne do oboru nezamítnutí V, pak jde o pouhé mlčení, které platnost nulové hypotézy jenom připouští.

Pravděpodobnosti chyb 1. a 2. druhu nyní zapíšeme takto:

$$P(T_0 \in W/H_0 \text{ platí}) = \alpha, P(T_0 \in V/H_1 \text{ platí}) = \beta.$$

Stanovení kritického oboru pro danou hladinu významnosti  $\alpha$ :

Označme  $t_{\min}$  (resp.  $t_{\max}$ ) nejmenší (resp. největší) hodnotu testového kritéria.

Kritický obor v případě oboustranné alternativy má tvar

$W = \{t_{\min}, K_{\alpha/2}(T) \cup (K_{1-\alpha/2}(T), t_{\max}]\}$ , kde  $K_{\alpha/2}(T)$  a  $K_{1-\alpha/2}(T)$  jsou kvantily rozložení, jímž se řídí testové kritérium  $T_0$ , je-li nulová hypotéza pravdivá.

Kritický obor v případě levostranné alternativy má tvar:

$$W = \{t_{\min}, K_{\alpha}(T)\}.$$

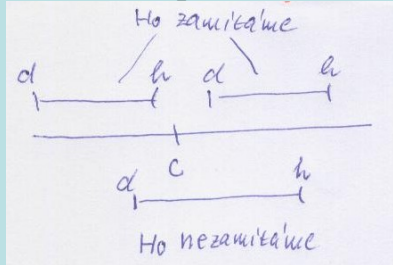
Kritický obor v případě pravostranné alternativy má tvar:

$$W = \{K_{1-\alpha}(T), t_{\max}\}.$$

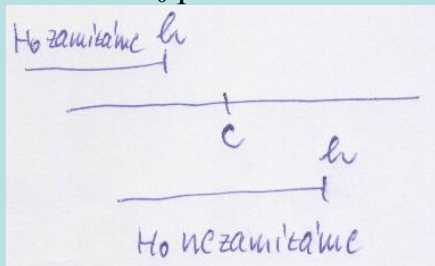
## Testování pomocí intervalu spolehlivosti

Sestrojíme  $100(1-\alpha)\%$  empirický interval spolehlivosti pro parametrickou funkci  $h(\theta)$ . Pokryje-li tento interval hodnotu  $c$ , pak  $H_0$  nezamítáme na hladině významnosti  $\alpha$ , v opačném případě  $H_0$  zamítáme na hladině významnosti  $\alpha$ .

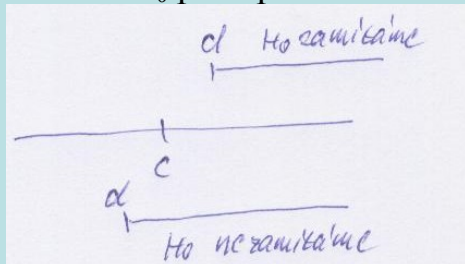
Pro test  $H_0$  proti oboustranné alternativě sestrojíme oboustranný interval spolehlivosti.



Pro test  $H_0$  proti levostranné alternativě sestrojíme pravostranný interval spolehlivosti.



Pro test  $H_0$  proti pravostranné alternativě sestrojíme levostranný interval spolehlivosti.



## Testování pomocí p-hodnoty

**p-hodnota** udává nejnižší možnou hladinu významnosti pro zamítnutí nulové hypotézy. Je to riziko, že bude zamítnuta  $H_0$  za předpokladu, že platí (riziko planého poplachu). Jestliže  $p\text{-hodnota} \leq \alpha$ , pak  $H_0$  zamítáme na hladině významnosti  $\alpha$ , je-li  $p\text{-hodnota} > \alpha$ , pak  $H_0$  nezamítáme na hladině významnosti  $\alpha$ .

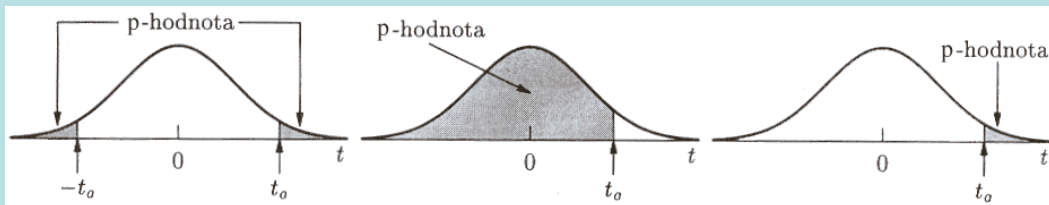
**Způsob výpočtu p-hodnoty:**

Pro oboustrannou alternativu  $p = 2 \min \{P(T_0 \leq t_0), P(T_0 \geq t_0)\}$ .

Pro levostrannou alternativu  $p = P(T_0 \leq t_0)$ .

Pro pravostrannou alternativu  $p = P(T_0 \geq t_0)$ .

Ilustrace významu p-hodnoty pro test nulové hypotézy proti oboustranné, levostranné a pravostranné alternativě:



(Zvonovitá křivka reprezentuje hustotu rozložení, kterým se řídí testové kritérium, je-li nulová hypotéza pravdivá.)

p-hodnota vyjadřuje pravděpodobnost, s jakou číselné realizace  $x_1, \dots, x_n$  náhodného výběru  $X_1, \dots, X_n$  podporují  $H_0$ , je-li pravdivá. Statistické programové systémy poskytují ve svých výstupech p-hodnotu. Její výpočet vyžaduje znalost distribuční funkce rozložení, kterým se řídí testové kritérium  $T_0$ , je-li  $H_0$  pravdivá.

## Doporučený postup při testování hypotéz

1. Stanovíme nulovou hypotézu a alternativní hypotézu. Přitom je vhodné zvolit jako alternativní hypotézu ten předpoklad, jehož přijetí znamená závažné opatření a mělo by k němu dojít jen s malým rizikem omylu.
2. Zvolíme hladinu významnosti  $\alpha$ . Zpravidla volíme  $\alpha = 0,05$ , méně často 0,1 nebo 0,01.
3. Najdeme vhodné testové kritérium a na základě zjištěných dat vypočítáme jeho realizaci.
4.
  - a) Testujeme-li pomocí kritického oboru, pak ho stanovíme. Jestliže realizace testového kritéria padla do kritického oboru, nulovou hypotézu zamítáme na hladině významnosti  $\alpha$  a přijímáme alternativní hypotézu. V opačném případě nulovou hypotézu nezamítáme na hladině významnosti  $\alpha$ .
  - b) Testujeme-li pomocí intervalu spolehlivosti, vypočteme empirický  $100(1-\alpha)\%$  interval spolehlivosti pro parametrickou funkci  $h(\mathbf{Q})$ . Pokud číslo  $c$  padne do tohoto intervalu, nulovou hypotézu nezamítáme na hladině významnosti  $\alpha$ . V opačném případě nulovou hypotézu zamítáme na hladině významnosti  $\alpha$  a přijímáme alternativní hypotézu.
  - c) Testujeme-li pomocí p-hodnoty, vypočteme ji a porovnáme ji s hladinou významnosti  $\alpha$ . Jestliže  $p \leq \alpha$ , pak nulovou hypotézu zamítáme na hladině významnosti  $\alpha$  a přijímáme alternativní hypotézu. Je-li  $p > \alpha$ , pak nulovou hypotézu nezamítáme na hladině významnosti  $\alpha$ .
5. Na základě rozhodnutí, které jsme učinili o nulové hypotéze, provedeme nějaké konkrétní opatření, např. seřídíme obráběcí stroj.

(Při testování hypotéz musíme mít k dispozici odpovídající nástroje, nejlépe vhodný statistický software. Nemáme-li ho k dispozici, musíme znát příslušné vzorce. Dále potřebujeme statistické tabulky a kalkulačku.)

**Příklad:** 10 x nezávisle na sobě byla změřena jistá konstanta  $\mu$ . Výsledky měření byly: 2,1, 1,8, 2,1, 2,4, 1,9, 2,1, 2,1, 1,8, 2,3, 2,2. Tyto výsledky považujeme za číselné realizace náhodného výběru  $X_1, \dots, X_{10}$  z rozložení  $N(\mu, 0,04)$ . Nějaká teorie tvrdí, že  $\mu = 1,95$ .

### 1. Oboustranná alternativa

Proti nulové hypotéze  $H_0: \mu = 1,95$  postavíme oboustrannou alternativu

$H_1: \mu \neq 1,95$ . Na hladině významnosti 0,05 testujte  $H_0$  proti  $H_1$  všemi třemi popsánymi způsoby.

### Řešení:

$$m = \frac{1}{10}(2,1 + 1,8 + 2,1 + 2,4 + 1,9 + 2,1 + 2,1 + 1,8 + 2,3 + 2,2) = 2,06, \sigma^2 = 0,04, n = 10, \alpha = 0,05, c = 1,95$$

a) Test provedeme pomocí kritického oboru.

Pro úlohy o střední hodnotě normálního rozložení při známém rozptylu používáme pivotovou statistiku  $U = \frac{\bar{M} - c}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$ .

Testové kritérium tedy bude

$$T_0 = \frac{\bar{M} - c}{\frac{\sigma}{\sqrt{n}}}$$

a bude mít rozložení  $N(0, 1)$ , pokud je nulová hypotéza pravdivá. Vypočítáme realizaci testového kritéria:

$$t_0 = \frac{2,06 - 1,95}{\frac{0,2}{\sqrt{10}}} = 1,74. \text{ Stanovíme kritický obor:}$$

$$W = \{ \min_{\alpha} K_{\alpha}^{-1}(T) \cup \max_{\alpha} K_{\alpha}^{-1}(T) \} = (-\infty, u_{\alpha/2}) \cup (u_{1-\alpha/2}, \infty) = (-\infty, u_{0,975}) \cup (u_{0,025}, \infty) = (-\infty, 1,96) \cup (1,96, \infty)$$

Protože  $1,74 \notin W$ ,  $H_0$  nezamítáme na hladině významnosti 0,05.

b) Test provedeme pomocí intervalu spolehlivosti.

Meze  $100(1-\alpha)\%$  empirického intervalu spolehlivosti pro střední hodnotu  $\mu$  při známém rozptylu  $\sigma^2$  jsou:

$$(d, h) = \left( m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}, m + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2} \right).$$

V našem případě dostáváme:

$$d = 2,06 - \frac{0,2}{\sqrt{10}} u_{0,975} = 2,06 - \frac{0,2}{\sqrt{10}} \cdot 1,96 = 1,936,$$

$$h = 2,06 + \frac{0,2}{\sqrt{10}} u_{0,975} = 2,06 + \frac{0,2}{\sqrt{10}} \cdot 1,96 = 2,184.$$

Protože  $1,95 \notin (1,936; 2,184)$ ,  $H_0$  nezamítáme na hladině významnosti 0,05.

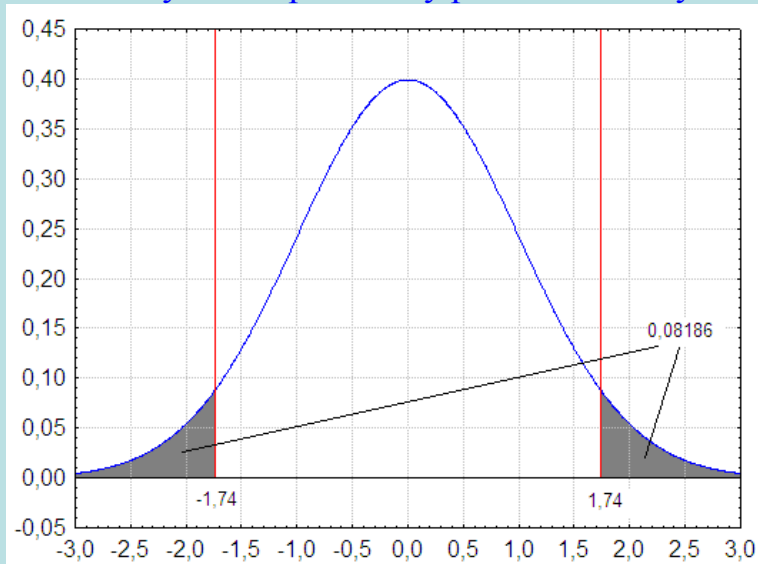
c) Test provedeme pomocí p-hodnoty.

Protože proti nulové hypotéze stavíme oboustrannou alternativu, použijeme vzorec

$$p = 2 \min \{P(T_0 \leq t_0), P(T_0 \geq t_0)\} = 2 \min \{P(T_0 \leq 1,74), P(T_0 \geq 1,74)\} = \\ = 2 \min \{ \Phi(1,74), 1 - \Phi(1,74) \} = 2 \min \{ 0,95907, 1 - 0,95907 \} = 0,08186.$$

Jelikož  $0,08186 > 0,05$ , nulovou hypotézu nezamítáme na hladině významnosti  $0,05$ .

Ilustrace významu p-hodnoty pro oboustranný test



## 2. Levostranná alternativa

Proti nulové hypotéze  $H_0: \mu = 1,95$  postavíme levostrannou alternativu

$H_1: \mu < 1,95$ . Na hladině významnosti 0,05 testujte  $H_0$  proti  $H_1$  všemi třemi popsánymi způsoby.

### Řešení:

a) Test provedeme pomocí kritického oboru.

Na rozdíl od oboustranné alternativy bude mít kritický obor tvar

$$W = \langle -\infty, u_{1-\alpha} \rangle = \langle -\infty, u_{0,95} \rangle = \langle -\infty, -1,64 \rangle$$

Protože  $1,74 \notin W$ ,  $H_0$  nezamítáme na hladině významnosti 0,05.

b) Test provedeme pomocí intervalu spolehlivosti.

Meze  $100(1-\alpha)\%$  empirického pravostranného intervalu spolehlivosti pro střední hodnotu  $\mu$  při známém rozptylu  $\sigma^2$  jsou:

$$(-\infty, h) = (-\infty, m + \frac{\sigma}{\sqrt{n}} u_{1-\alpha}).$$

$$\text{V našem případě dostáváme: } h = 2,06 + \frac{0,2}{\sqrt{10}} u_{0,95} = 2,06 + \frac{0,2}{\sqrt{10}} \cdot 1,645 = 2,164.$$

Protože  $1,95 \notin (-\infty; 2,164)$ ,  $H_0$  nezamítáme na hladině významnosti 0,05.



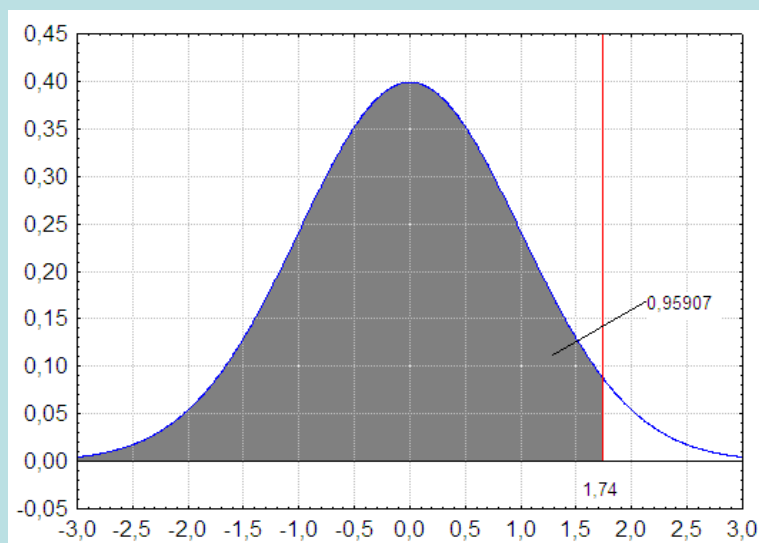
c) Test provedeme pomocí p-hodnoty.

Protože proti nulové hypotéze stavíme levostrannou alternativu, použijeme vzorec

$$p = P(T_0 \leq t_0) = \Phi(1,74) = 0,95907.$$

Jelikož  $0,95907 > 0,05$ , nulovou hypotézu nezamítáme na hladině významnosti 0,05.

Ilustrace významu p-hodnoty pro levostranný test



### 3. Pravostranná alternativa

Proti nulové hypotéze  $H_0: \mu = 1,95$  postavíme pravostrannou alternativu

$H_1: \mu > 1,95$ . Na hladině významnosti 0,05 testujte  $H_0$  proti  $H_1$  všemi třemi popsanými způsoby.

#### Řešení:

a) Test provedeme pomocí kritického oboru.

Na rozdíl od oboustranné alternativy bude mít kritický obor tvar

$$W = \langle u_{1-\alpha} | \infty \rangle = \langle u_{0,95} | \infty \rangle = \langle 1,645 | \infty \rangle.$$

Protože  $1,74 \notin W$ ,  $H_0$  zamítáme na hladině významnosti 0,05 ve prospěch pravostranné alternativy.

b) Test provedeme pomocí intervalu spolehlivosti.

Meze  $100(1-\alpha)\%$  empirického jednostranného intervalu spolehlivosti pro střední hodnotu  $\mu$  při známém rozptylu  $\sigma^2$  jsou:

$$(d, \infty) = \left( m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha}, \infty \right).$$

$$\text{V našem případě dostáváme: } d = 2,06 - \frac{0,2}{\sqrt{10}} u_{0,95} = 2,06 - \frac{0,2}{\sqrt{10}} \cdot 1,645 = 1,956.$$

Protože  $1,95 \notin (1,956, \infty)$ ,  $H_0$  zamítáme na hladině významnosti 0,05 ve prospěch pravostranné alternativy.

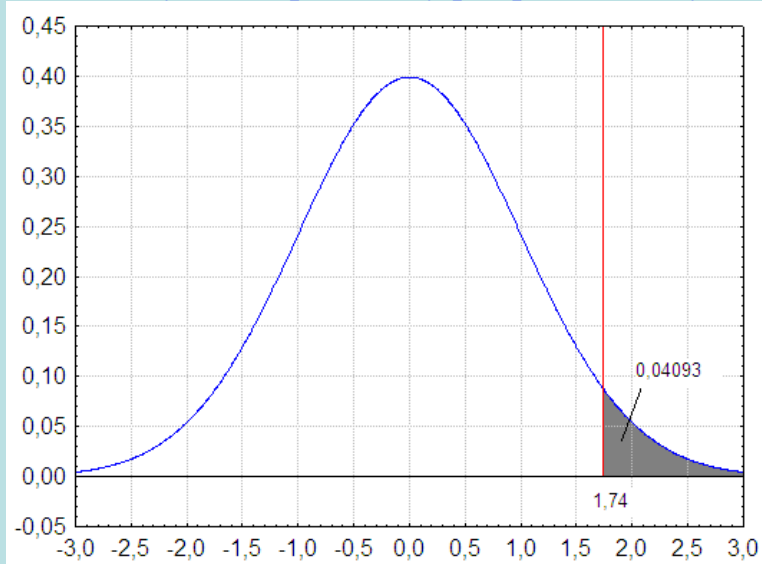
c) Test provedeme pomocí p-hodnoty.

Protože proti nulové hypotéze stavíme pravostrannou alternativu, použijeme vzorec

$$p = P(T_0 \geq t_0) = 1 - \Phi(1,74) = 1 - 0,95907 = 0,04093.$$

Jelikož  $0,04093 \leq 0,05$ , nulovou hypotézu zamítáme na hladině významnosti 0,05 ve prospěch pravostranné alternativy.

Ilustrace významu p-hodnoty pro pravostranný test



## Testy normality dat

K ověřování normality dat slouží celá řada testů, které jsou podrobně popsány ve statistické literatuře. Zde se omezíme na dva testy, které jsou implementovány v systému STATISTICA, a to Kolmogorovův – Smirnovův test a jeho Lilieforsovu variantu a Shapirův – Wilksův test. K závěrům těchto testů však přistupujeme s určitou opatrností. Máme-li k dispozici rozsáhlejší datový soubor (orientačně  $n > 30$ ) a test zamítne na obvyklé hladině významnosti 0,01 nebo 0,05 hypotézu o normalitě, i když vzhled diagnostických grafů svědčí jenom o lehkém porušení normality, nedopustíme se závažné chyby, pokud použijeme statistickou metodu založenou na normalitě dat.

### Kolmogorovův – Smirnovův test a jeho Lilieforsova varianta

Testujeme hypotézu, která tvrdí, že náhodný výběr  $X_1, \dots, X_n$  pochází z normálního rozložení s parametry  $\mu$  a  $\sigma^2$ .

Distribuční funkci tohoto rozložení označme  $\Phi_T(x)$ .

Necht'  $F_n(x)$  je výběrová distribuční funkce.

Testovou statistikou je statistika  $D_n = \sup_x |F_n(x) - \Phi_T(x)|$ .

Nulovou hypotézu zamítáme na hladině významnosti  $\alpha$ , když  $D_n \geq D_n(\alpha)$ , kde  $D_n(\alpha)$  je tabelovaná kritická hodnota.

Pro  $n \geq 30$  lze  $D_n(\alpha)$  aproximovat výrazem  $\sqrt{\frac{1}{2n} \ln \frac{2}{\alpha}}$ .

V případě, že neznáme parametry  $\mu$  a  $\sigma^2$  normálního rozložení, musíme je odhadnout z dat (střední hodnotu odhadneme pomocí  $m$  a rozptyl pomocí  $s^2$ ). Tím se změní rozložení testové statistiky  $D_n$ . Příslušné modifikované kvantily byly určeny pomocí simulačních studií. V této situaci používáme **Lilieforsovu variantu Kolmogorovova – Smirnovova testu**.

## Shapirův – Wilksův test normality dat

Testujeme hypotézu, která tvrdí, že náhodný výběr  $X_1, \dots, X_n$  pochází z normálního rozložení  $N(\mu, \sigma^2)$ .

Testová statistika má tvar:

$$W = \frac{\sum_{i=1}^m a_i^{(n)} (X_{(i)} - \bar{X})^2}{\sum_{i=1}^m (X_i - \bar{X})^2},$$

kde  $m = n/2$  pro  $n$  sudé a  $m = (n-1)/2$  pro  $n$  liché. Koeficienty  $a_i^{(n)}$  jsou tabelovány.

Na testovou statistiku  $W$  lze pohlížet jako na korelační koeficient mezi uspořádanými pozorováními a jim odpovídajícími kvantily standardizovaného normálního rozložení. V případě, že data vykazují perfektní shodu s normálním rozložením, bude mít  $W$  hodnotu 1. Hypotézu o normalitě tedy zamítneme na hladině významnosti  $\alpha$ , když se na této hladině neprokáže korelace mezi daty a jim odpovídajícími kvantily rozložení  $N(0,1)$ .

Lze také říci, že  $S - W$  test je založen na zjištění, zda body v Q-Q grafu jsou významně odlišné od regresní přímky proložené těmito body.

( $S-W$  test se používá především pro výběry menších rozsahů,  $n < 50$ , ale v systému STATISTICA je implementováno jeho rozšíření i na výběry velkých rozsahů, kolem 2000.)

### Příklad:

Jsou dány hodnoty 10, 12, 8, 9, 16. Pomocí Lilieforsova testu a S – W testu zjistěte na hladině významnosti 0,05, zda tato data pocházejí z normálního rozložení.

### Řešení:

Vytvoříme nový datový soubor o jedné proměnné nazvané X a pěti případech. Do proměnné X zapíšeme uvedené hodnoty. V menu vybereme Statistika – Základní statistiky/tabulky – Tabulky četností – OK, Proměnné X – OK. Na záložce zvolíme Normalita a zaškrtneme Lilieforsův test a Shapiro – Wilksův W test – Testy normality.

Proměnná	Testy normality (Tabulka1)			
	N	max  Lilliefors	W	p
X	5	0,224	0,912	0,482

Vidíme, že testová statistika K-S testu je  $d = 0,22409$ , odpovídající Lilieforsova p-hodnota je větší než 0,2, tedy hypotézu o normalitě nezamítáme na hladině významnosti 0,05.

Testová statistika S-W testu je  $W = 0,9124$ , odpovídající p-hodnota je 0,48215, tedy hypotézu o normalitě nezamítáme na hladině významnosti 0,05.

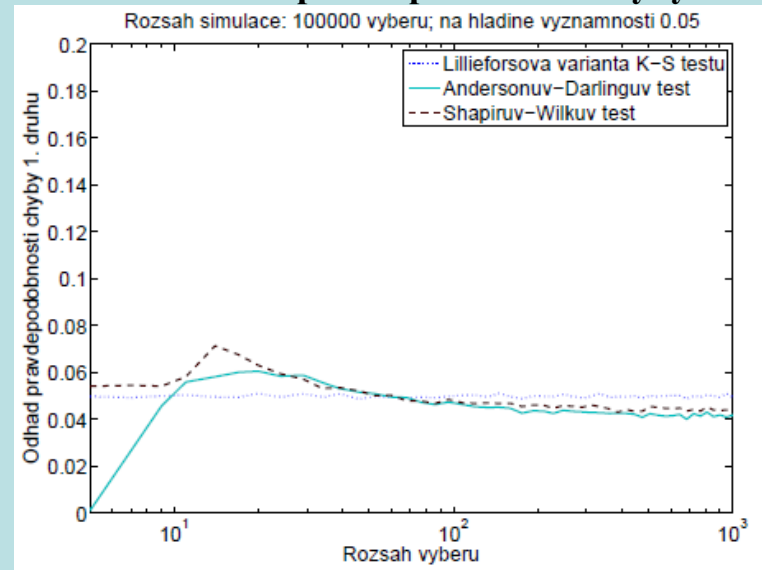
## Srovnání S-W testu a Lilieforsovy varianty K-S testu pomocí simulačních studií

Simulační studie byly provedeny v bakalářské práci **Marka Haičmana Simulace a testy normality** (zde byl zkoumán i Andersonův – Darlingův test).

### Odhad pravděpodobnosti chyby 1. druhu

Bylo vygenerováno 100 000 náhodných výběrů z normálního rozložení, jejichž rozsahy se pohybovaly od 5 do 1000. Na tyto výběry byly aplikovány oba testy (s hladinou významnosti 0,05) a byla stanovena relativní četnost těch případů, kdy došlo k neoprávněnému zamítnutí pravdivé nulové hypotézy. Tato relativní četnost je považována za odhad pravděpodobnosti chyby 1. druhu.

### Závislost odhadu pravděpodobnosti chyby 1. druhu na rozsahu výběru (hodnoty na vodorovné ose jsou logaritmovány)



### Výsledek:

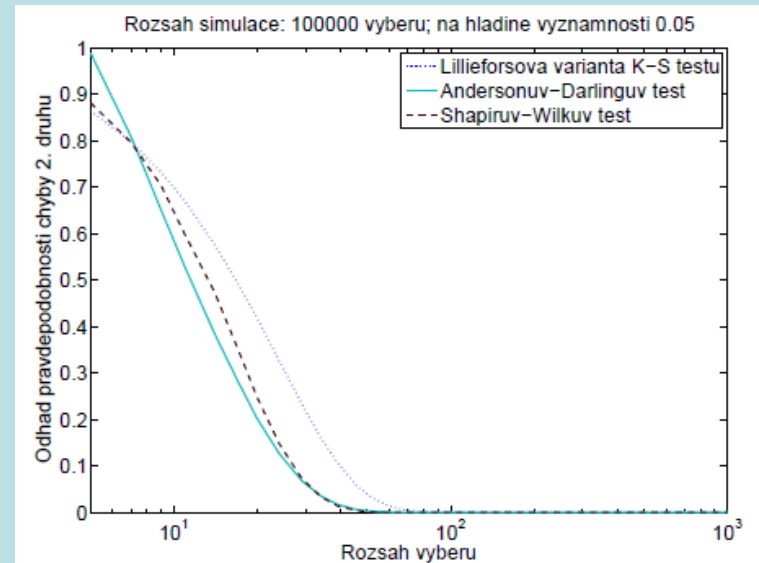
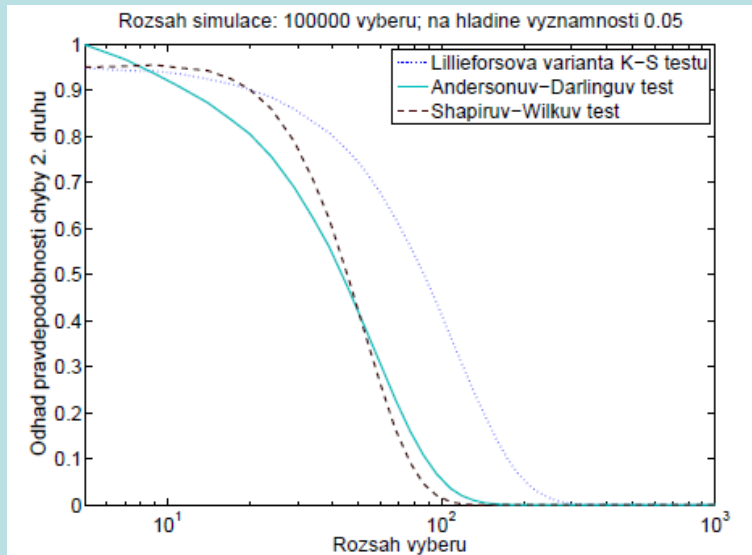
Lilieforsův test má pravděpodobnost chyby 1. druhu nezávislou na rozsahu výběru, udržuje se na 5 %.

S-W test má do velikosti výběru 60 vyšší pravděpodobnost chyby 1. druhu, poté poklesne pod 5 % a již nevystoupí nad 5 %.

## Odhad pravděpodobnosti chyby 2. druhu

Pro toto zkoumání byla vybrána následující rozložení: rovnoměrné spojité, exponenciální, logaritmicko – normální, Studentovo s jedním, třemi a pěti stupni volnosti. Pro každé z těchto rozložení bylo vygenerováno 100 000 náhodných výběrů o rozsazích 5 až 1 000. Při aplikaci S-W testu a Lilieforsova testu byla zjišťována relativní četnost těch případů, kdy test nezamítl nepravdivou nulovou hypotézu. Tato relativní četnost je považována za odhad pravděpodobnosti chyby 2. druhu.

**Ilustrace pro rovnoměrné spojité rozložení a exponenciální rozložení: závislost odhadu pravděpodobnosti chyby 2. druhu na rozsahu výběru (hodnoty na vodorovné ose jsou logaritmovány)**



## Výsledek:

Lilieforsův test nejméně chybuje u velmi malých výběrů, orientačně do 10 prvků.

S-W test se pro výběry větších rozsahů (nad 60) vesměs nedopouští chyby. K chybám však dochází i pro velmi rozsáhlé výběry ze Studentova rozložení.



### Stanovení hranice 20 % odhadu pravděpodobnosti chyby 2. druhu

Zde byl hledán rozsah výběru z rovnoměrného, exponenciálního, logaritmicko – normálního a Studentova rozložení tak, aby odhadu pravděpodobnosti chyby 2. druhu byl nanejvýš 20 %.

**Tabulka minimálních rozsahů výběrů, pro něž je odhad pravděpodobnosti chyby 2. druhu nejvýše 20 %:**

Test normality	Norm	Rovno.	Expo.	Logn.	Stud(1)	Stud(3)	Stud(5)
Anderson-Darling	–	72	21	15	16	87	247
Lilliefors	–	143	32	21	18	121	377
Shapiro-Wilk	–	65	22	17	19	89	221

### Výsledek:

S-W test je možno použít na výběry menších rozsahů než Lillieforsův test.

U výběrů, jejichž rozsah je menší než 15, nemá příliš smysl testovat hypotézu o normalitě, neboť pravděpodobnost chyby 2. druhu je příliš vysoká (nad 70 %).

## Parametrické úlohy o jednom náhodném výběru z normálního rozložení

### Motivace:

K nejčastěji používaným statistickým metodám patří konstrukce intervalů spolehlivosti pro parametry normálního rozložení či testování hypotéz o těchto parametrech. Normální rozložení je charakterizováno dvěma parametry – střední hodnotou  $\mu$  a rozptylem  $\sigma^2$ . Budeme tedy řešit úlohy, které se týkají těchto dvou parametrů. K tomu slouží např. jednovýběrový z-test, t-test či test o rozptylu. Můžeme také mít k dispozici náhodný výběr z dvourozměrného rozložení s vektorem středních hodnot

$\begin{pmatrix} \mu^1 \\ \mu^2 \end{pmatrix}$  a naším úkolem bude posoudit rozdílnost středních hodnot  $\begin{pmatrix} \mu^1 \\ \mu^2 \end{pmatrix}$ . K řešení tohoto problému slouží párový t-test.

### Osnova:

- rozložení statistik odvozených z výběrového průměru a výběrového rozptylu
- vzorce pro meze intervalů spolehlivosti pro střední hodnotu a rozptyl
- jednotlivé typy testů pro parametry normálního rozložení  
(z-test, jednovýběrový t-test, test o rozptylu, párový t-test)

**Rozložení statistik odvozených z výběrového průměru a výběrového rozptylu**

Nechť  $X_1, \dots, X_n$  je náhodný výběr z rozložení  $N(\mu, \sigma^2)$ . Pak platí

a)  $M \sim N(\mu, \frac{\sigma^2}{n})$ , tedy  $U = \frac{M - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$ .

(Pivotová statistika U slouží k řešení úloh o  $\mu$ , když  $\sigma^2$  známe.)

b)  $K = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ .

(Pivotová statistika K slouží k řešení úloh o  $\sigma^2$ , když  $\mu$  neznáme.)

c)  $\sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi^2(n)$ .

(Tato pivotová statistika slouží k řešení úloh o  $\sigma^2$ , když  $\mu$  známe.)

d)  $T = \frac{M - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$ .

(Pivotová statistika T slouží k řešení úloh o  $\mu$ , když  $\sigma^2$  neznáme.)

## Vysvětlení

ad a) Výběrový průměr  $M$  je lineární kombinace náhodných veličin s normálním rozložením, má tedy normální rozložení s parametry  $E(M) = \mu$ ,  $D(M) = \sigma^2/n$ . Statistika  $U$  se získá standardizací  $M$ .

ad b) Vhodnou úpravou výběrového rozptylu  $S^2$ , kde použijeme obrat  $X_i - M = (X_i - \mu) - (M - \mu)$ , lze statistiku  $K$  vyjádřit jako součet kvadrátů  $n - 1$  stochasticky nezávislých náhodných veličin se standardizovaným normálním rozložením. Tento součet se řídí rozložením  $\chi^2(n-1)$ .

ad c) Tato statistika je součet kvadrátů  $n$  stochasticky nezávislých náhodných veličin se standardizovaným normálním rozložením, řídí se tedy rozložením  $\chi^2(n)$ .

ad d)  $U \sim N(0, 1)$ ,  $K \sim \chi^2(n-1)$  jsou stochasticky nezávislé, protože  $M$  a  $S^2$  jsou stochasticky nezávislé, tudíž statistika

$$T = \frac{U}{\sqrt{\frac{K}{n-1}}} \sim t(n-1).$$

**Příklad:** Hmotnost balíčku krystalového cukru baleného na automatické lince se řídí normálním rozložením se střední hodnotou 1002 g a směrodatnou odchylkou 8 g. Kontrolor náhodně vybírá 9 balíčků z jedné série a zjišťuje, zda jejich průměrná hmotnost je alespoň 999 g. Pokud ne, podnik musí zaplatit pokutu 20 000 Kč. Jaká je pravděpodobnost, že podnik bude muset zaplatit pokutu?

**Řešení:**

$$X \sim N(1002, 64), M \sim N\left(1002, \frac{64}{9}\right)$$

$$P(M \leq 999) = P\left(\frac{M - 1002}{\sqrt{\frac{64}{9}}} \leq \frac{999 - 1002}{\sqrt{\frac{64}{9}}}\right) = P\left(Z \leq \frac{-3}{\frac{8}{3}}\right) = P(Z \leq -1,125) = \Phi(-1,125) = 1 - \Phi(1,125) = 1 - 0,86705 = 0,13295 = 12,9\%$$

Pravděpodobnost, že podnik bude platit pokutu, je asi 12,9%.

**Řešení pomocí systému STATISTICA:**

Využijeme toho, že STATISTICA pomocí funkce INormal(x;mu;sigma) umí vypočítat hodnotu distribuční funkce normálního rozložení se střední hodnotou mu a směrodatnou odchylkou sigma. Tedy  $P(M < 999) = \Phi\left(\frac{999 - 1002}{\sqrt{\frac{64}{9}}}\right)$ , kde  $\Phi$  je distribuční funkce rozložení  $N(1002, 64/9)$ .

Otevřeme nový datový soubor o jedné proměnné a jednom případě. Dvakrát klikneme na název proměnné Prom1. Do Dlouhého jména této proměnné napíšeme = INormal(999;1002;8/3).

V proměnné Prom1 se objeví hodnota 0,130295.

## Vzorce pro meze 100(1- $\alpha$ )% empirických intervalů spolehlivosti pro $\mu$ a $\sigma^2$

a) Interval spolehlivosti pro  $\mu$ , když  $\sigma^2$  známe (využití pivotové statistiky U)

$$\text{Oboustranný: } (d, h) = \left( m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2}, m + \frac{\sigma}{\sqrt{n}} u_{1-\alpha/2} \right)$$

$$\text{Levostranný: } (d, \infty) = \left( m - \frac{\sigma}{\sqrt{n}} u_{1-\alpha}, \infty \right)$$

$$\text{Pravostranný: } (-\infty, h) = \left( -\infty, m + \frac{\sigma}{\sqrt{n}} u_{1-\alpha} \right)$$

b) Interval spolehlivosti pro  $\mu$ , když  $\sigma^2$  neznáme (využití pivotové statistiky T)

$$\text{Oboustranný: } (d, h) = \left( m - \frac{S}{\sqrt{n}} t_{1-\alpha/2}(n-1), m + \frac{S}{\sqrt{n}} t_{1-\alpha/2}(n-1) \right)$$

$$\text{Levostranný: } (d, \infty) = \left( m - \frac{S}{\sqrt{n}} t_{1-\alpha}(n-1), \infty \right)$$

$$\text{Pravostranný: } (-\infty, h) = \left( -\infty, m + \frac{S}{\sqrt{n}} t_{1-\alpha}(n-1) \right)$$

c) Interval spolehlivosti pro  $\sigma^2$ , když  $\mu$  neznáme (využití pivotové statistiky K)

$$\text{Oboustranný: } (d, h) = \left( \frac{(n-1)s^2}{\chi_{1-\alpha/2}^2(n-1)}, \frac{(n-1)s^2}{\chi_{\alpha/2}^2(n-1)} \right)$$

$$\text{Levostranný: } (d, \infty) = \left( \frac{(n-1)s^2}{\chi_{1-\alpha}^2(n-1)}, \infty \right)$$

$$\text{Pravostranný: } (-\infty, h) = \left( -\infty, \frac{(n-1)s^2}{\chi_{\alpha}^2(n-1)} \right)$$

d) Interval spolehlivosti pro  $\sigma^2$ , když  $\mu$  známe (využití pivotové statistiky  $\sum_{i=1}^n \frac{(x_i - \mu)^2}{\sigma^2}$ )

$$\text{Oboustranný: } (d, h) = \left( \frac{\sum_{i=1}^n (x_i - \mu)^2}{\chi_{1-\alpha/2}^2(n)}, \frac{\sum_{i=1}^n (x_i - \mu)^2}{\chi_{\alpha/2}^2(n)} \right)$$

$$\text{Levostranný: } (d, \infty) = \left( \frac{\sum_{i=1}^n (x_i - \mu)^2}{\chi_{1-\alpha}^2(n)}, \infty \right)$$

$$\text{Pravostranný: } (-\infty, h) = \left( -\infty, \frac{\sum_{i=1}^n (x_i - \mu)^2}{\chi_{\alpha}^2(n)} \right)$$

**Příklad:** 10 krát nezávisle na sobě byla změřena jistá konstanta  $\mu$ . Výsledky měření byly: 2,1, 1,8, 2,1, 2,4, 1,9, 2,1, 2,1, 1,8, 2,3, 2,2. Tyto výsledky považujeme za číselné realizace náhodného výběru  $X_1, \dots, X_{10}$  z rozložení  $N(\mu, \sigma^2)$ , kde parametry  $\mu, \sigma^2$  neznáme. Najděte 95% empirický interval spolehlivosti jak pro  $\mu$ , tak pro  $\sigma^2$  a to

- a) oboustranný,
- b) levostranný,
- c) pravostranný.

**Řešení:**  $m = 2,06, s^2 = 0,0404, s = 0,2011, \alpha = 0,05, t_{0,975}(9) = 2,2622, t_{0,95}(9) = 1,8331, \chi^2_{0,975}(9) = 19,023, \chi^2_{0,025}(9) = 2,7, \chi^2_{0,95}(9) = 16,919, \chi^2_{0,05}(9) = 3,325$

ad a) **Oboustranný interval spolehlivosti pro střední hodnotu  $\mu$**

$$d = m - \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1) = 2,06 - \frac{0,2011}{\sqrt{10}} 2,2622 = 1,92$$

$$h = m + \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1) = 2,06 + \frac{0,2011}{\sqrt{10}} 2,2622 = 2,20$$

$1,92 < \mu < 2,20$  s pravděpodobností aspoň 0,95.

**Oboustranný interval spolehlivosti pro rozptyl  $\sigma^2$**

$$d = \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)} = \frac{9 \cdot 0,0404}{19,023} = 0,191$$

$$h = \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)} = \frac{9 \cdot 0,0404}{2,7} = 0,1347$$

$0,0191 < \sigma^2 < 0,1347$  s pravděpodobností aspoň 0,95.



ad b) **Levostranný interval spolehlivosti pro střední hodnotu  $\mu$**

$$d = m - \frac{s}{\sqrt{n}} t_{1-\alpha}(n-1) = 2,06 - \frac{0,201}{\sqrt{10}} 1,8331 = 1,94$$

$1,94 < \mu$  s pravděpodobností aspoň 0,95.

**Levostranný interval spolehlivosti pro rozptyl  $\sigma^2$**

$$d = \frac{n-1}{c} \frac{s^2}{\chi^2_{1-\alpha}(n-1)} = \frac{9}{1,9} \frac{0,0404}{21} = 0,21$$

$\sigma^2 > 0,0215$  s pravděpodobností aspoň 0,95.

ad c) **Pravostranný interval spolehlivosti pro střední hodnotu  $\mu$**

$$h = m + \frac{s}{\sqrt{n}} t_{1-\alpha}(n-1) = 2,06 + \frac{0,201}{\sqrt{10}} 1,8331 = 2,18$$

$\mu < 2,18$  s pravděpodobností aspoň 0,95.

**Pravostranný interval spolehlivosti pro rozptyl  $\sigma^2$**

$$h = \frac{n-1}{c} \frac{s^2}{\chi^2_{\alpha}(n-1)} = \frac{9}{1,9} \frac{0,0404}{9} = 0,09$$

$\sigma^2 < 0,1094$  s pravděpodobností aspoň 0,95.

## Řešení pomocí systému STATISTICA:

Vytvoříme nový datový soubor o jedné proměnné X a 10 případech. Do proměnné X napíšeme dané hodnoty.

Statistika – Základní statistiky a tabulky – Popisné statistiky – OK – Proměnné X – OK – Detailní výsledky – zaškrtneme

Meze spolehl. prům. a Meze sp. směr. odch. (ostatní volby zrušíme) – pro oboustranný 95% interval spolehlivosti

ponecháme implicitní hodnotu pro Interval 95,00, pro jednostranné intervaly změňme hodnotu na 90,00.

Výsledky pro oboustranné 95% intervaly spolehlivosti pro střední hodnotu  $\mu$ , pro směrodatnou odchylku  $\sigma$  a rozptyl  $\sigma^2$ :

Proměnná	Int. spolehl.	Int. spolehl.	Spolehliv. Sm. Odch.	Spolehliv. Sm. Odch.	NPror	NPror
	-95,00%	95,00%	-95,00%	+95,00%	= $v^2$	= $v^2$
X	1,916	2,203	0,138	0,367	0,019	0,134

Vidíme, že

$1,92 < \mu < 2,20$  s pravděpodobností aspoň 0,95,

$0,1383 < \sigma < 0,3671$  s pravděpodobností aspoň 0,95.

$0,0191 < \sigma^2 < 0,1348$  s pravděpodobností aspoň 0,95.

Výsledky pro jednostranné 95% intervaly spolehlivosti pro střední hodnotu  $\mu$ , pro směrodatnou odchylku  $\sigma$  a rozptyl  $\sigma^2$ :

	Int. spolehliv. -90,000	Int. spolehliv. 90,000	Spolehliv. Sm. Odc. -90,000	Spolehliv. Sm. Odc. +90,000	NPron. = $v3^2$	NPron. = $v4^2$
Proměření X	1,943	2,176	0,146	0,330	0,021	0,109

Vidíme, že

$\mu > 1,94$  s pravděpodobností aspoň 0,95,

$\mu < 2,20$  s pravděpodobností aspoň 0,95,

$\sigma > 0,1467$  s pravděpodobností aspoň 0,95,

$\sigma < 0,3309$  s pravděpodobností aspoň 0,95,

$\sigma^2 > 0,0215$  s pravděpodobností aspoň 0,95,

$\sigma^2 < 0,1095$  s pravděpodobností aspoň 0,95,

## Jednotlivé typy testů pro parametry normálního rozložení

a) Necht'  $X_1, \dots, X_n$  je náhodný výběr  $N(\mu, \sigma^2)$ , kde  $\sigma^2$  známe. Necht'  $n \geq 2$  a  $c$  je konstanta.

Test  $H_0: \mu = c$  proti  $H_1: \mu \neq c$  se nazývá **jednovýběrový z-test**.

b) Necht'  $X_1, \dots, X_n$  je náhodný výběr  $N(\mu, \sigma^2)$ , kde  $\sigma^2$  neznáme. Necht'  $n \geq 2$  a  $c$  je konstanta.

Test  $H_0: \mu = c$  proti  $H_1: \mu \neq c$  se nazývá **jednovýběrový t-test**.

c) Necht'  $X_1, \dots, X_n$  je náhodný výběr  $N(\mu, \sigma^2)$ , kde  $\mu$  neznáme. Necht'  $n \geq 2$  a  $c$  je konstanta.

Test  $H_0: \sigma^2 = c$  proti  $H_1: \sigma^2 \neq c$  se nazývá **test o rozptylu**.

## Provedení testů o parametrech $\mu, \sigma^2$ pomocí kritického oboru

### a) Provedení jednovýběrového z-testu

Vypočteme realizaci testového kritéria  $t_0 = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$ . Stanovíme kritický obor  $W$ . Pokud  $t_0 \in W$ ,  $H_0$  zamítáme na hladině významnosti  $\alpha$  a přijímáme  $H_1$ .

**Oboustranný test:** Testujeme  $H_0: \mu = c$  proti  $H_1: \mu \neq c$ . Kritický obor má tvar:  $W = (-\infty, -u_{1-\alpha/2}] \cup [u_{1-\alpha/2}, \infty)$ .

**Levostranný test:** Testujeme  $H_0: \mu = c$  proti  $H_1: \mu < c$ . Kritický obor má tvar:  $W = (-\infty, -u_{1-\alpha})$ .

**Pravostranný test:** Testujeme  $H_0: \mu = c$  proti  $H_1: \mu > c$ . Kritický obor má tvar:  $W = [u_{1-\alpha}, \infty)$ .

### b) Provedení jednovýběrového t-testu

Vypočteme realizaci testového kritéria  $t_0 = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$ . Stanovíme kritický obor  $W$ . Pokud  $t_0 \in W$ ,  $H_0$  zamítáme na hladině významnosti  $\alpha$  a přijímáme  $H_1$ .

**Oboustranný test:** Testujeme  $H_0: \mu = c$  proti  $H_1: \mu \neq c$ . Kritický obor má tvar:  $W = (-\infty, -t_{1-\alpha/2, n-1}] \cup [t_{1-\alpha/2, n-1}, \infty)$ .

**Levostranný test:** Testujeme  $H_0: \mu = c$  proti  $H_1: \mu < c$ . Kritický obor má tvar:  $W = (-\infty, -t_{1-\alpha, n-1})$ .

**Pravostranný test:** Testujeme  $H_0: \mu = c$  proti  $H_1: \mu > c$ . Kritický obor má tvar:  $W = [t_{1-\alpha, n-1}, \infty)$ .

### c) Provedení testu o rozptylu

Vypočteme realizaci testového kritéria  $t_0 = \frac{s^2}{c} \cdot (n-1)$ . Stanovíme kritický obor  $W$ . Pokud  $t_0 \in W$ ,  $H_0$  zamítáme na hladině významnosti  $\alpha$  a přijímáme  $H_1$ .

**Oboustranný test:** Testujeme  $H_0: \sigma^2 = c$  proti  $H_1: \sigma^2 \neq c$ . Kritický obor má tvar:

$$W = \left(0, \chi_{\alpha/2}^2(n-1)\right) \cup \left(\chi_{1-\alpha/2}^2(n-1), \infty\right)$$

**Levostranný test:** Testujeme  $H_0: \sigma^2 = c$  proti  $H_1: \sigma^2 < c$ . Kritický obor má tvar:

$$W = \left(0, \chi_{1-\alpha}^2(n-1)\right)$$

**Pravostranný test:** Testujeme  $H_0: \sigma^2 = c$  proti  $H_1: \sigma^2 > c$ . Kritický obor má tvar:

$$W = \left(\chi_{1-\alpha}^2(n-1), \infty\right)$$

**Příklad:** Podle údajů na obalu čokolády by její čistá hmotnost měla být 125 g. Výrobce dostal několik stížností od kupujících, ve kterých tvrdili, že hmotnost čokolád je nižší než deklarovaných 125 g. Z tohoto důvodu oddělení kontroly náhodně vybralo 50 čokolád a zjistilo, že jejich průměrná hmotnost je 122 g a směrodatná odchylka 8,6 g. Za předpokladu, že hmotnost čokolád se řídí normálním rozložením, můžeme na hladině významnosti 0,01 považovat stížnosti kupujících za oprávněné?

**Řešení:**  $X_1, \dots, X_{50}$  je náhodný výběr z  $N(\mu, \sigma^2)$ . Testujeme hypotézu  $H_0: \mu = 125$  proti levostranné alternativě  $H_1: \mu < 125$ . Protože neznáme rozptyl  $\sigma^2$ , použijeme jednovýběrový t-test.

Testové kritérium 
$$\frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{122 - 125}{\frac{8,6}{\sqrt{50}}} = -2,66$$

Kritický obor  $W = \left\{ \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \leq -t_{n-1; 1-\alpha} \right\} = \left\{ \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \leq -t_{49; 0,99} \right\} = \left\{ \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \leq -2,404 \right\}$

Jelikož testové kritérium se realizuje v kritickém oboru, zamítáme nulovou hypotézu na hladině významnosti 0,01. Stížnosti kupujících tedy lze považovat za oprávněné.

### Výpočet pomocí systému STATISTICA:

Statistiky – Základní statistiky a tabulky – Testy rozdílů: r, %, průměry – OK – vybereme Rozdíl mezi dvěma průměry (normální rozdělení) – zaškrtneme Výběrový průměr vs. Střední hodnota a zvolíme jednostr. – do políčka Pr1 napíšeme 122, do políčka SmOd1 napíšeme 8,6, do políčka N1 napíšeme 50, do políčka Pr2 napíšeme 125 - Výpočet. Dostaneme p-hodnotu 0,0086, tedy zamítáme nulovou hypotézu na hladině významnosti 0,01

## Náhodný výběr z dvourozměrného rozložení

Nechť  $\left(\begin{smallmatrix} X_1 \\ Y_1 \end{smallmatrix}\right), \dots, \left(\begin{smallmatrix} X_n \\ Y_n \end{smallmatrix}\right)$  je náhodný výběr z dvourozměrného rozložení, přičemž  $n \geq 2$ . Označíme  $\mu = \mu_1 - \mu_2$  a zavedeme

**rozdílový náhodný výběr**  $Z_1 = X_1 - Y_1, \dots, Z_n = X_n - Y_n$ , o němž předpokládáme, že se řídí normálním rozložením.

Vypočteme  $M = \frac{1}{n} \sum_{i=1}^n Z_i$ ,  $S^2 = \frac{1}{n} \sum_{i=1}^n Z_i^2 - M^2$ .

## Vzorec pro meze 100(1- $\alpha$ )% empirického intervalu spolehlivosti pro střední hodnotu rozdílového náhodného výběru

Oboustranný:  $(d, h) = \left(m - \frac{S}{\sqrt{n}} t_{1-\alpha/2}(n-1), m + \frac{S}{\sqrt{n}} t_{1-\alpha/2}(n-1)\right)$

Levostranný:  $(d, \infty) = \left(m - \frac{S}{\sqrt{n}} t_{1-\alpha}(n-1), \infty\right)$

Pravostranný:  $(-\infty, h) = \left(-\infty, m + \frac{S}{\sqrt{n}} t_{1-\alpha}(n-1)\right)$



**Příklad:** Dvěma rozdílnými laboratorními metodami se zjišťoval obsah chemické látky v roztoku (v procentech). Bylo vybráno 5 vzorků a proměřeno oběma metodami. Výsledky měření jsou obsaženy v tabulce:

číslo vzorku	1	2	3	4	5
1. metoda	2,3	1,9	2,1	2,4	2,6
2. metoda	2,4	2,0	2,0	2,3	2,5

Za předpokladu, že data mají normální rozložení, sestrojte 90% empirický interval spolehlivosti pro rozdíl středních hodnot výsledků obou metod.

**Řešení:**

Přejdeme k rozdílovému náhodnému výběru, jehož realizace jsou: -0,1 -0,1 0,1 0,1 0,1. Vypočteme  $m = 0,02$ ,  $s^2 = 0,012$ ,  $s = 0,109545$ . Předpokládáme, že tato data pocházejí z normálního rozložení  $N(\mu, \sigma^2)$ . Vypočteme meze 90% oboustranného intervalu spolehlivosti pro  $\mu$  při neznámém  $\sigma$ :

$$d = \bar{x} - t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} = 0,2 - \frac{0,109545}{\sqrt{5}} \cdot 2,1318 = 0,0844$$

$$h = \bar{x} + t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} = 0,2 + \frac{0,109545}{\sqrt{5}} \cdot 2,1318 = 0,24$$

$-0,0844 < \mu < 0,1244$  s pravděpodobností aspoň 0,9.

### Výpočet pomocí systému STATISTICA:

Vytvoříme nový datový soubor o 3 proměnných a 5 případech. Do 1. proměnné X napíšeme hodnoty pro 1. metodu, do 2. proměnné Y hodnoty pro 2. metodu a do 3. proměnné Z rozdíly mezi X a Y.

Statistiky – Základní statistiky a tabulky – Popisné statistiky, OK - Proměnné Z, Detailní výsledky – zaškrtneme Meze spolehl. Prům. – Interval 90% - Výpočet. Dostaneme tabulku:

	Popisné statistiky (ch)	
	Int. spoleh.	Int. spoleh.
Promě.	-90,000	90,000
Z	-0,084	0,124

Vidíme tedy, že  $-0,0844 < \mu < 0,1244$  s pravděpodobností aspoň 0,9.

### Párový t-test

Nechť  $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \dots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$  je náhodný výběr z dvourozměrného rozložení, přičemž  $n > 1$ . Označíme  $\mu_1$  a zavedeme

rozdílový náhodný výběr  $Z_1, \dots, Z_n$ , jehož výběrový průměr je  $M = \frac{1}{n} \sum_{i=1}^n Z_i$  a výběrový rozptyl je

$S^2 = \frac{1}{n-1} \sum_{i=1}^n (Z_i - M)^2$ . Předpokládáme, že tento náhodný výběr pochází z normálního rozložení. Test hypotézy o rozdílu středních hodnot  $\mu_1$  se nazývá **párový t-test** a provádí se stejně jako jednovýběrový t-test aplikovaný na rozdílový náhodný výběr  $Z_1, \dots, Z_n$ .

### Provedení párového t-testu

Vypočteme realizaci testového kritéria  $t_0 = \frac{\bar{z} - c}{\frac{s}{\sqrt{n}}}$ . Stanovíme kritický obor  $W$ . Pokud  $t_0 \in W$ ,  $H_0$  zamítáme na hladině významnosti  $\alpha$  a přijímáme  $H_1$ .

**Oboustranný test:** Testujeme  $H_0: \mu = c$  proti  $H_1: \mu \neq c$ . Kritický obor má tvar:  $W = (-\infty, -t_{1-\alpha/2, n-1}) \cup (t_{1-\alpha/2, n-1}, \infty)$ .

**Levostranný test:** Testujeme  $H_0: \mu = c$  proti  $H_1: \mu < c$ . Kritický obor má tvar:  $W = (-\infty, -t_{1-\alpha, n-1})$ .

**Pravostranný test:** Testujeme  $H_0: \mu = c$  proti  $H_1: \mu > c$ . Kritický obor má tvar:  $W = (t_{1-\alpha, n-1}, \infty)$ .

**Příklad:** V následující tabulce jsou údaje o výnosnosti dosažené 12 náhodně vybranými firmami při investování do mezinárodního podnikání (veličina X) a do domácího podnikání (veličina Y):

č.firmy	1	2	3	4	5	6	7	8	9	10	11	12
X	10	12	14	12	12	17	9	15	9	11	7	15
Y	11	14	15	11	13	16	10	13	11	17	9	19

(Výnosnost je vyjádřena v procentech a představuje podíl na zisku vložených investic za rok.)

Za předpokladu, že data pocházejí z dvourozměrného rozložení a jejich rozdíl se řídí normálním rozložením, na hladině významnosti 0,1 testujte hypotézu, že neexistuje rozdíl mezi střední hodnotou výnosnosti investic do mezinárodního a domácího podnikání proti oboustranné alternativě.

Testování proved'te

a) pomocí intervalu spolehlivosti, b) pomocí kritického oboru.

(Pro úsporu času známe realizace výběrového průměru  $m = 0$  a výběrového rozptylu  $s^2 = 4,78$  rozdílového náhodného výběru  $Z_i = X_i - Y_i$ ,  $i = 1, \dots, 12$ .)

**Řešení:**

Testujeme  $H_0: \mu = 0$  proti  $H_1: \mu \neq 0$

ad a) 90% interval spolehlivosti pro střední hodnotu  $\mu$  při neznámém rozptylu  $\sigma^2$  má meze:

$$d = -\frac{s}{\sqrt{n}} t_{0,95}(n-1) = -\frac{\sqrt{4,78}}{\sqrt{12}} \cdot 1,795 = -1,67$$

$$h = +\frac{s}{\sqrt{n}} t_{0,95}(n-1) = +\frac{\sqrt{4,78}}{\sqrt{12}} \cdot 1,795 = 1,67$$

Protože číslo  $c = 0$  neleží v intervalu  $(-2,4677; -0,1989)$ ,  $H_0$  zamítáme na hladině významnosti 0,1.

ad b) Vypočítáme realizaci testové statistiky  $t_0 = \frac{m - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{0 - 0}{\frac{\sqrt{4,78}}{\sqrt{12}}} = 0$

Stanovíme kritický obor  $W = \{t_0 \mid t_0 < -1,7959 \text{ nebo } t_0 > 1,7959\}$

Protože testová statistika se realizuje v kritickém oboru,  $H_0$  zamítáme na hladině významnosti 0,1.

### Výpočet pomocí systému STATISTICA:

Vytvoříme nový datový soubor o 2 proměnných a 12 případech. Do 1. proměnné X napíšeme hodnoty pro mezinárodní podnikání, do 2. proměnné hodnoty pro domácí podnikání.

Statistiky – Základní statistiky a tabulky – t-test pro závislé vzorky, OK - Proměnné X, Y – OK – Výpočet. Dostaneme tabulku:

Promě	t-test pro závislé vzorky (investování)							
	Prům	Sm.od	N	Rozd	Sm.od rozdíl	t	s	p
X	11,91	2,937						
Y	13,25	3,048	12	-1,33	2,188	-2,11	1	0,058

Vypočtenou p-hodnotu 0,05849 porovnáme se zvolenou hladinou významnosti  $\alpha = 0,1$ . Protože  $p \leq \alpha$ , zamítáme nulovou hypotézu na hladině významnosti 0,1.

## Parametrické úlohy o dvou nezávislých náhodných výběrech z normálních rozložení

**Motivace:** Máme-li k dispozici dva nezávislé náhodné výběry z normálních rozložení, je naším úkolem porovnat střední hodnoty či rozptyly těchto rozložení. Zpravidla konstruujeme intervaly spolehlivosti pro rozdíl středních hodnot respektive hodnotíme shodu středních hodnot pomocí dvouvýběrového t-testu či dvouvýběrového z-testu a shodu rozptylů pomocí F-testu.

### Osnova:

- rozložení statistik odvozených ze dvou výběrových průměrů a rozptylů
- vzorce pro meze intervalů spolehlivosti pro rozdíl středních hodnot a podíl rozptylů
- jednotlivé typy testů pro parametry dvou normálních rozložení  
(dvouvýběrový z-test, dvouvýběrový t-test, F-test)
- Cohenův koeficient věcného účinku

## Rozložení statistik odvozených z výběrových průměrů a výběrových rozptylů normálních rozložení

Předpokládáme, že

$X_{11}, \dots, X_{1n_1}$  je náhodný výběr z rozložení  $N(\mu_1, \sigma_1^2)$ ,

$X_{21}, \dots, X_{2n_2}$  je náhodný výběr z rozložení  $N(\mu_2, \sigma_2^2)$ ,

přičemž  $n_1 \geq 2$  a  $n_2 \geq 2$  a oba výběry jsou stochasticky nezávislé.

Označme

$M_1, M_2$  výběrové průměry,

$S_1^2, S_2^2$  výběrové rozptyly a

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$
 vážený průměr výběrových rozptylů.

Pak platí:

a) Statistiky  $M_1 - M_2$  a  $S^2$  jsou stochasticky nezávislé.

$$b) U = \frac{M_1 - M_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1).$$

(Pivotová statistika U slouží k řešení úloh o  $\mu_1 - \mu_2$ , když  $\sigma_1^2$  a  $\sigma_2^2$  známe.)

$$c) \text{ Jestliže } \sigma_1^2 = \sigma_2^2 =: \sigma^2, \text{ pak } K = \frac{(n_1 + n_2 - 2) S^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2).$$

(Pivotová statistika K slouží k řešení úloh o neznámém společném rozptylu  $\sigma^2$ .)

$$d) \text{ Jestliže } \sigma_1^2 = \sigma_2^2 =: \sigma^2, \text{ pak } T = \frac{M_1 - M_2 - (\mu_1 - \mu_2)}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2).$$

(Pivotová statistika T slouží k řešení úloh o  $\mu_1 - \mu_2$ , když  $\sigma_1^2$  a  $\sigma_2^2$  neznáme, ale víme, že jsou shodné.)

$$e) F = \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1).$$

(Pivotová statistika F slouží k řešení úloh o  $\sigma_1^2/\sigma_2^2$ .)



### Vysvětlení:

ad a) Neuvádíme, viz např. J. Anděl: Matematická statistika.

ad b)  $M_1 - M_2$  je lineární kombinace náhodných veličin s normálním rozložením, má tedy normální rozložení s parametry

$$E(M_1 - M_2) = \mu_1 - \mu_2,$$

$$D(M_1 - M_2) = \sigma_1^2/n_1 + \sigma_2^2/n_2.$$

U se získá standardizací  $M_1 - M_2$ .

ad c)  $K_1 = \frac{(n_1 - 1)S_1^2}{\sigma^2} \sim \chi^2(n_1 - 1)$  a  $K_2 = \frac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2(n_2 - 1)$  jsou stochasticky nezávislé náhodné veličiny, tedy

$$K = K_1 + K_2 \sim \chi^2(n_1 + n_2 - 2).$$

ad d)  $U = \frac{M_1 - M_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma^2/n_1 + \sigma^2/n_2}} \sim N(0, 1)$ ,  $K = \frac{(n_1 + n_2 - 2)S^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$  jsou stochasticky nezávislé, protože

$M_1 - M_2$  a  $S^2$  jsou stochasticky nezávislé.  $T = \frac{U}{\frac{S}{\sqrt{n_1 + n_2 - 2}}} = \frac{M_1 - M_2 - (\mu_1 - \mu_2)}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$ .

ad e)  $K_1 = \frac{(n_1 - 1)S_1^2}{\sigma^2} \sim \chi^2(n_1 - 1)$  a  $K_2 = \frac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2(n_2 - 1)$  jsou stochasticky nezávislé náhodné veličiny, tedy

$$F = \frac{K_1/n_1}{K_2/n_2} = \frac{S_1^2/S_2^2}{\sigma^2/\sigma^2} \sim F(n_1 - 1, n_2 - 1).$$

**Příklad:** Necht' jsou dány dva nezávislé náhodné výběry, první pochází z rozložení  $N(0,28; 0,09)$  a má rozsah 16, druhý pochází z rozložení  $N(0,25; 0,04)$  a má rozsah 25. Jaká je pravděpodobnost, že výběrový průměr 1. výběru bude větší než výběrový průměr 2. výběru?

**Řešení:**

$$P(M_1 > M_2) = P\left(\frac{M_1 - M_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq \frac{0 - 0,03}{\sqrt{\frac{0,09}{16} + \frac{0,04}{25}}}\right) = P(U \leq -0,529) = 0,2995 = 29,95\%$$

S pravděpodobností přibližně 63,7% je výběrový průměr 1. výběru větší než výběrový průměr 2. výběru.

**Výpočet pomocí systému STATISTICA:**

Statistika  $M_1 - M_2$  se podle bodu (a) řídí rozložením  $N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$ ,

kde  $\mu_1 - \mu_2 = 0,28 - 0,25 = 0,03$ ,  $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{0,09}{16} + \frac{0,04}{25} = 0,0072$ , tj. statistika  $M_1 - M_2 \sim N(0,03; 0,007225)$ .

Otevřeme nový datový soubor o jedné proměnné a jednom případě. Do Dlouhého jména této proměnné napíšeme

`= 1-INormal(0;0,03;sqrt(0,007225))`. V proměnné Prom1 se objeví hodnota 0,637934:

	1
Prom1	0,637

## Intervaly spolehlivosti pro parametrické funkce $\mu_1 - \mu_2, \sigma_1^2/\sigma_2^2$

Uvedeme přehled vzorců pro meze  $100(1-\alpha)\%$  empirických intervalů spolehlivosti pro parametrické funkce  $\mu_1 - \mu_2, \sigma_1^2/\sigma_2^2$ .

a) Interval spolehlivosti pro  $\mu_1 - \mu_2$ , když  $\sigma_1^2, \sigma_2^2$  známe (využití pivotové statistiky U)

$$\text{Oboustranný: } (d, h) = (m_1 - m_2 - \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} u_{1-\alpha/2}, m_1 - m_2 + \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} u_{1-\alpha/2})$$

$$\text{Levostranný: } (d, \infty) = (m_1 - m_2 - \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} u_{1-\alpha}, \infty)$$

$$\text{Pravostranný: } (-\infty, h) = (-\infty, m_1 - m_2 + \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} u_{1-\alpha})$$

b) Interval spolehlivosti pro  $\mu_1 - \mu_2$ , když  $\sigma_1^2, \sigma_2^2$  neznáme, ale víme, že jsou shodné (využití pivotové statistiky T)

Oboustranný:

$$(d, h) = (m_1 - m_2 - S^* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{1-\alpha/2}(n_1+n_2-2), m_1 - m_2 + S^* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{1-\alpha/2}(n_1+n_2-2))$$

$$\text{Levostranný: } (d, \infty) = (m_1 - m_2 - S^* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{1-\alpha}(n_1+n_2-2), \infty)$$

$$\text{Pravostranný: } (-\infty, h) = (-\infty, m_1 - m_2 + S^* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{1-\alpha}(n_1+n_2-2))$$

c) Interval spolehlivosti pro společný neznámý rozptyl  $\sigma^2$  (využití pivotové statistiky K)

$$\text{Oboustranný: } (d, h) = \left( \frac{(n_1 + n_2 - 2)s_*^2}{\chi_{1-\alpha/2}^2(n_1 + n_2 - 2)}, \frac{(n_1 + n_2 - 2)s_*^2}{\chi_{\alpha/2}^2(n_1 + n_2 - 2)} \right)$$

$$\text{Levostranný: } (d, \infty) = \left( \frac{(n_1 + n_2 - 2)s_*^2}{\chi_{1-\alpha}^2(n_1 + n_2 - 2)}, \infty \right)$$

$$\text{Pravostranný: } (-\infty, h) = \left( -\infty, \frac{(n_1 + n_2 - 2)s_*^2}{\chi_{\alpha}^2(n_1 + n_2 - 2)} \right)$$

d) Interval spolehlivosti pro podíl rozptylů  $\frac{\sigma_1^2}{\sigma_2^2}$  (využití pivotové statistiky F)

$$\text{Oboustranný: } (d, h) = \left( \frac{s_1^2/s_2^2}{F_{\alpha/2}(n_1-1, n_2-1)}, \frac{s_1^2/s_2^2}{F_{\alpha/2}(n_1-1, n_2-1)} \right)$$

$$\text{Levostranný: } (d, \infty) = \left( \frac{s_1^2/s_2^2}{F_{\alpha}(n_1-1, n_2-1)}, \infty \right)$$

$$\text{Pravostranný: } (-\infty, h) = \left( -\infty, \frac{s_1^2/s_2^2}{F_{\alpha}(n_1-1, n_2-1)} \right)$$

**Upozornění:** Nemá-li v bodě (b) splněn předpoklad o shodě rozptylů, lze sestavit aspoň přibližný  $100(1-\alpha)\%$  interval spolehlivosti pro  $\mu_1 - \mu_2$ .

V tomto případě má statistika T přibližně rozložení  $t(\nu)$ , kde počet stupňů volnosti  $\nu = \frac{s_1^2/n_1 + s_2^2/n_2}{s_1^2/n_1 + s_2^2/n_2}$ . Nemá-li v celé

číslo, použijeme v tabulkách kvantilů Studentova rozložení lineární interpolaci.

**Příklad:** Ve dvou nádržích se zkoumal obsah chlóru (v g/l). Z první nádrže bylo odebráno 25 vzorků, z druhé nádrže 10 vzorků. Byly vypočteny realizace výběrových průměrů a rozptylů:  $m_1 = 34,48$ ,  $m_2 = 35,59$ ,  $s_1^2 = 1,7482$ ,  $s_2^2 = 1,7121$ . Hodnoty zjištěné z odebraných vzorků považujeme za realizace dvou nezávislých náhodných výběrů z rozložení  $N(\mu_1, \sigma^2)$  a  $N(\mu_2, \sigma^2)$ . Sestrojte 95% empirický interval spolehlivosti pro rozdíl středních hodnot  $\mu_1 - \mu_2$ .

**Řešení:**

Úloha vede na vzorec (b) s využitím statistiky T. Vypočteme vážený průměr výběrových rozptylů a najdeme odpovídající kvantily Studentova rozložení:

$$S_*^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{24 \cdot 1,7482 + 9 \cdot 1,7121}{33} = 1,738, \quad t_{0,975}(33) = 2,035$$

Dosadíme do vzorců pro dolní a horní mez intervalu spolehlivosti:

$$d = m_1 - m_2 - S_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{1-\alpha/2}(n_1+n_2-2) =$$

$$= 34,48 - 35,59 - \sqrt{1,738} \cdot \sqrt{\frac{1}{25} + \frac{1}{10}} \cdot 2,035 = -2,114$$

$$h = m_1 - m_2 + S_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{1-\alpha/2}(n_1+n_2-2) =$$

$$= 34,48 - 35,59 + \sqrt{1,738} \cdot \sqrt{\frac{1}{25} + \frac{1}{10}} \cdot 2,035 = -0,106$$

$-2,114 \text{ g/l} < \mu_1 - \mu_2 < -0,106 \text{ g/l}$  s pravděpodobností aspoň 0,95.

## Výpočet pomocí systému STATISTICA:

Otevřeme nový datový soubor o dvou proměnných d a h a jednom případě.

Do Dlouhého jména proměnné d napíšeme

```
=34,48-35,59-sqrt((24*1,7482+9*1,7121)/33)*sqrt((1/25)+(1/10))*VStudent(0,975;33)
```

Do Dlouhého jména proměnné h napíšeme

```
=34,48-35,59+
```

```
sqrt((24*1,7482+9*1,7121)/33)*sqrt((1/25)+(1/10))*VStudent(0,975;33)
```

	d	h
1	-2,114	-0,106

S pravděpodobností aspoň 0,95 tedy  $-2,114 \text{ g/l} < \mu_1 - \mu_2 < -0,106 \text{ g/l}$ .

**Příklad:** V předešlém příkladě nyní předpokládáme, že dané dva náhodné výběry pocházejí z rozložení  $N(\mu_1, \sigma_1^2)$  a  $N(\mu_2, \sigma_2^2)$ . Sestrojte 95% empirický interval spolehlivosti pro podíl rozptylů.

**Řešení:**

Úloha vede na vzorec (d) s využitím statistiky F.

$$d = \frac{s_1^2/s_2^2}{F_{1-\alpha/2}(n_1-1, n_2-1)} = \frac{17482/17121}{F_{0,975}(24;9)} = \frac{17482/17121}{36142} = 0,28$$

$$h = \frac{s_1^2/s_2^2}{F_{\alpha/2}(n_1-1, n_2-1)} = \frac{17482/17121}{F_{0,025}(24;9)} = \frac{17482/17121}{127027} = 2,76$$

$0,28 < \sigma_1^2 / \sigma_2^2 < 2,76$  s pravděpodobností aspoň 0,95.

**Výpočet pomocí systému STATISTICA:**

Otevřeme nový datový soubor o dvou proměnných d a h a jednom případě.

Do Dlouhého jména proměnné d napíšeme

$$=(1,7482/1,7121)/VF(0,975;24;9)$$

(Funkce VF(x;ný;omega) počítá x-kvantil Fisherova – Snedecorova rozložení F(ný, omega).)

Do Dlouhého jména proměnné h napíšeme

$$=(1,7482/1,7121)/VF(0,025;24;9)$$

	1	2
	d	h
1	0,282	2,759

S pravděpodobností aspoň 0,95 tedy platí:  $0,28 < \sigma_1^2 / \sigma_2^2 < 2,76$ .

## Jednotlivé typy testů o parametrických funkcích $\mu_1 - \mu_2, \sigma_1^2/\sigma_2^2$

a) Necht'  $X_{11}, \dots, X_{1n_1}$  je náhodný výběr z rozložení  $N(\mu_1, \sigma_1^2)$  a  $X_{21}, \dots, X_{2n_2}$  je na něm nezávislý náhodný výběr z rozložení  $N(\mu_2, \sigma_2^2)$ , přičemž  $n_1 \geq 2, n_2 \geq 2$  a  $\sigma_1^2, \sigma_2^2$  známe. Necht'  $c$  je konstanta.

Test  $H_0: \mu_1 - \mu_2 = c$  proti  $H_1: \mu_1 - \mu_2 \neq c$  se nazývá **dvouvýběrový z-test**.

b) Necht'  $X_{11}, \dots, X_{1n_1}$  je náhodný výběr z rozložení  $N(\mu_1, \sigma^2)$  a  $X_{21}, \dots, X_{2n_2}$  je na něm nezávislý náhodný výběr z rozložení  $N(\mu_2, \sigma^2)$ , přičemž  $n_1 \geq 2$  a  $n_2 \geq 2$  a  $\sigma^2$  neznáme. Necht'  $c$  je konstanta.

Test  $H_0: \mu_1 - \mu_2 = c$  proti  $H_1: \mu_1 - \mu_2 \neq c$  se nazývá **dvouvýběrový t-test**.

c) Necht'  $X_{11}, \dots, X_{1n_1}$  je náhodný výběr z rozložení  $N(\mu_1, \sigma_1^2)$  a  $X_{21}, \dots, X_{2n_2}$  je na něm nezávislý náhodný výběr rozložení  $N(\mu_2, \sigma_2^2)$ , přičemž  $n_1 \geq 2$  a  $n_2 \geq 2$ . Test  $H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$  proti  $H_1: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$  se nazývá **F-test**.



## Provedení testů o parametrických funkcích $\mu_1 - \mu_2, \sigma_1^2/\sigma_2^2$ pomocí kritického oboru

### a) Provedení dvouvýběrového z-testu

Vypočteme realizaci  $t_0$  testového kritéria  $T_0 = \frac{\bar{M}_1 - \bar{M}_2 - c}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ . Stanovíme kritický obor  $W$ . Pokud  $t_0 \in W$ ,  $H_0$  zamítáme na

hladině významnosti  $\alpha$  a přijímáme  $H_1$ .

**Oboustranný test:** Testujeme  $H_0: \mu_1 - \mu_2 = c$  proti  $H_1: \mu_1 - \mu_2 \neq c$ . Kritický obor má tvar:  $W = (-\infty, -t_{1-\alpha/2}) \cup (t_{1-\alpha/2}, \infty)$ .

**Levostranný test:** Testujeme  $H_0: \mu_1 - \mu_2 = c$  proti  $H_1: \mu_1 - \mu_2 < c$ . Kritický obor má tvar:  $W = (-\infty, -t_{1-\alpha})$ .

**Pravostranný test:** Testujeme  $H_0: \mu_1 - \mu_2 = c$  proti  $H_1: \mu_1 - \mu_2 > c$ . Kritický obor má tvar:  $W = (t_{1-\alpha}, \infty)$ .

### b) Provedení dvouvýběrového t-testu

Vypočteme realizaci  $t_0$  testového kritéria  $T_0 = \frac{\bar{M}_1 - \bar{M}_2 - c}{S^* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ . Stanovíme kritický obor  $W$ . Pokud  $t_0 \in W$ ,  $H_0$  zamítáme na

hladině významnosti  $\alpha$  a přijímáme  $H_1$ .

**Oboustranný test:** Testujeme  $H_0: \mu_1 - \mu_2 = c$  proti  $H_1: \mu_1 - \mu_2 \neq c$ . Kritický obor má tvar:

$$W = (-\infty, -t_{1-\alpha/2} \sqrt{\frac{n_1 n_2}{n_1 + n_2 - 2}}) \cup (t_{1-\alpha/2} \sqrt{\frac{n_1 n_2}{n_1 + n_2 - 2}}, \infty)$$

**Levostranný test:** Testujeme  $H_0: \mu_1 - \mu_2 = c$  proti  $H_1: \mu_1 - \mu_2 < c$ . Kritický obor má tvar:  $W = (-\infty, -t_{1-\alpha} \sqrt{\frac{n_1 n_2}{n_1 + n_2 - 2}})$ .

**Pravostranný test:** Testujeme  $H_0: \mu_1 - \mu_2 = c$  proti  $H_1: \mu_1 - \mu_2 > c$ . Kritický obor má tvar:  $W = (t_{1-\alpha} \sqrt{\frac{n_1 n_2}{n_1 + n_2 - 2}}, \infty)$ .

### c) Provedení F-testu

Vypočteme realizaci testového kritéria  $t_0 = \frac{s^2}{\sigma_0^2}$ . Stanovíme kritický obor  $W$ . Pokud  $t_0 \in W$ ,  $H_0$  zamítáme na hladině významnosti  $\alpha$  a přijímáme  $H_1$ .

**Oboustranný test:** Testujeme  $H_0: \sigma^2 = 1$  proti  $H_1: \sigma^2 \neq 1$ . Kritický obor má tvar:

$$W = \left\{ F_{\alpha/2, n_1-1, n_2-1} < \frac{s^2}{\sigma_0^2} < F_{1-\alpha/2, n_1-1, n_2-1}, \infty \right\}$$

**Levostranný test:** Testujeme  $H_0: \sigma^2 = 1$  proti  $H_1: \sigma^2 < 1$ . Kritický obor má tvar:  $W = \left\{ F_{1-\alpha, n_1-1, n_2-1}, \infty \right\}$ .

**Pravostranný test:** Testujeme  $H_0: \sigma^2 = 1$  proti  $H_1: \sigma^2 > 1$ . Kritický obor má tvar:  $W = \left\{ F_{1-\alpha, n_1-1, n_2-1}, \infty \right\}$ .

**Příklad:** V restauraci "U bílého koníčka" měřili ve 20 případech čas obsluhy zákazníka. Výsledky v minutách: 6, 8, 11, 4, 7, 6, 10, 6, 9, 8, 5, 12, 13, 10, 9, 8, 7, 11, 10, 5. V restauraci "Zlatý lev" bylo dané pozorování uskutečněno v 15 případech s těmito výsledky: 9, 11, 10, 7, 6, 4, 8, 13, 5, 15, 8, 5, 6, 8, 7. Za předpokladu, že uvedené hodnoty pocházejí ze dvou normálních rozložení, na hladině významnosti 0,05 testujte hypotézu, že střední hodnoty doby obsluhy jsou v obou restauracích stejné.

**Řešení:**

Na hladině významnosti 0,05 testujeme nulovou hypotézu  $H_0: \mu_1 - \mu_2 = 0$  proti oboustranné alternativě  $H_1: \mu_1 - \mu_2 \neq 0$ . Je to úloha na dvouvýběrový t-test. Před provedením tohoto testu je však nutné pomocí F-testu ověřit shodu rozptylů. Na hladině významnosti 0,05 tedy testujeme  $H_0: \sigma^2 = 1$  proti  $H_1: \sigma^2 \neq 1$ . Nejprve vypočteme  $m_1 = 8,25$ ,  $m_2 = 8,13$ ,  $s_1^2 = 6,307$ ,  $s_2^2 = 9,41$ ,  $S^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{19 \cdot 6,307 + 14 \cdot 9,41}{33} = 7,623$ . Podle vzorce (c) vypočteme realizaci testové statistiky:

$$t_0 = \frac{m_1 - m_2}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{8,25 - 8,13}{7,623 \sqrt{\frac{1}{20} + \frac{1}{15}}} = 0,2$$

Stanovíme kritický obor:  $W = (F_{\alpha/2, n_1-1, n_2-1}, F_{1-\alpha/2, n_1-1, n_2-1}) = (F_{0,025, 19, 14}, F_{0,975, 19, 14}) = (0,7149, 2,64928607)$

Protože se testová statistika nerealizuje v kritickém oboru, nulovou hypotézu nezamítáme na hladině významnosti 0,05. Rozptyly tedy můžeme považovat za shodné. Nyní se vrátíme k dvouvýběrovému t-testu. Podle vzorce (b) vypočteme realizaci testové statistiky:

$$t_0 = \frac{m_1 - m_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{8,25 - 8,13}{7,623 \sqrt{\frac{1}{20} + \frac{1}{15}}} = 0,2$$

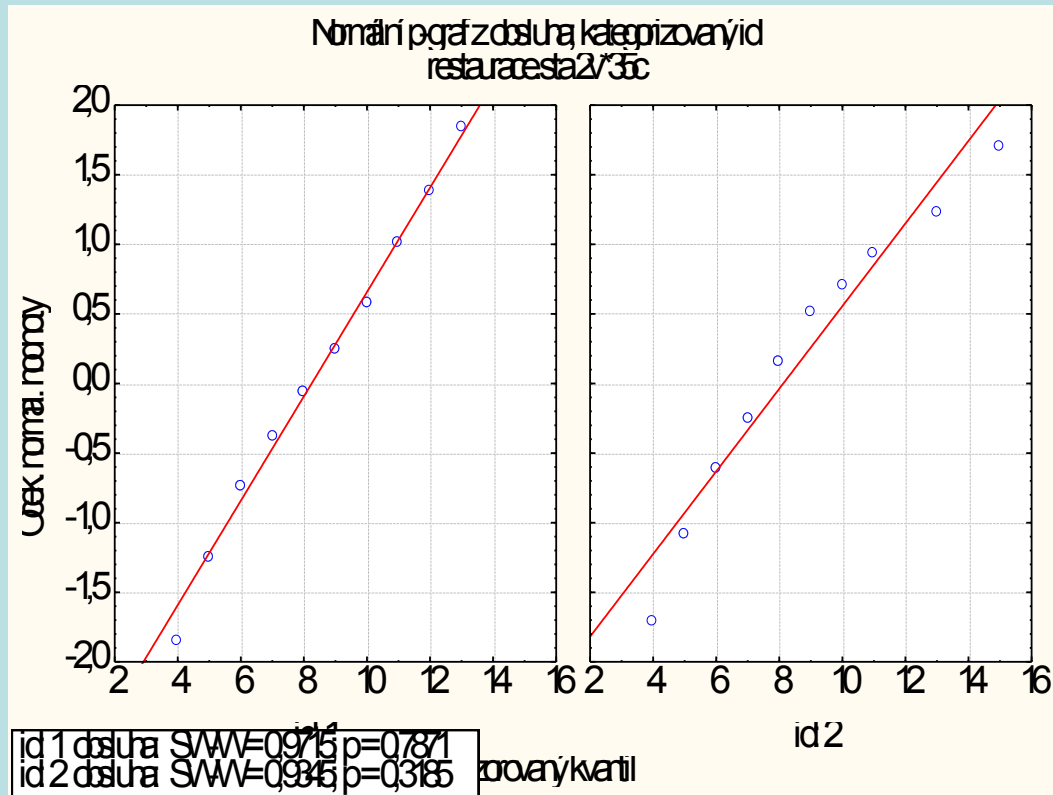
Stanovíme kritický obor:  $W = (-t_{\alpha/2, n_1+n_2-2}, t_{\alpha/2, n_1+n_2-2}) = (-t_{0,975, 33}, t_{0,975, 33}) = (-2,035, 2,035)$

Protože testová statistika se nerealizuje v kritickém oboru, nulovou hypotézu nezamítáme na hladině významnosti 0,05.

## Výpočet pomoci systému STATISTICA:

Otevřeme nový datový soubor o dvou proměnných a 35 případech. První proměnnou nazveme OBSLUHA, druhou ID. Do proměnné OBSLUHA napíšeme nejprve doby obsluhy v první restauraci a poté doby obsluhy ve druhé restauraci. Do proměnné ID, která slouží k rozlišení první a druhé restaurace, napíšeme 20 krát jedničku a 15 krát dvojku.

Pomocí NP-grafu ověříme normalitu dat v obou skupinách. Grafy – 2D Grafy – Normální pravděpodobnostní grafy – zaškrtneme S-W test - Proměnné OBSLUHA, OK, Kategorizovaný – Kategorie X, zaškrtneme Zapnuto, Změnit proměnnou – ID, OK. Dostaneme graf



V obou případech se tečky odchylují od přímky jenom málo a p-hodnoty S-W testu převyšují 0,05. Předpoklad o normálním rozložení dat v obou skupinách je oprávněný.

Nyní provedeme dvouvýběrový t-test současně s testem o shodě rozptylů:

Statistika – Základní statistiky a tabulky – t-test, nezávislé, dle skupin – OK, Proměnné – Závislé proměnné OBSLUHA,

Grupovací proměnná ID – OK.

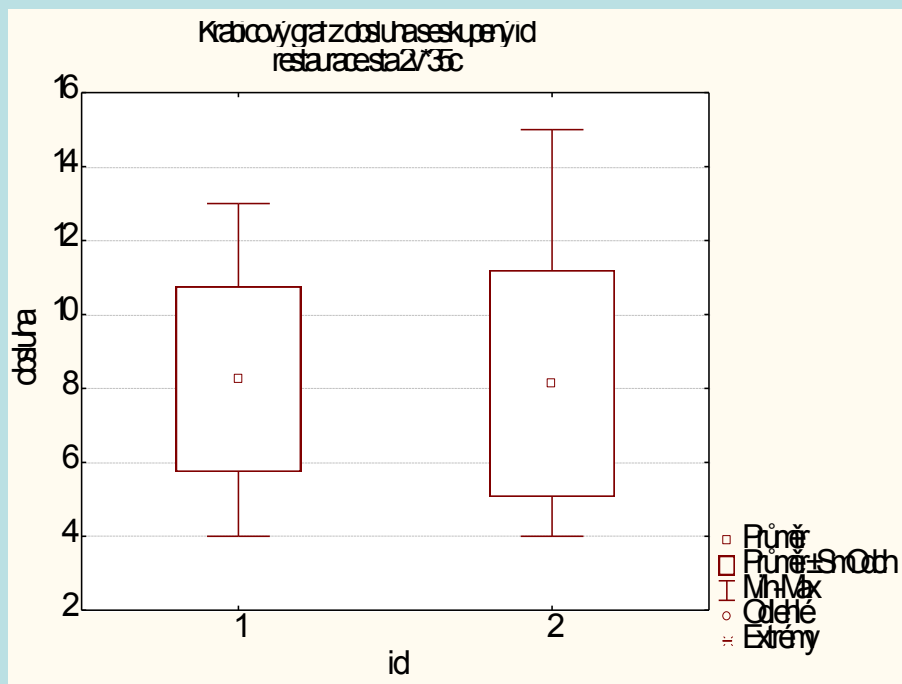
Po kliknutí na tlačítko Souhrn dostaneme tabulku

t-testy; grupovano: ID (restaurace)											
Skup. 1: 1											
Skup. 2: 2											
Proměr	Prům. 1	Prům. 2	t	s\	p	Poč.p 1	Poč.pl 2	Sm.od. 1	Sm.od. 2	F-pon rozpty	p rozpty
OBSLU	8,250	8,133	0,123	3	0,902	2	1	2,510	3,067	1,492	0,410

Vidíme, že testová statistika pro test shody rozptylů se realizuje hodnotou 1,492952 (je to převrácená hodnota k číslu 0,6702, které jsme vypočítali při ručním postupu), odpovídající p-hodnota je 0,41044, tedy na hladině významnosti 0,05 nezamítáme hypotézu o shodě rozptylů. (Upozornění: v případě zamítnutí hypotézy o shodě rozptylů je zapotřebí v tabulce t-testu pro nezávislé vzorky dle skupin zaškrtnout volbu Test se samostatnými odhady rozptylu.)

Dále z tabulky plyne, že testová statistika pro test shody středních hodnot se realizuje hodnotou 0,12373, počet stupňů volnosti je 33, odpovídající p-hodnota 0,902279, tedy hypotézu o shodě středních hodnot nezamítáme na hladině významnosti 0,05. Znamená to, že s rizikem omylu nejvýše 5% se neprokázal rozdíl ve středních hodnotách dob obsluhy v restauracích "U bílého koníčka" a „Zlatý lev“.

Tabulku ještě doplníme krabicovými diagramy. Na záložce Details zaškrtneme krabicový graf a vybereme volbu Průměr/SmOdch/Min-Max.



Z grafu je vidět, že průměrná doba obsluhy v první restauraci je nepatrně delší a má menší variabilitu než ve druhé restauraci. Extrémní ani odlehlé hodnoty se zde nevyskytují.

## Upozornění:

V případě, že známe realizace obou výběrových průměrů a směrodatných odchylek, můžeme pro provedení dvouvýběrového t-testu v systému STATISTICA použít aplikaci Tesy rozdílů. Postup si ukážeme na příkaldě s dobou obsluhy ve dvou restauracích

Statistiky – Základní statistiky a tabulky – Testy rozdílů: r, %, průměry – OK – vybereme Rozdíl mezi dvěma průměry (normální rozdělení) – do políčka Pr1 napíšeme 8,25, do políčka SmOd1 napíšeme 2,5105, do políčka N1 napíšeme 20, do políčka Pr2 napíšeme 8,25, do políčka SmOd2 napíšeme 3,0675, do políčka N2 napíšeme 15 – Výpočet. Dostaneme p-hodnotu 0,9023, tedy nezamítáme nulovou hypotézu na hladině významnosti 0,05.

The screenshot shows the 'Testy rozdílů: r, %, průměry: restaurace.sta' dialog box. The 'Rozdíl mezi dvěma průměry (normální rozdělení)' section is active, showing the following values:

Parameter	Value
Pr1	8,25
SmOd1	2,5105
N1	20
Pr2	8,1333
SmOd2	3,0675
N2	15
p-value	,9023

The 'Oboustr.' radio button is selected, and the 'Výpočet' button is visible. Other sections include 'Rozdíl mezi dvěma korelačními koeficienty' (r1: 0,00, N1: 10, p: 1,0000) and 'Rozdíl mezi dvěma poměry' (P 1: ,50000, N1: 10, p: 1,0000).

## Nepovinná část: Cohenův koeficient věcného účinku – doplnění významu dvouvýběrového t-testu:

Nechť  $X_{11}, \dots, X_{1n_1}$  je náhodný výběr z rozložení  $N(\mu_1, \sigma^2)$  a  $X_{21}, \dots, X_{2n_2}$  je na něm nezávislý náhodný výběr rozložení  $N(\mu_2, \sigma^2)$ , přičemž  $n_1 \geq 2$  a  $n_2 \geq 2$  a  $\sigma^2$  neznáme. Necht'  $c$  je konstanta.

Testujeme  $H_0: \mu_1 - \mu_2 = c$  proti  $H_1: \mu_1 - \mu_2 \neq c$ . Označme  $m_1, m_2$  realizace výběrových průměrů hodnot dané veličiny v těchto dvou skupinách,  $s_1^2, s_2^2$  realizace výběrových rozptylů a  $S_*^2 = \frac{n_1 - 1}{n_1 + n_2 - 2} s_1^2 + \frac{n_2 - 1}{n_1 + n_2 - 2} s_2^2$  realizaci váženého průměru výběrových rozptylů.

Cohenův koeficient  $d$  vypočteme podle vzorce:  $d = \frac{|m_1 - m_2|}{S_*}$ .

Tento koeficient slouží k posouzení velikosti rozdílu průměrů, který je standardizován pomocí odmocniny z váženého průměru výběrových rozptylů. Jedná se o tzv. **věcnou významnost** neboli **velikost účinku** skupiny na variabilitu hodnot sledované náhodné veličiny. Velikost účinku hodnotíme podle následující tabulky:

Hodnota $d$	účinek
aspoň 0,8	velký
mezi 0,5 až 0,8	střední
mezi 0,2 až 0,5	malý
pod 0,2	zanedbatelný

(Uvedené hodnoty nemají samozřejmě absolutní platnost, posouzení, jaký účinek považujeme za velký či malý, závisí na kontextu.)

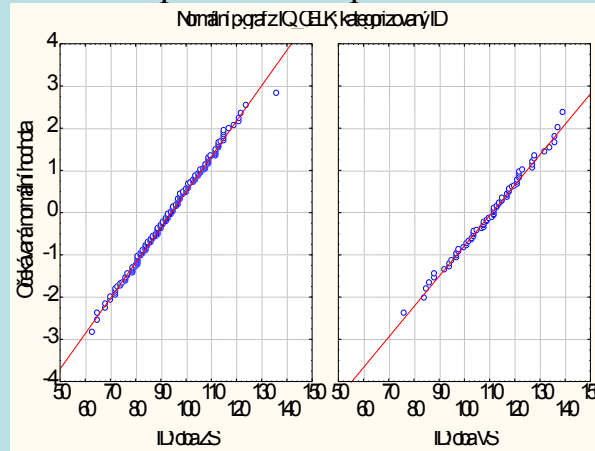
Je zapotřebí si uvědomit, že při dostatečně velkých rozsazích náhodných výběrů i malý rozdíl ve výběrových průměrech způsobí zamítnutí nulové hypotézy na hladině významnosti  $\alpha$ , i když z věcného hlediska tak malý rozdíl nemá význam. Naopak, máme-li výběry malých rozsahů, pak i značně velký rozdíl ve výběrových průměrech nemusí vést k zamítnutí nulové hypotézy na hladině významnosti  $\alpha$ .



## Příklad:

Máme k dispozici údaje o celkovém IQ 856 žáků ZŠ. Zajímáme se jednak o skupinu dětí, jejichž oba rodiče mají pouze základní vzdělání (je jich 296) a jednak o skupinu dětí, jejichž oba rodiče mají vysokoškolské vzdělání (těch je 75). Na hladině významnosti 0,05 budeme testovat hypotézu, že střední hodnota celkového IQ je v obou skupinách stejná a také vypočteme Cohenův koeficient věcného účinku.

**Řešení:** Normalitu dat v obou skupinách posoudíme pomocí N-P plotu:



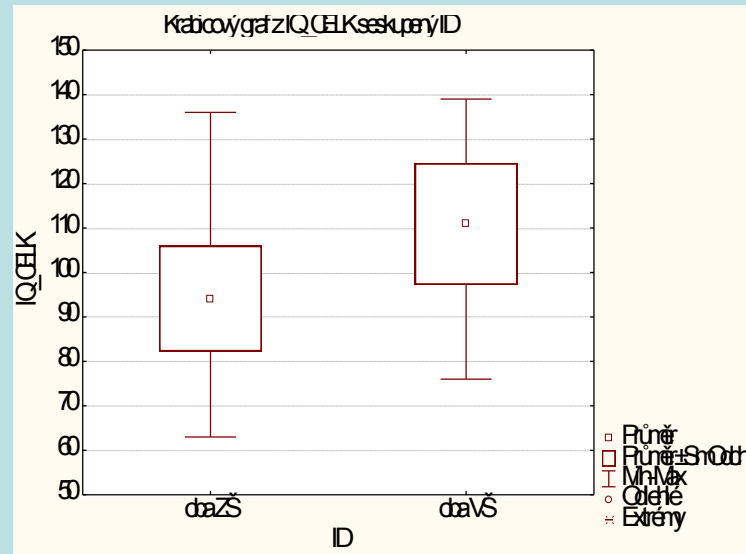
Vzhled N- P plotů v obou skupinách podporuje domněnku o normalitě dat.

Provedeme dvouvýběrový t-test:

t-testy; grupováno: ZŠ a VS (IQ)											
Skup. 1: oba ZŠ											
Skup. 2: oba VS											
Promě.	Prům.	Prům.	t	sv	p	Poč. p.	Poč. p.	Sm. od.	Sm. od.	F-pon	p
IQ CE	oba Z	oba V				oba Z	oba V	oba Z	oba V	Rozpt	Rozpt
IQ CE	94,13	110,9	-10,6	36	0,000	29	7	11,82	13,60	1,322	0,110

Hypotézu o shodě středních hodnot zamítáme na hladině významnosti 0,05, protože odpovídající p-hodnota je velmi blízká 0 (hypotézu o shodě rozptylů nezamítáme na hladině významnosti 0,05, p-hodnota F-testu je 0,110124, což je větší než 0,05).

Krabicový diagram:



Vidíme, že průměrné celkové IQ dětí v 1. skupině je 94,1, zatímco ve 2. skupině 110,9. Vliv skupiny na variabilitu hodnot celkového IQ posoudíme pomocí Cohenova koeficientu.

	1	2	3	4	5	6	7
	n1	n2	m1	m2	s1	s2	d
1	29	7	94,13	110,9	11,82	13,60	1,374

Cohenův koeficient nabývá hodnoty 1,37, tudíž vliv skupiny na variabilitu hodnot celkového IQ lze považovat za velký.

## Parametrické úlohy o jednom a dvou výběrech z alternativního rozložení

### Osnova:

Případ jednoho náhodného výběru

- asymptotické rozložení statistiky odvozené z výběrového průměru alternativního rozložení
- vzorec pro meze intervalu spolehlivosti pro parametr alternativního rozložení
- testování hypotézy o parametru alternativního rozložení

Případ dvou nezávislých náhodných výběrů

- asymptotické rozložení statistiky odvozené z výběrových průměrů dvou nezávislých alternativních rozložení
- vzorec pro meze intervalu spolehlivosti pro rozdíl parametrů dvou alternativních rozložení
- testování hypotézy o rozdílu parametrů dvou alternativních rozložení

**Případ jednoho náhodného výběru:** S náhodným výběrem rozsahu  $n$  z alternativního rozložení se setkáváme v situaci, kdy provádíme  $n$  opakovaných nezávislých pokusů a v každém z těchto pokusů sledujeme nastoupení úspěchu. Pravděpodobnost úspěchu je pro všechny pokusy stejná. Náhodná veličina  $X_i$  nabude hodnoty 1, pokud v  $i$ -tém pokusu nastal úspěch a hodnoty 0, pokud v  $i$ -tém pokusu úspěch nenastal,  $i = 1, 2, \dots, n$ . Realizací náhodného výběru  $X_1, \dots, X_n$  je tedy posloupnost 0 a 1.

### Opakování:

**Alternativní rozložení:** Náhodná veličina  $X$  udává počet úspěchů v jednom pokusu, přičemž pravděpodobnost úspěchu je  $q$ . Píšeme  $X \sim A(q)$ .

$$\pi(x) = \begin{cases} q^x (1-q)^{1-x} & \text{neboli } \pi(x) = \begin{cases} q^x (1-q)^{1-x} & \text{pro } x=0,1 \\ 0 & \text{jinak} \end{cases} \end{cases}$$

**Binomické rozložení:** Náhodná veličina  $X$  udává počet úspěchů v posloupnosti  $n$  nezávislých opakovaných pokusů, přičemž pravděpodobnost úspěchu je v každém pokusu  $q$ . Píšeme  $X \sim \text{Bi}(n, q)$ .

$$\pi(x) = \begin{cases} \binom{n}{x} q^x (1-q)^{n-x} & \text{pro } x=0, \dots, n \\ 0 & \text{jinak} \end{cases}$$

$$E(X) = nq, D(X) = nq(1-q)$$

(Alternativní rozložení je speciálním případem binomického rozložení pro  $n = 1$ .)

Jsou-li  $X_1, \dots, X_n$  stochasticky nezávislé náhodné veličiny,  $X_i \sim A(q)$ ,  $i = 1, \dots, n$ , pak  $X = \sum_{i=1}^n X_i \sim \text{Bi}(n, q)$ .

### Centrální limitní věta:

Jsou-li náhodné veličiny  $X_1, \dots, X_n$  stochasticky nezávislé a všechny mají stejné rozložení se střední hodnotou  $\mu$  a rozptylem  $\sigma^2$ , pak pro velká  $n$  ( $n \geq 30$ ) lze rozložení součtu  $\sum_{i=1}^n X_i$  aproximovat normálním rozložením  $N(n\mu, n\sigma^2)$ . Zkráceně píšeme

$\sum_{i=1}^n X_i \approx N(n\mu, n\sigma^2)$ . Pokud součet  $\sum_{i=1}^n X_i$  standardizujeme, tj. vytvoříme náhodnou veličinu  $U_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}}$ , pak rozložení této náhodné veličiny lze aproximovat standardizovaným normálním rozložením. Zkráceně píšeme  $U_n \approx N(0,1)$

### Asymptotické rozložení statistiky odvozené z výběrového průměru.

Nechť  $X_1, \dots, X_n$  je náhodný výběr z rozložení  $A(\theta)$  a necht' je splněna podmínka  $I(\theta) > 0$ . Pak statistika  $U = \frac{M - n\theta}{\sqrt{\frac{M}{n} - \theta^2}}$

konverguje v distribuci k náhodné veličině se standardizovaným normálním rozložením. (Říkáme, že  $U$  má asymptoticky rozložení  $N(0,1)$  a píšeme  $U \approx N(0,1)$ .)

### Vysvětlení:

Protože  $X_1, \dots, X_n$  je náhodný výběr z rozložení  $A(\theta)$ , bude mít statistika  $Y_n = \sum_{i=1}^n X_i$  (výběrový úhrn) rozložení  $Bi(n, \theta)$ .  $Y_n$

má střední hodnotu  $E(Y_n) = n\theta$  a rozptyl  $D(Y_n) = n\theta(1-\theta)$ . Podle centrální limitní věty se standardizovaná statistika  $U = \frac{Y_n - n\theta}{\sqrt{n\theta(1-\theta)}}$  asymptoticky řídí standardizovaným normálním rozložením  $N(0,1)$ . Pokud čitatele i jmenovatele podělíme  $n$ ,

dostaneme vyjádření:  $U = \frac{Y_n/n - \theta}{\sqrt{\frac{1}{n} \sum_{i=1}^n X_i - \theta^2}} = \frac{M/n - \theta}{\sqrt{\frac{M}{n} - \theta^2}}$

## Vzorec pro meze 100(1- $\alpha$ )% asymptotického empirického intervalu spolehlivosti pro parametr $\mu$ .

Meze 100(1- $\alpha$ )% asymptotického empirického intervalu spolehlivosti pro parametr  $\mu$  jsou:

$$d = \bar{x} - \frac{z_{1-\alpha/2} \sqrt{\hat{\sigma}^2}}{\sqrt{n}}, h = \bar{x} + \frac{z_{1-\alpha/2} \sqrt{\hat{\sigma}^2}}{\sqrt{n}}.$$

### Vysvětlení:

Pokud rozptyl  $\sigma^2 = \hat{\sigma}^2$  nahradíme odhadem  $\frac{M_2 - M_1^2}{n}$ , konvergence náhodné veličiny  $U$  k veličině s rozložením  $N(0,1)$  se neporuší. Tedy

$$\forall \epsilon > 0: 1 - P\left(-u_{1-\alpha/2} < \frac{M_2 - M_1^2}{\sqrt{n}} - M_1 < u_{1-\alpha/2}\right) \rightarrow 1 - P\left(-u_{1-\alpha/2} < M_1 - \frac{M_1^2 - M_2}{\sqrt{n}} < u_{1-\alpha/2}\right)$$

### Příklad:

Náhodně bylo vybráno 100 osob a zjištěno, že 34 z nich nakupuje v internetových obchodech. Najděte 95% asymptotický interval spolehlivosti pro pravděpodobnost, že náhodně vybraná osoba nakupuje v internetových obchodech.

### Řešení:

Zavedeme náhodné veličiny  $X_1, \dots, X_{100}$ , přičemž  $X_i = 1$ , když  $i$ -tá osoba nakupuje v internetových obchodech a  $X_i = 0$  jinak,  $i = 1, \dots, 100$ . Tyto náhodné veličiny tvoří náhodný výběr z rozložení  $A(\theta)$ .

$n = 100$ ,  $m = 34/100$ ,  $\alpha = 0,05$ ,  $u_{1-\alpha/2} = u_{0,975} = 1,96$ .

Ověření podmínky  $n\theta(1-\theta) > 9$ : parametr  $\theta$  neznáme, musíme ho nahradit výběrovým průměrem. Pak  $100 \cdot 0,34 \cdot 0,66 = 22,44 > 9$ .

$$d = 34 - \frac{0,34 \cdot 34}{100} = 33,19 \quad d = 34 + \frac{0,34 \cdot 34}{100} = 43,32$$

S pravděpodobností přibližně 0,95 tedy  $0,2472 < \theta < 0,4328$ . Znamená to, že s pravděpodobností přibližně 95% je v uvažované populaci nejméně 24,7% a nejvíce 43,3% osob, které nakupují v internetových obchodech.

## Výpočet pomocí systému STATISTICA:

### a) Přesný způsob

Otevřeme nový datový soubor se dvěma proměnnými a jedním případu.

První proměnnou nazveme d a do jejího Dlouhého jména napíšeme

$$=0,34-\text{sqrt}(0,34*0,66/100)*\text{VNormal}(0,975;0;1)$$

Druhou proměnnou nazveme h a do jejího Dlouhého jména napíšeme

$$=0,34+\text{sqrt}(0,34*0,66/100)*\text{VNormal}(0,975;0;1)$$

Dostaneme výsledek:

	1	2
	d	h
1	0,247	0,432

Vidíme, že s pravděpodobností aspoň 0,95 se pravděpodobnost nákupu v inetrnetových obchodech bude pohybovat v mezích 0,2471 až 0,4328.

### b) Přibližný způsob, použitelný pro dostatečně velký rozsah výběru

Do nového datového souboru o jedné proměnné X a 100 případech uložíme 34 jedniček (nakupování v internetových obchodech) a 66 nul.

Statistika – Základní statistiky a tabulky – Popisné statistiky – OK – Proměnné X – OK – Detailní výsledky – zaškrtneme Meze spolehl. prům. – ponecháme implicitní hodnotu pro Interval 95,00 – Výpočet.

Dostaneme tabulku:

Promě.	Popisné statistiky (Tabulka3)			
	N platn.	Prum.	Int. spoleh.	Int. spoleh.
X	10	0,340	0,245	0,434

Dospěli jsme k výsledku, že s pravděpodobností aspoň 0,95 se pravděpodobnost nákupu v inetrnetových obchodech bude pohybovat v mezích 0,2455 až 0,4345.



### c) Výpočet pomocí modulu Analýza síly testu

Statistiky – Analýza síly testu – Odhad intervalu – Jeden podíl, Z, Chí-kvadrát test – OK – Pozorovaný podíl  $p$ : 0,34, Velikost vzorku: 100, Spolehlivost: 0,95 – Vypočítat.

Dostaneme tabulku:

	Hodn
Podíl vzorku $p$	0,34
Velikost vz. ve s	100,0
Interval spolehliv	0,95
Meze spolehlivo	
$P_i$ (přesně):	
Dolní mez	0,24
Horní mez	0,44
$P_i$ (přibližně):	
Dolní mez	0,25
Horní mez	0,44
$P_i$ (původ.):	
Dolní mez	0,24
Horní mez	0,43

Zajímá nás výsledek uvedený v dolní části tabulky, tj.  $P_i$  (původ.). Zjistíme, že s pravděpodobností aspoň 0,95 se pravděpodobnost nákupu v internetových obchodech bude pohybovat v mezích 0,2455 až 0,4345.

**Příklad:** Kolik osob musíme vybrat, abychom podíl modrookých osob v populaci odhadli se spolehlivostí 90% a šířka intervalu spolehlivosti byla nanejvýš a) 0,06, b) 0,01?

**Řešení:**

Šířka  $100(1-\alpha)\%$  asymptotického empirického intervalu spolehlivosti pro parametr  $\theta$ :

$$h = \left( \hat{p} + \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} - \left( \hat{p} - \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) \right) = \frac{2\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$$

Požadujeme, aby  $h - d \leq \Delta$ , tedy  $2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq \Delta$ . Odtud vyjádříme  $n \geq \frac{4\hat{p}(1-\hat{p})}{\Delta^2}$ .

Předpokládejme, že nemáme žádné předběžné informace o podílu modrookých osob v populaci. Musíme tedy zvolit takové  $m$ , aby šířka intervalu spolehlivosti byla maximální. Maximalizujeme výraz  $\frac{4m(1-m)}{\Delta^2}$ . Derivujeme podle  $m$  a položíme rovno 0:  $1 - 2m = 0 \Rightarrow m = 0,5$ . V tomto případě volíme relativní četnost  $m = 0,5$ .

$$\text{ad a) } n \geq \frac{4m(1-m)}{\Delta^2} = \frac{4 \cdot 0,5 \cdot 0,5 \cdot 0,95^2}{0,06^2} = \frac{4 \cdot 0,5 \cdot 0,5 \cdot 0,9025}{0,0036} = 506,25 \approx 507$$

Uvedenou podmínku tedy splníme, když vybereme aspoň 507 osob.

$$\text{ad b) } n \geq \frac{4m(1-m)}{\Delta^2} = \frac{4 \cdot 0,5 \cdot 0,5 \cdot 0,95^2}{0,01^2} = \frac{4 \cdot 0,5 \cdot 0,5 \cdot 0,9025}{0,01} = 18056,25 \approx 18057$$

Chceme-li dosáhnout podstatně užšího intervalu spolehlivosti, musíme vybrat aspoň 18 057 osob.

**Modifikace:** Předpokládejme, že v populaci je nanejvýš 30% modrookých osob. Pak relativní četnost  $m = 0,3$ .

$$\text{ad a) } n \geq \frac{4m^2}{\lambda} = \frac{4 \cdot 0,3^2 \cdot 0,95^2}{0,06} = \frac{4 \cdot 0,3 \cdot 0,645}{0,06} = 3,4$$

V tomto případě stačí vybrat 632 osob.

Ve srovnání s předešlým případem vidíme, že rozsah výběru skutečně klesl.

ad b)

$$n \geq \frac{4m^2}{\lambda} = \frac{4 \cdot 0,3^2 \cdot 0,95^2}{0,01} = \frac{4 \cdot 0,3 \cdot 0,645}{0,01} = 27,6$$

V tomto případě musíme vybrat aspoň 22 731 osob.

## Testování hypotézy o parametru $\theta$

Nechť  $X_1, \dots, X_n$  je náhodný výběr z rozložení  $A(\theta)$  a necht' je splněna podmínka  $I_{\theta}$ .

Na asymptotické hladině významnosti  $\alpha$  testujeme hypotézu

$H_0: \theta = c$  proti alternativě  $H_1: \theta \neq c$  (resp.  $H_1: \theta < c$  resp.  $H_1: \theta > c$ ).

Testovým kritériem je statistika  $T_0 = \frac{\bar{X}_n - c}{\sqrt{\frac{1}{n} I_{\theta}}}$ , která v případě platnosti nulové hypotézy má asymptoticky rozložení  $N(0,1)$ .

Kritický obor má tvar  $W = (-\infty, -u_{1-\alpha/2}) \cup (u_{1-\alpha/2}, \infty)$  (resp.  $W = (-\infty, -u_{1-\alpha})$  resp.  $W = (u_{1-\alpha}, \infty)$ ).

(Testování hypotézy o parametru  $\theta$  lze samozřejmě provést i pomocí  $100(1-\alpha)\%$  asymptotického intervalu spolehlivosti nebo pomocí p-hodnoty.)

**Příklad:** Podíl zmetků při výrobě určité součástky činí  $Q = 0,01$ . Bylo náhodně vybráno 1000 výrobků a zjistilo se, že mezi nimi je 16 zmetků. Na asymptotické hladině významnosti 0,05 testujte hypotézu  $H_0: Q = 0,01$  proti oboustranné alternativě  $H_1: Q \neq 0,01$ .

### Řešení:

Zavedeme náhodné veličiny  $X_1, \dots, X_{1000}$ , přičemž  $X_i = 1$ , když  $i$ -tý výrobek byl zmetek a  $X_i = 0$  jinak,  $i = 1, \dots, 1000$ . Tyto náhodné veličiny tvoří náhodný výběr z rozložení  $A(Q)$ .

Testujeme hypotézu  $H_0: Q = 0,01$  proti alternativě  $H_1: Q \neq 0,01$ .

Známe:  $n = 1000$ ,  $m = 16$ ,  $c = 0,01$ ,  $\alpha = 0,05$ ,  $u_{1-\alpha/2} = u_{0,975} = 1,96$

Ověření podmínky  $1000 \cdot 0,01 \cdot 0,99 = 9,9 > 9$ .

#### a) Testování pomocí kritického oboru:

Realizace testového kritéria:  $t_0 = \frac{m - nQ}{\sqrt{nQ(1-Q)}} = \frac{16 - 10}{\sqrt{1000 \cdot 0,01 \cdot 0,99}} = 1,907$ .

Kritický obor:  $W = (-\infty, -1,96) \cup (1,96, \infty)$ . Protože  $1,907 \notin W$ ,  $H_0$  nezamítáme na asymptotické hladině významnosti 0,05.

#### b) Testování pomocí intervalu spolehlivosti

$d = -\frac{m}{n} + u_{-\alpha} \sqrt{\frac{m}{n} \left(1 - \frac{m}{n}\right)} = 0,01 - 1,96 \cdot \sqrt{0,01 \cdot 0,99} = 0,008$

$h = +\frac{m}{n} + u_{-\alpha} \sqrt{\frac{m}{n} \left(1 - \frac{m}{n}\right)} = 0,01 + 1,96 \cdot \sqrt{0,01 \cdot 0,99} = 0,023$

Protože číslo  $c = 0,01$  leží v intervalu 0,0082 až 0,0238,  $H_0$  nezamítáme na asymptotické hladině významnosti 0,05.

#### c) Testování pomocí p-hodnoty

Protože testujeme nulovou hypotézu proti oboustranné alternativě, vypočteme p-hodnotu podle vzorce:

$p = 2 \min \{ \Phi(1,907), 1 - \Phi(1,907) \} = 2 \min \{ 0,97104, 1 - 0,97104 \} = 0,05792$ .

Protože vypočtená p-hodnota je větší než hladina významnosti 0,05,  $H_0$  nezamítáme na asymptotické hladině významnosti 0,05.

## Výpočet pomocí systému STATISTICA

### a) Využití aplikace Testy rozdílů

Statistiky – Základní statistiky a tabulky – Testy rozdílů: r, %, průměry – OK – vybereme Rozdíl mezi dvěma poměry – do políčka P 1 napíšeme 0,016, do políčka N1 napíšeme 1000, do políčka P 2 napíšeme 0,01, do políčka N2 napíšeme 32767 (větší hodnotu systém neumožní) - Výpočet. Dostaneme p-hodnotu 0,0626, tedy nezamítáme nulovou hypotézu na hladině významnosti 0,05.

The screenshot shows the 'Testy rozdílů: r, %, průměry: Tabulka3' dialog box. It is divided into three sections for different types of tests. The 'Rozdíl mezi dvěma poměry' section is active, showing input values for P1 (0,01600), N1 (1000), P2 (0,01000), and N2 (32767), resulting in a p-value of 0,0626. The 'Rozdíl mezi dvěma korelačními koeficienty' section shows r1 and r2 both set to 0,00, N1 and N2 to 10, and a p-value of 1,0000. The 'Rozdíl mezi dvěma průměry (normální rozdělení)' section shows Pr1 and Pr2 both set to 0, SmOd1 and SmOd2 both set to 1, N1 and N2 to 10, and a p-value of 1,0000. The 'Výběrový průměr vs. střední hodnota' checkbox is unchecked. The 'Jednostr.' radio button is selected in the first two sections, while the 'Oboustr.' radio button is selected in the third section. A 'Výpočet' button is highlighted in the active section, and a 'Storno' button is visible in the top right.

Section	Parameter	Value
Rozdíl mezi dvěma korelačními koeficienty	r1	0,00
	r2	0,00
	N1	10
	N2	10
	p	1,0000
Rozdíl mezi dvěma průměry (normální rozdělení)	Pr1	0,
	Pr2	0,
	SmOd1	1,
	SmOd2	1,
	N1	10
Rozdíl mezi dvěma poměry	P 1	0,01600
	P 2	0,01000
	N1	1000
	N2	32767
	p	0,0626

## b) Využití modulu Analýza síly testu

Statistiky – Analýza síly testu – Odhad intervalu – Jeden podíl, Z, Chí-kvadrát test – OK – Pozorovaný podíl  $p$ : 0,016, Velikost vzorku: 1000, Spolehlivost: 0,95 – Vypočítat.

Dostaneme tabulku:

	Hodnota
Podíl vzorku $p$	0,01
Velikost vz. ve s	1000,0
Interval spolehliv	0,95
Meze spolehlivo	
$P_i$ (přesně):	
Dolní mez	0,00
Horní mez	0,02
$P_i$ (přibližně):	
Dolní mez	0,00
Horní mez	0,02
$P_i$ (původ.):	
Dolní mez	0,00
Horní mez	0,02

Zajímá nás výsledek uvedený v dolní části tabulky, tj.  $P_i$  (původ.). Zjistíme, že s pravděpodobností aspoň 0,95 se pravděpodobnost vyrobení zmetku bude pohybovat v mezích 0,00822 až 0,0238. Protože tento interval obsahuje číslo 0,01, nelze nulovou hypotézu zamítnout na asymptotické hladině významnosti 0,05.

**Příklad:** Nový léčebný postup považujeme za úspěšný, pokud po jeho ukončení bude dosaženo zlepšení zdravotního stavu u alespoň 50% zúčastněných pacientů. Nová terapie byla vyzkoušena u 40 pacientů a ke zlepšení došlo u 24 osob, tj. u 60%. Je možné na asymptotické hladině významnosti 0,05 zamítnout hypotézu, že tato terapie nedosahuje úspěšnosti aspoň 50%?

### Řešení:

Zavedeme náhodné veličiny  $X_1, \dots, X_{40}$ , přičemž  
 $X_i = 1$ , když terapie u  $i$ -tého pacienta byl úspěšná a  
 $X_i = 0$  jinak,  
 $i = 1, \dots, 40$ .

Tyto náhodné veličiny tvoří náhodný výběr z rozložení  $A(\theta)$ .

Testujeme hypotézu  $H_0: \theta \leq 0,5$  proti pravostranné alternativě  $H_1: \theta > 0,5$ .

Známe:  $n = 40$ ,  $\hat{\theta} = \frac{24}{40} = 0,6$ ,  $c = 0,5$ ,  $\alpha = 0,05$ ,  $u_{1-\alpha} = u_{0,95} = 1,645$

Ověření podmínky  $n\theta_0(1-\theta_0) \geq 9$ :  $40 \cdot 0,5 \cdot 0,5 = 10 > 9$ .

Realizace testového kritéria:  $t_0 = \frac{\hat{\theta} - c}{\sqrt{\frac{c(1-c)}{n}}} = \frac{0,6 - 0,5}{\sqrt{\frac{0,5 \cdot 0,5}{40}}} = 1,2649$

Kritický obor:  $W = \{ \theta : \theta > u_{0,95} = 1,645 \}$ .

Protože  $1,2649 \notin W$ ,  $H_0$  nezamítáme na asymptotické hladině významnosti 0,05.



## Výpočet pomocí systému STATISTICA:

The screenshot shows the 'Testy rozdílů: r, %, průměry: Tabulka9' dialog box in STATISTICA. It contains three sections for different types of tests, each with input fields and a 'Výpočet' button.

- Top section: Rozdíl mezi dvěma korelačními koeficienty**
  - Inputs: r1: 0,00, N1: 10, r2: 0,00, N2: 10, p: 1,0000
  - Options:  Jednostr.,  Oboustr.
  - Buttons: 'Storno' and 'Výpočet'
- Middle section: Rozdíl mezi dvěma průměry (normální rozdělení)**
  - Inputs: Pr1: 0, SmOd1: 1, N1: 10, Pr2: 0, SmOd2: 1, N2: 10, p: 1,0000
  - Options:  Jednostr.,  Oboustr.
  - Checkbox:  Výběrový průměr vs. střední hodnota
  - Buttons: 'Výpočet'
- Bottom section: Rozdíl mezi dvěma poměry**
  - Inputs: P 1: ,60000, N1: 40, P 2: ,50000, N2: 32767, p: ,1031
  - Options:  Jednostr.,  Oboustr.
  - Buttons: 'Výpočet'

Vypočtená p-hodnota jednostranného testu je 0,1031, tedy větší než asymptotická hladina významnosti 0,05.  $H_0$  nezamítáme na asymptotické hladině významnosti 0,05.

**Případ dvou nezávislých výběrů z alternativních rozložení:** Provádíme opakovaně nezávisle  $n_1$ -krát jeden náhodný pokus a nezávisle na tom  $n_2$ -krát druhý náhodný pokus. V první sérii pokusů sledujeme nějaký jev, který v každém pokusu může nastat s pravděpodobností  $\theta_1$  a ve druhé sérii pokusů sledujeme nějaký jiný jev, jehož pravděpodobnost nastoupení je  $\theta_2$ . Parametry  $\theta_1, \theta_2$  neznáme. Naším úkolem bude konstruovat interval spolehlivosti pro parametrickou funkci  $\theta_1 - \theta_2$  nebo testovat hypotézu o této parametrické funkci, a to pomocí dvou nezávislých náhodných výběrů z alternativních rozložení  $A_{\theta_1}, A_{\theta_2}$ .

### Asymptotické rozložení statistiky odvozené ze dvou výběrových průměrů alternativních rozložení

Nechť  $X_{11}, \dots, X_{1n_1}$  je náhodný výběr z alternativního rozložení  $A(\theta_1)$  a  $X_{21}, \dots, X_{2n_2}$  je na něm nezávislý náhodný výběr alternativního rozložení  $A(\theta_2)$  a necht' jsou splněny podmínky  $n_1 \theta_1 (1 - \theta_1) > 9$  a  $n_2 \theta_2 (1 - \theta_2) > 9$ . Označme  $M_1, M_2$  výběrové průměry.

Pak statistika 
$$U = \frac{M_1 - M_2 - (\theta_1 - \theta_2)}{\sqrt{\frac{\theta_1(1-\theta_1)}{n_1} + \frac{\theta_2(1-\theta_2)}{n_2}}} \approx N(0, 1).$$

**Vysvětlení:** Analogicky jako v případě jednoho náhodného výběru z alternativního rozložení.

## Vzorec pro meze 100(1- $\alpha$ )% asymptotického empirického intervalu spolehlivosti pro parametrickou funkci $g(\theta)$ .

Meze 100(1- $\alpha$ )% asymptotického empirického intervalu spolehlivosti pro  $g(\theta)$  jsou:

$$d = \hat{g}(\hat{\theta}) - \frac{\sqrt{m(1-\alpha)} \cdot \hat{m}(1-\alpha)}{\sqrt{\frac{1}{I_1} + \frac{1}{I_2}}} u_{1-\alpha/2}, \quad h = \hat{g}(\hat{\theta}) + \frac{\sqrt{m(1-\alpha)} \cdot \hat{m}(1-\alpha)}{\sqrt{\frac{1}{I_1} + \frac{1}{I_2}}} u_{1-\alpha/2}$$

**Vysvětlení:** Pokud rozptyl  $\text{DM}_i^2 = \frac{1}{I_i}$  nahradíme odhadem  $\frac{M_i^2}{\zeta_i^2}$ ,  $i = 1, 2$ , konvergence náhodné veličiny  $U$

k veličině s rozložením  $N(0,1)$  se neporuší. Tedy

$$\forall \epsilon > 0 \quad \left( \frac{\sqrt{m(1-\alpha)} \cdot \hat{m}(1-\alpha)}{\sqrt{\frac{1}{I_1} + \frac{1}{I_2}}} < \frac{M_1^2}{\zeta_1^2} + \frac{M_2^2}{\zeta_2^2} < \frac{\sqrt{m(1-\alpha)} \cdot \hat{m}(1-\alpha)}{\sqrt{\frac{1}{I_1} + \frac{1}{I_2}}} + \epsilon \right) = \\ P\left( \frac{M_1^2}{\zeta_1^2} + \frac{M_2^2}{\zeta_2^2} < \frac{\sqrt{m(1-\alpha)} \cdot \hat{m}(1-\alpha)}{\sqrt{\frac{1}{I_1} + \frac{1}{I_2}}} + \epsilon \right)$$

**Příklad:** Management supermarketu vyhlásil týden slev a sledoval, zda toto vyhlášení má vliv na podíl větších nákupů (nad 500 Kč). Na základě náhodného výběru 200 zákazníků v týdnu bez slev bylo zjištěno 97 velkých nákupů, zatímco v týdnu se slevou z 300 náhodně vybraných zákazníků učinilo velký nákup 162 zákazníků. Sestrojte 95% asymptotický interval spolehlivosti pro rozdíl pravděpodobností uskutečnění většího nákupu v týdnu bez slevy a v týdnu se slevou.

**Řešení:**

Zavedeme náhodnou veličinu  $X_{1i}$ , která bude nabývat hodnoty 1, když v týdnu bez slevy  $i$ -tý náhodně vybraný zákazník uskuteční větší nákup a hodnoty 0 jinak,  $i = 1, \dots, 200$ . Náhodné veličiny  $X_{1,1}, \dots, X_{1,200}$  tvoří náhodný výběr z rozložení  $A_{p_1}$ . Dále zavedeme náhodnou veličinu  $X_{2i}$ , která bude nabývat hodnoty 1, když v týdnu se slevou  $i$ -tý náhodně vybraný zákazník uskuteční větší nákup a hodnoty 0 jinak,  $i = 1, \dots, 300$ . Náhodné veličiny  $X_{2,1}, \dots, X_{2,300}$  tvoří náhodný výběr z rozložení  $A_{p_2}$ .

$$n_1 = 200, n_2 = 300, m_1 = 97/200 = 0,485, m_2 = 162/300 = 0,54.$$

Ověření podmínek  $n_1 p_1 (1 - p_1) > 9$  a  $n_2 p_2 (1 - p_2) > 9$ : Parametry  $p_1$  a  $p_2$  neznáme, nahradíme je odhady  $m_1$  a  $m_2$ , tedy  $97 \cdot (1 - 97/200) = 49,955 > 9$ ,  $162 \cdot (1 - 162/300) = 74,52 > 9$ .

Meze  $100(1-\alpha)\%$  asymptotického empirického intervalu spolehlivosti pro parametrickou funkci  $p_1 - p_2$  jsou:

$$d = \left( \frac{m_1(1 - m_1)}{n_1} + \frac{m_2(1 - m_2)}{n_2} \right) u_{1-\alpha/2} = \left( \frac{97 \cdot 103}{200} + \frac{162 \cdot 138}{300} \right) 1,96 = 0,1443$$

$$h = \left( \frac{m_1(1 - m_1)}{n_1} + \frac{m_2(1 - m_2)}{n_2} \right) u_{\alpha/2} = \left( \frac{97 \cdot 103}{200} + \frac{162 \cdot 138}{300} \right) 1,96 = 0,0343$$

Zjistili jsme tedy, že s pravděpodobností přibližně 0,95:  $-0,1443 < p_1 - p_2 < 0,0343$ .

## Testování hypotézy o parametrické funkci $q$

Nechť  $X_{11}, \dots, X_{1n_1}$  je náhodný výběr z alternativního rozložení  $A(q)$  a  $X_{21}, \dots, X_{2n_2}$  je na něm nezávislý náhodný výběr alternativního rozložení  $A(q)$  a necht' jsou splněny podmínky  $n_1 q (1-q) > 9$  a  $n_2 q (1-q) > 9$ . Na asymptotické hladině významnosti  $\alpha$  testujeme nulovou hypotézu  $H_0: q = c$  proti alternativě  $H_1: q \neq c$  (resp.  $H_1: q < c$  resp.  $H_1: q > c$ ).

Testovým kritériem je statistika

$T_0 = \frac{M_1 - \frac{1}{2} + \frac{1}{2} \frac{M_2 - 1}{n_2}}{\sqrt{\frac{M_1(1-M_1/n_1) + (M_2-1)(1-(M_2-1)/n_2)}{n_1 + n_2 - 1}}}$ , která v případě platnosti nulové hypotézy má asymptoticky rozložení  $N(0,1)$ .

Kritický obor má tvar  $W = (-\infty, u_{1-\alpha/2}) \cup (u_{\alpha/2}, \infty)$

(resp.  $W = (-\infty, u_{1-\alpha})$  resp.  $W = (u_{\alpha}, \infty)$ ).

(Testování hypotézy o parametrické funkci  $q$  lze provést též pomocí  $100(1-\alpha)\%$  asymptotického intervalu spolehlivosti nebo pomocí p-hodnoty.)

### Poznámka: Postup při testování hypotézy $H_0: \mu_1 = \mu_2$ .

Je-li  $c = 0$ , pak označme  $M_* = \frac{n_1 M_1 + n_2 M_2}{n_1 + n_2}$  vážený průměr výběrových průměrů. Jako testová statistika slouží

$T_0 = \frac{M_1 - M_2}{\sqrt{M_* \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$ , která v případě platnosti nulové hypotézy má asymptoticky rozložení  $N(0,1)$ .

Kritický obor má tvar  $W_{\alpha/2, \alpha/2, \infty}$  (resp.  $W_{\alpha/2, \alpha/2}$  resp.  $W_{\alpha/2, 0}$ ).

Testová statistika  $T_0$  vznikne standardizací statistiky  $M_1 - M_2$ , kde neznámé parametry  $\mu_1, \mu_2$  nahradíme společným odhadem  $M_*$ .

**Příklad:** Pro údaje z příkladu o slevách v supermarketu testujte na asymptotické hladině významnosti 0,05 hypotézu, že týden se slevami nezvýší pravděpodobnost uskutečnění většího nákupu.

**Řešení:**

Testujeme hypotézu  $\mu = 0$  proti levostranné alternativě  $H_1: \mu < 0$  na asymptotické hladině významnosti 0,05.

$n_1 = 200$ ,  $n_2 = 300$ ,  $m_1 = 97/200$ ,  $m_2 = 162/300$ ,  $m_* = (97 + 162)/500 = 0,518$ .

Podmínky dobré aproximace byly ověřeny v předešlém příkladu.

**Testování pomocí intervalu spolehlivosti:**

Pro levostrannou alternativu používáme pravostranný interval spolehlivosti:

$$h = \mu + \frac{m_1(1-\alpha) + m_2(1-\alpha)}{1_1 + 1_2} u_{1-\alpha} = 0 + 0 + \frac{0,97 + 0,162}{200 + 300} 1,64 = 0,02$$

Protože číslo  $c = 0$  je obsaženo v intervalu  $(-\infty, 0,02]$ ,  $H_0$  nezamítáme na asymptotické hladině významnosti 0,05.

**Testování pomocí kritického oboru:**

Realizace testového kritéria:

$$t_0 = \frac{m_1}{n_1} - \frac{m_2}{n_2} = \frac{97}{200} - \frac{162}{300} = -0,18$$

Kritický obor je  $W = (-\infty, -1,64] = (-\infty, -0,54]$ . Protože testové kritérium nepatří do kritického oboru,  $H_0$  nezamítáme na asymptotické hladině významnosti 0,05.

**Testování pomocí p-hodnoty:**

Pro levostrannou alternativu se p-hodnota počítá podle vzorce  $p = P(T_0 \leq t_0)$ :

$$p = P_0 < 0,05 = 0,05 = 0,05 = 0,386 = 13$$

Protože p-hodnota je větší než 0,05,  $H_0$  nezamítáme na asymptotické hladině významnosti 0,05.

## Výpočet pomocí systému STATISTICA:

Statistiky – Základní statistiky a tabulky – Testy rozdílů: r, %, průměry – OK – vybereme Rozdíl mezi dvěma poměry – do políčka P 1 napíšeme 0,485, do políčka N1 napíšeme 200, do políčka P 2 napíšeme 0,54, do políčka N2 napíšeme 300 – zaškrtneme Jednostr. - Výpočet. Dostaneme p-hodnotu 0,1142, tedy nezamítáme nulovou hypotézu na hladině významnosti 0,05.

Testy rozdílů: r, %, průměry: tram\_bus

Poslat/tisknout výsledky každ. výpočtu do okna protokolu

Rozdíl mezi dvěma korelačními koeficienty

r1: 0,00 N1: 10 r2: 0,00 N2: 10 p: 1,0000  Jednostr.  Oboustr.

Rozdíl mezi dvěma průměry (normální rozdělení)

Pr1: 0, SmOd1: 1, N1: 10 Pr2: 0, SmOd2: 1, N2: 10 p: 1,0000  Jednostr.  Oboustr.

Výběrový průměr vs. střední hodnota

Rozdíl mezi dvěma poměry

P 1: .48500 N1: 200 P 2: .54000 N2: 300 p: .1142  Jednostr.  Oboustr.



## Parametrické úlohy o více nezávislých náhodných výběrech

### Osnova:

Porovnání aspoň tří nezávislých náhodných výběrů z normálních rozložení (jednofaktorová analýza rozptylu)

- testování hypotézy o shodě středních hodnot
- testování hypotézy o shodě rozptylů (testy homogenity rozptylů)
- zkoumání vlastností testů homogenity pomocí simulačních studií
- post-hoc metody mnohonásobného porovnávání

Porovnání aspoň tří nezávislých náhodných výběrů z alternativních rozložení

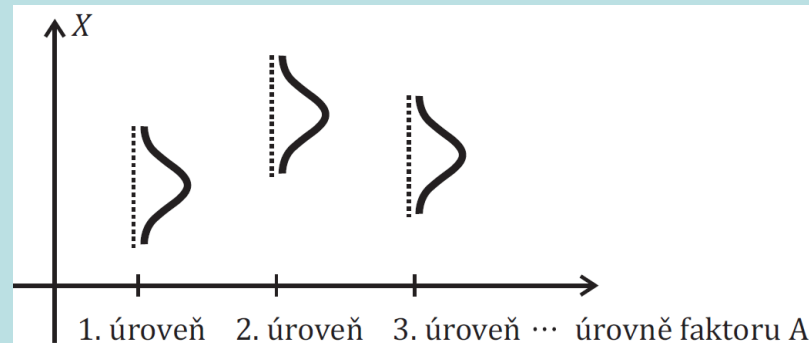
- test homogenity binomických rozložení
- mnohonásobné porovnávání

## I. Příklad $r \geq 3$ nezávislých náhodných výběrů z normálních rozložení (Analýza rozptylu jednoduchého třídění)

**Motivace:** Zajímáme se o problém, zda lze určitým faktorem (tj. nominální náhodnou veličinou A) vysvětlit variabilitu pozorovaných hodnot náhodné veličiny X, která je intervalového či poměrového typu. Např. zkoumáme, zda metoda výuky určitého předmětu (faktor A) ovlivňuje počet bodů dosažených studenty v závěrečném testu (náhodná veličina X). Předpokládáme, že faktor A má  $r \geq 3$  úrovní a přitom i-té úrovni odpovídá  $n_i$  pozorování  $X_{i1}, \dots, X_{in_i}$ , které tvoří náhodný výběr z rozložení  $N(\mu_i, \sigma^2)$ ,  $i = 1, \dots, r$  a jednotlivé náhodné výběry jsou stochasticky nezávislé, tedy  $X_{ij} = \mu_i + \varepsilon_{ij}$ , kde  $\varepsilon_{ij}$  jsou stochasticky nezávislé náhodné veličiny s rozložením  $N(0, \sigma^2)$ ,  $i = 1, \dots, r$ ,  $j = 1, \dots, n_i$ . Výsledky lze zapsat do tabulky

faktor A	výsledky
úroveň 1	$X_{11}, \dots, X_{1n_1}$
úroveň 2	$X_{21}, \dots, X_{2n_2}$
...	...
úroveň r	$X_{r1}, \dots, X_{rn_r}$

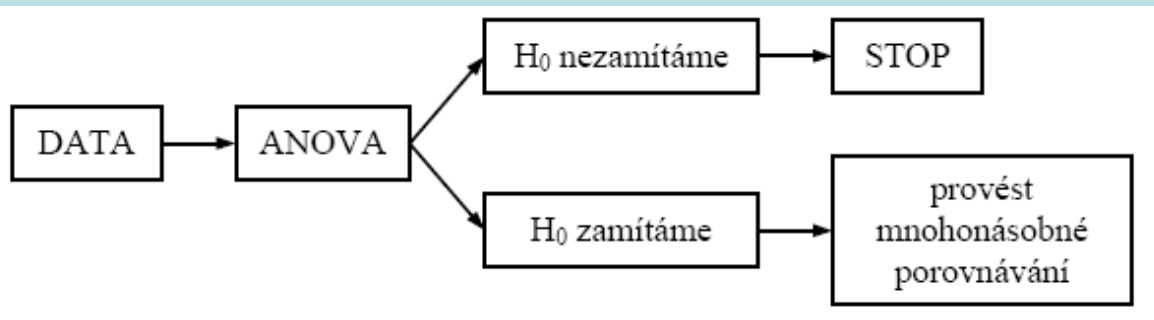
Ilustrace:



Na hladině významnosti  $\alpha$  testujeme nulovou hypotézu, která tvrdí, že všechny střední hodnoty jsou stejné, tj.  $H_0: \mu_1 = \dots = \mu_r$  proti alternativní hypotéze  $H_1$ , která tvrdí, že aspoň jedna dvojice středních hodnot se liší.

Jedná se tedy o zobecnění dvouvýběrového t-testu a na první pohled se zdá, že stačí utvořit  $\binom{r}{2}$  dvojic náhodných výběrů a na každou dvojici aplikovat dvouvýběrový t-test. Hypotézu o shodě všech středních hodnot bychom pak zamítli, pokud aspoň v jednom případě z  $\binom{r}{2}$  porovnávání se prokáže odlišnost středních hodnot. Odtud je vidět, že k neoprávněnému zamítnutí nulové hypotézy (tj. k chybě 1. druhu) může dojít s pravděpodobností větší než  $\alpha$ . Proto ve 30. letech 20. století vytvořil R. A. Fisher metodu ANOVA (analýza rozptylu, v popsané situaci konkrétně analýza rozptylu jednoduchého třídění), která uvedenou podmínku splňuje.

Pokud na hladině významnosti  $\alpha$  zamítneme nulovou hypotézu, zajímá nás, které dvojice středních hodnot se od sebe liší. K řešení tohoto problému slouží metody mnohonásobného porovnávání, např. Scheffého nebo Tukeyova metoda.



## Označení:

V analýze rozptylu jednoduchého třídění se používá tzv. tečková notace.

$n_{i1}^r \dots n_{i1}^r$  ... celkový rozsah všech r výběrů

$X_{i1}^r \dots X_{i1}^r$  ... součet hodnot v i-tém výběru

$M_{i1}^r \dots M_{i1}^r$  ... výběrový průměr v i-tém výběru

$X_{i1}^r \dots X_{i1}^r$  ... součet hodnot všech výběrů

$M_{i1}^r \dots M_{i1}^r$  ... celkový průměr všech r výběrů

Zavedeme součty čtverců

$S_T = \sum_{i=1}^r \sum_{j=1}^n X_{ij} - M_{..}^2$  ... **celkový součet čtverců** (charakterizuje variabilitu jednotlivých pozorování kolem celkového průměru),

počet stupňů volnosti  $f_T = n - 1$ ,

$S_A = \sum_{i=1}^r n_i M_{i.} - M_{..}^2$  ... **skupinový součet čtverců** (charakterizuje variabilitu mezi jednotlivými náhodnými výběry),

počet stupňů volnosti  $f_A = r - 1$ .

Sčítanec  $M_{i.}$  představuje bodový odhad efektu  $\alpha_i$ .

$S_E = \sum_{i=1}^r \sum_{j=1}^n X_{ij} - M_{i.}^2$  ... **reziduální součet čtverců** (charakterizuje variabilitu uvnitř jednotlivých výběrů),

počet stupňů volnosti  $f_E = n - r$ .

Lze dokázat, že  $S_T = S_A + S_E$ .

(Důkaz je proveden např. ve skriptech Budíková, Mikoláš, Osecký: Popisná statistika v poznámce 5.20.)

## Testování hypotézy o shodě středních hodnot

Náhodné veličiny  $X_{ij}$  se řídí modelem

$$M_0: X_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

pro  $i = 1, \dots, r, j = 1, \dots, n_i$ , přičemž

$\varepsilon_{ij}$  jsou stochasticky nezávislé náhodné veličiny s rozložením  $N(0, \sigma^2)$ ,

$\mu$  je společná část střední hodnoty závisle proměnné veličiny,

$\alpha_i$  je efekt faktoru A na úrovni i.

Parametry  $\mu, \alpha_i$  neznáme.

Požadujeme, aby platila tzv. **reparametrizační rovnice**:  $\sum_{i=1}^r \alpha_i = 0$ .

(Pokud je třídění vyvážené, tj. pokud mají všechny výběry stejný rozsah:  $n_1 = n_2 = \dots = n_r$ , pak lze použít zjednodušenou podmínku  $\sum_{i=1}^r \alpha_i = 0$ .)

Kdyby nezáleželo na faktoru A, platila by hypotéza  $\alpha_1 = \dots = \alpha_r = 0$  a dostali bychom model

$$M1: X_{ij} = \mu + \varepsilon_{ij}.$$

Během analýzy rozptylu tedy zkoumáme, zda výběrové průměry  $M_1, \dots, M_r$  se od sebe liší pouze v mezích náhodného kolísání kolem celkového průměru  $M$  nebo zda se projevuje vliv faktoru A.

Rozdíl mezi modely  $M_0$  a  $M_1$  ověřujeme pomocí testové statistiky

$F_A = \frac{S_A/f_A}{S_E/f_E}$ , která se řídí rozložením  $F(r-1, n-r)$ , je-li model  $M_1$  správný. Hypotézu o nevýznamnosti faktoru A tedy zamítneme na hladině významnosti  $\alpha$ , když platí:  $F_A \geq F_{1-\alpha}(r-1, n-r)$ .

Výsledky výpočtů zapisujeme do **tabulky analýzy rozptylu jednoduchého třídění**.

Zdroj variability	součet čtverců	stupně volnosti	podíl	$F_A$
skupiny	$S_A$	$f_A = r - 1$	$S_A/f_A$	$\frac{S_A/f_A}{S_E/f_E}$
reziduální	$S_E$	$f_E = n - r$	$S_E/f_E$	-
celkový	$S_T$	$f_T = n - 1$	-	-

Sílu závislosti náhodné veličiny X na faktoru A můžeme měřit pomocí **poměru determinace**:  $R^2 = \frac{S_A}{S_T}$ . Nabývá hodnot z intervalu  $\langle 0, 1 \rangle$ .

## Testování hypotézy o shodě rozptylů

Před provedením analýzy rozptylu je zapotřebí ověřit předpoklad o shodě rozptylů v daných  $r$  výběrech.

a) **Levenův test:** Položme  $Z_{ij} = \frac{X_{ij} - \bar{X}_i}{s_i}$ . Označíme

$$M_{Zi} = \frac{1}{n_i} \sum_{j=1}^{n_i} Z_{ij},$$

$$M_Z = \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^{n_i} Z_{ij},$$

$$S_{ZE} = \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^{n_i} (Z_{ij} - M_{Zi})^2,$$

$$S_{ZA} = \frac{1}{n} \sum_{i=1}^r n_i (M_{Zi} - M_Z)^2$$

Platí-li hypotéza o shodě rozptylů, pak statistika

$$F_{ZA} = \frac{S_{ZA}/r}{S_{ZE}/(n-r)} \approx F(r-1, n-r).$$

Hypotézu o shodě rozptylů tedy zamítáme na asymptotické hladině významnosti  $\alpha$ , když  $F_{ZA} \geq F_{1-\alpha}(r-1, n-r)$ .

(Levenův test je vlastně založen na analýze rozptylu absolutních hodnot centrovaných pozorování. Vzhledem k tomu, že náhodné veličiny  $X_{ij} - M_i$  nejsou stochasticky nezávislé a absolutní hodnoty těchto veličin nemají normální rozložení, je Levenův test pouze aproximační.)



b) **Brownův – Forsytheův test** je modifikací Levenova testu. Modifikace spočívá v tom, že místo výběrového průměru  $i$ -tého výběru se při výpočtu veličiny  $Z_{ij}$  používá medián  $i$ -tého výběru.

c) **Bartlettův test**: Platí-li hypotéza o shodě rozptylů a rozsahy všech výběrů jsou větší než 6, pak statistika  $B = C \frac{1}{n-r} \ln S^2 - \sum_{i=1}^r \frac{n_i}{n} \ln S_i^2$  se asymptoticky řídí rozložením  $\chi^2_{r-1}$ . Přitom konstanta  $C = \frac{1}{3(r-1)} \left( \sum_{i=1}^r \frac{1}{n_i} - \frac{1}{n-r} \right)$  a  $S_i^2$  je vážený průměr výběrových rozptylů.  
 $H_0$  zamítáme na asymptotické hladině významnosti  $\alpha$ , když  $B$  se realizuje v kritickém oboru  $W = \chi^2_{r-1, \alpha}$ .

## Zkoumání vlastností uvedených tří testů

Pro odhad pravděpodobnosti chyby 1. druhu bylo vždy vygenerováno 100 000 náhodných výběrů, a to postupně z těchto rozložení:

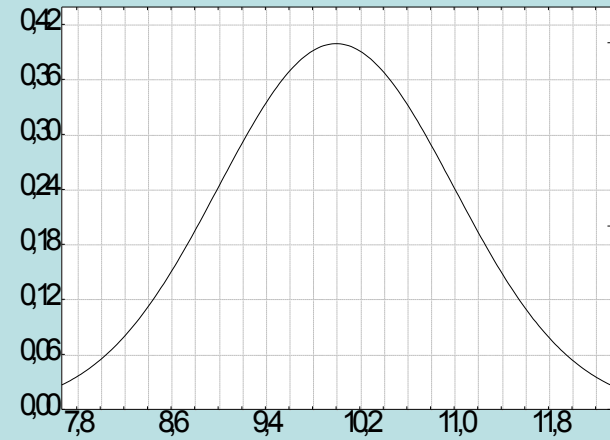
$N(10; 1)$ ,  $t(10)$ ,  $LN(1; 0,4)$ ,  $Ex(0,85)$ .

Všechny výběry měly stejný rozsah od 3 do 11 s krokem 2, počet výběrů byl od 2 do 10 s krokem 2.

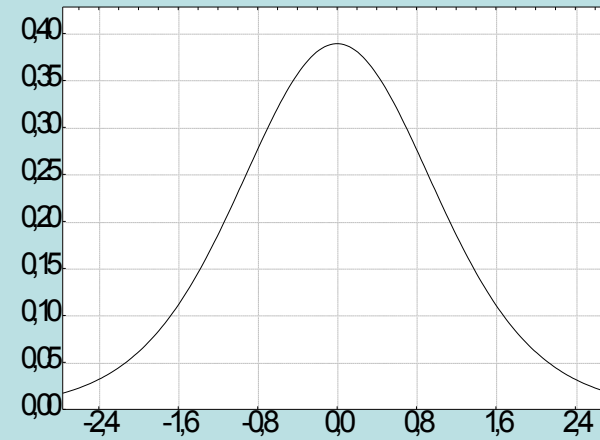
Jako odhad pravděpodobnosti chyby 1. druhu sloužila relativní četnost těch případů, kdy se na hladině významnosti 0,05 zamítla nulová hypotéza o shodě rozptylů. Simulace byly provedeny v programu MathCad.

## Grafy hustot zkoumaných rozložení

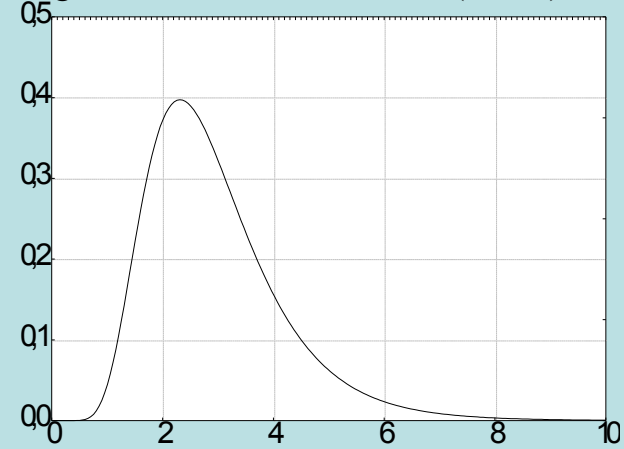
Normální rozložení  $N(10; 1)$



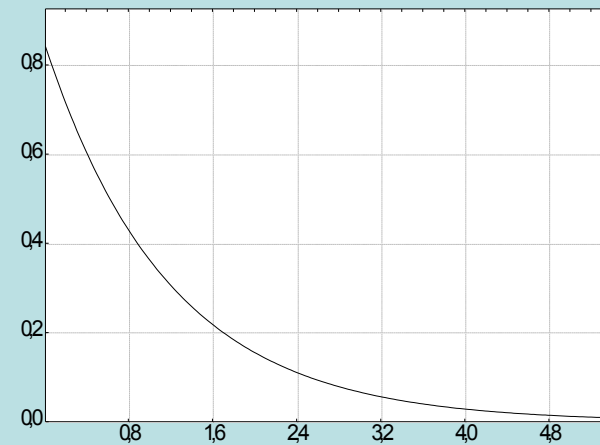
Studentovo rozložení  $t(10)$



Log – normální rozložení  $LN(1; 0,4)$



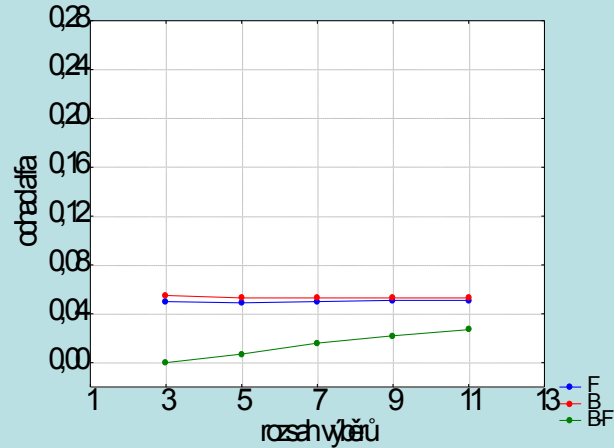
Exponenciální rozložení  $Ex(0,85)$



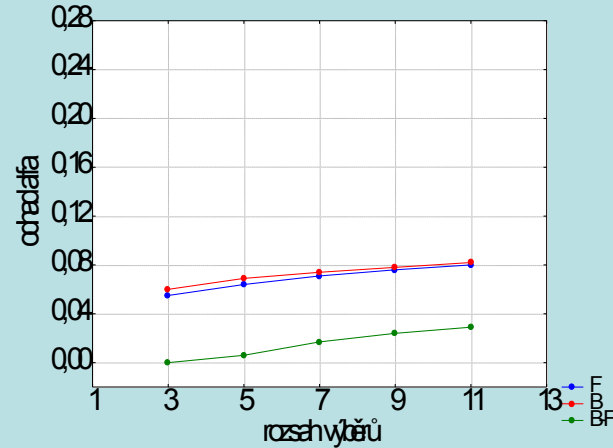
## Případ dvou nezávislých náhodných výběrů

Nejprve bylo provedeno srovnání F-testu s Bartlettovým testem a Brownovým – Forsytheovým testem pro **dva nezávislé náhodné výběry**. V grafech se modrá barva vztahuje k F-testu, červená k Bartlettovu testu a zelená k Brownovu – Forsytheovu testu.

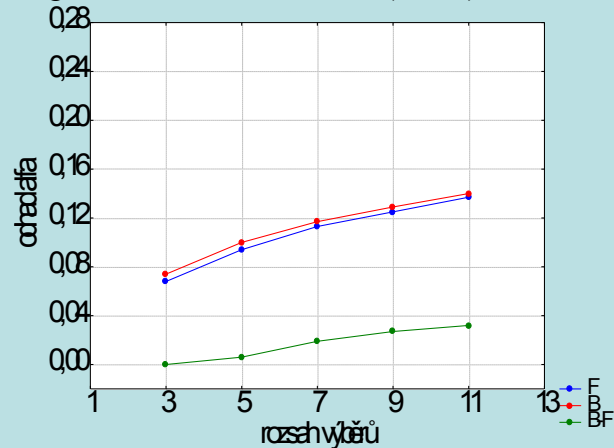
Normální rozložení  $N(10; 1)$



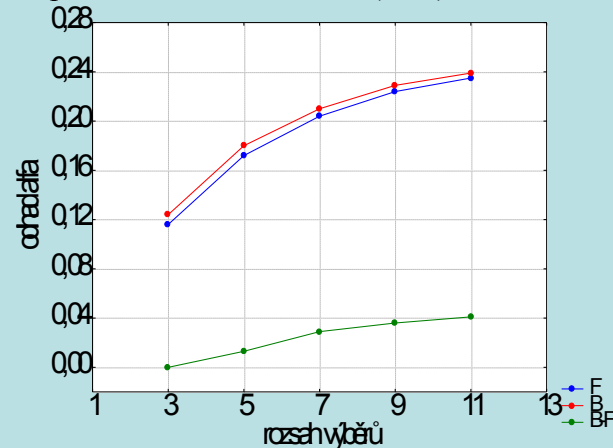
Studentovo rozložení  $t(10)$



Log - normální rozložení  $LN(1; 0,4)$



Exponenciální rozložení  $Ex(0,85)$



**Komentář:** Podle očekávání je nejnižších odhadů pravděpodobnosti chyby 1. druhu dosahováno pro výběry z normálního rozložení, kdy všechny testy udrží odhad pod hladinou významnosti 0,05. S postupným „vzdalováním se“ od normality relativní četnost neoprávněného zamítnutí nulové hypotézy roste, nejvyšší je pro výběry z exponenciálního rozložení, kde se pro F-test a Bartlettův test blíží k 0,24.

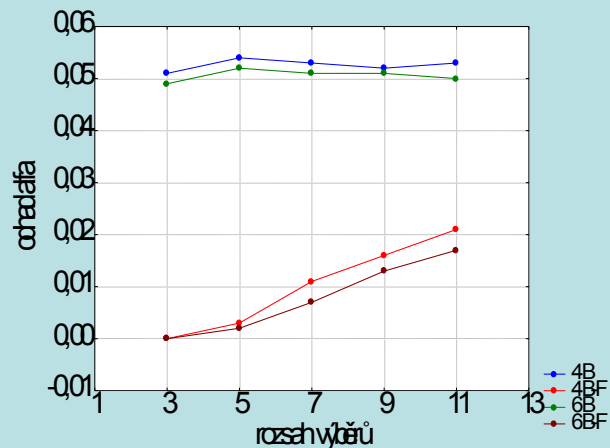
Pro všechna zkoumaná rozložení dávají F-test a Bartlettův test srovnatelné výsledky, u F-testu pozorujeme poněkud nižší odhad. Jednoznačně nejlepší výsledky jsou dosahovány při použití B-F testu, který i pro výběry z exponenciálního rozložení poskytuje odhad pravděpodobnosti chyby 1. druhu dostatečně hluboko pod 0,05.

## Případ více než dvou nezávislých náhodných výběrů

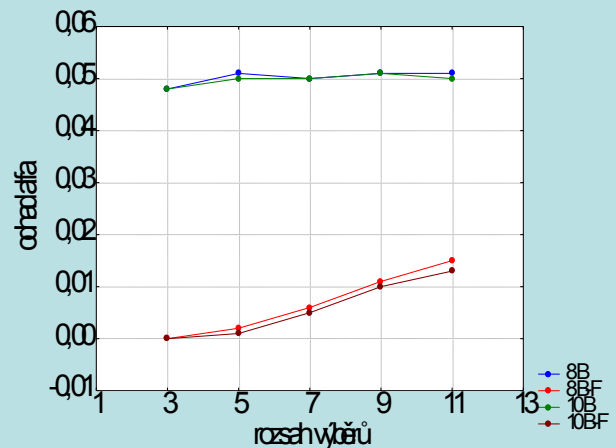
Dále jsme se zabývali srovnáním Bartlettova testu s Brownovým – Forsytheovým testem pro 4, 6, 8 a 10 nezávislých náhodných výběrů, jejichž rozsahy byly 3, 5, 7, 9, 11. Kvůli větší přehlednosti jsou grafy závislosti odhadu na rozsahu výběrů uvedeny zvlášť pro 4 a 6 výběrů a poté pro 8 a 10 výběrů. V grafech se modrá a zelená barva vztahuje k Bartlettovu testu, červená a hnědá pak k Brownovu – Forsytheovu testu.

### a) Normální rozložení $N(10; 1)$

Počet výběrů 4 a 6



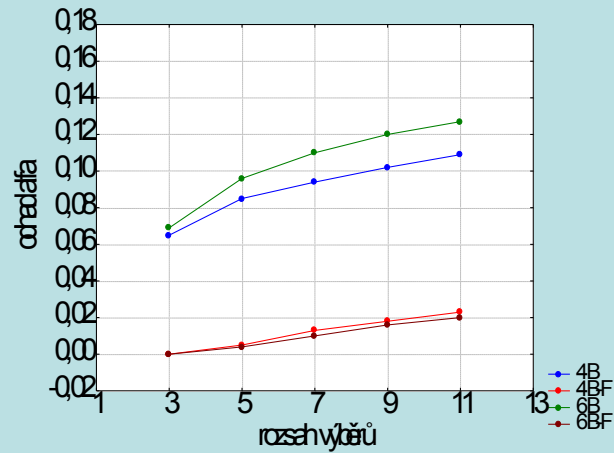
Počet výběrů 8 a 10



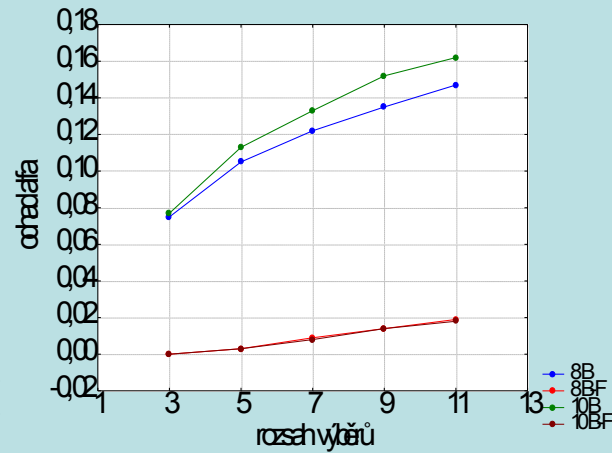
Pro výběry z normálního rozložení dává Bartlettův test odhady velmi blízké hladině významnosti 0,05. Není zde pozorovatelná závislost na rozsahu výběrů. Brownův – Forsytheův test neoprávněně zamítá nulovou hypotézu s podstatně menší relativní četností, která nepřesáhne 0,021.

## b) Studentovo rozložení t(10)

Počet výběrů 4 a 6



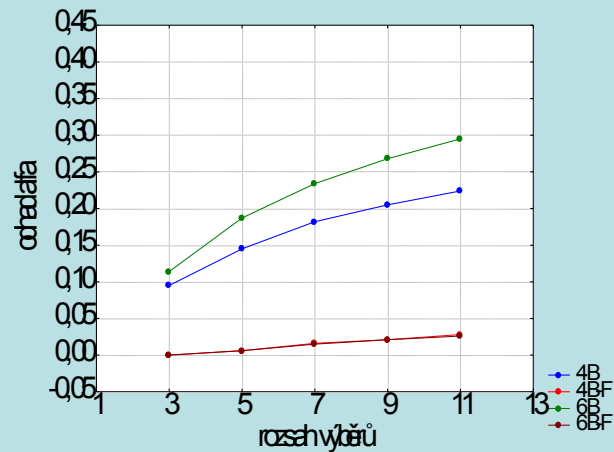
Počet výběrů 8 a 10



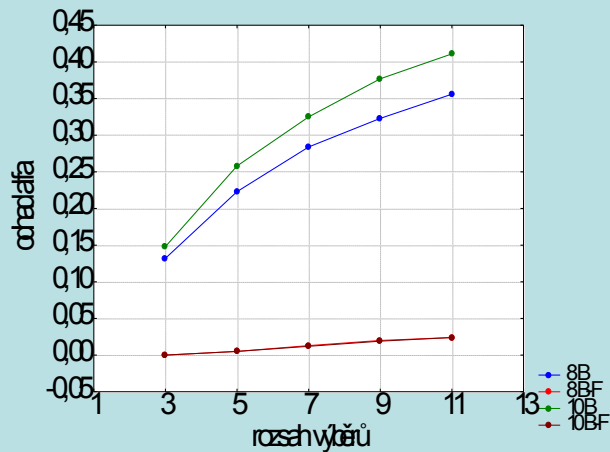
Pro výběry ze Studentova rozložení jsou výsledky Bartlettova testu již ovlivněny porušením předpokladu normality. Získané odhady narůstají se zvětšujícím se rozsahem výběrů a v nejméně příznivém případě, tj. pro 10 nezávislých náhodných výběrů o rozsahu 11, odhad pravděpodobnosti chyby 1. druhu převyšuje 0,16. Brownův – Forsytheův test neoprávněně zamítá nulovou hypotézu s relativní četností, která nepřesáhne 0,023. Rozdíly mezi odhady pro různé počty výběrů jsou u B-F testu zanedbatelně malé.

c) Logaritmicko – normální rozložení LN(1; 0,4)

Počet výběrů 4 a 6



Počet výběrů 8 a 10

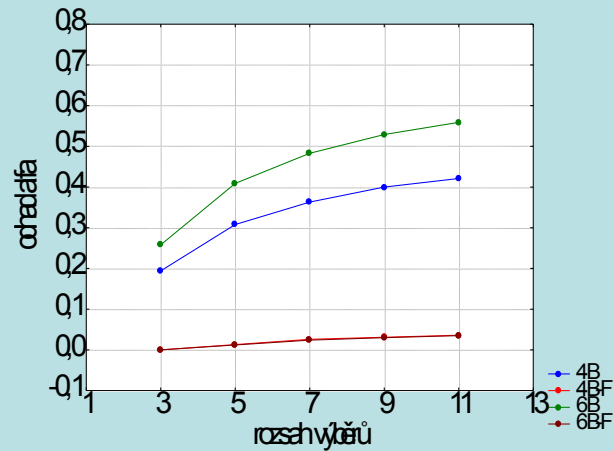


Pro výběry z logaritmicko - normálního rozložení odhad pravděpodobnosti chyby 1. druhu získaný Bartlettovým testem velmi výrazně narůstá, zvláště pro větší počet rozsáhlejších výběrů. Zde je dokonce o něco vyšší než 0,42, tudíž použití Bartlettova testu skutečně nelze doporučit. Daleko lepší výsledky poskytuje Brownův – Forsytheův test, kde odhady zůstávají pod 0,03.

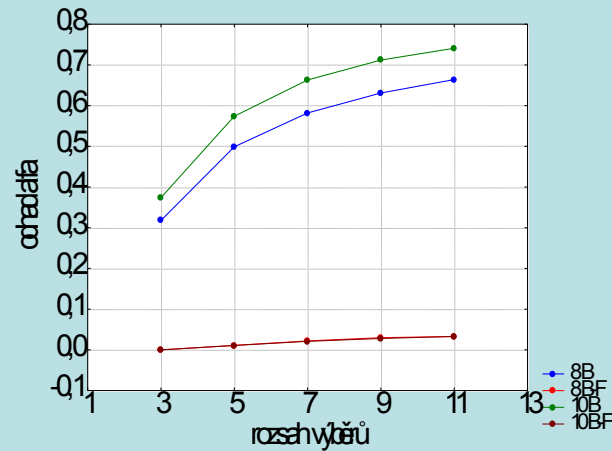


#### d) Exponenciální rozložení $Ex(0,85)$

Počet výběrů 4 a 6



Počet výběrů 8 a 10



Vidíme, že použití Bartlettova testu pro výběry z exponenciálního rozložení nelze vůbec doporučit. Odhad pravděpodobnosti chyby 1. druhu je neúnosně velký, v nejméně příznivém případě – pro 10 nezávislých náhodných výběrů o rozsahu 11 - se tento odhad blíží 0,75. Naproti tomu odhady získané Brownovým – Forsytheovým testem jsou nanejvýš 0,035, což ještě zdaleka nedosahuje hladiny významnosti 0,05.

## Komentář

Výsledky našich simulačních studií vedou k závěru, že pro testy homogenity rozptylů je vhodné používat Brownův – Forsytheův test, a to jak pro dva, tak pro více nezávislých náhodných výběrů. Ukazuje se, že tento test lze aplikovat i na výběry, které pocházejí z výrazně nenormálních rozložení. To lze vysvětlit tím, že při jeho konstrukci jsou použity výběrové mediány jednotlivých výběrů, přičemž medián – na rozdíl od průměru – je robustní vůči odlehlým či extrémním hodnotám.

U Brownova – Forsytheova testu odhad pravděpodobnosti chyby 1. druhu ve všech případech zůstal pod hladinou významnosti 0,05, nejhorší výsledek byl 0,036 pro 4 nezávislé výběry z exponenciálního rozložení. Bartlettův test zcela selhává pro výběry z nesymetrických rozložení. Např. pro 10 nezávislých výběrů z exponenciálního rozložení, jejichž rozsah byl 11, se odhad pravděpodobnosti chyby 1. druhu blížil číslu 0,8.

Výhodou Brownova – Forsytheova testu je rovněž skutečnost, že velikosti odhadů vykazují jen velmi nepatrnou závislost na počtu výběrů.

Brownův – Forsytheův test je implementován např. v systémech STATISTICA či MINITAB, Bartlettův test najdeme v systému MINITAB, F-test pak v obou zmíněných systémech.

## Post – hoc metody mnohonásobného porovnávání

Zamítne-li na hladině významnosti  $\alpha$  hypotézu o shodě středních hodnot, chceme zjistit, které dvojice středních hodnot se liší na dané hladině významnosti  $\alpha$ , tj. na hladině významnosti  $\alpha$  testujeme  $H_0: \mu_l = \mu_k$  proti  $H_1: \mu_l \neq \mu_k$  pro všechna  $l, k = 1, \dots, r, l \neq k$ .

a) Mají-li všechny výběry též rozsah  $p$  (říkáme, že třídění je vyvážené), použijeme **Tukeyovu metodu**.

Testová statistika má tvar  $\frac{|M_k - M_l|}{S \sqrt{p}}$ . Rovnost středních hodnot  $\mu_k$  a  $\mu_l$  zamítne na hladině významnosti  $\alpha$ , když

$\frac{|M_k - M_l|}{S \sqrt{p}} \geq q_{1-\alpha}(r, n-p)$ , kde hodnoty  $q_{1-\alpha}(r, n-p)$  jsou kvantily studentizovaného rozpětí a najdeme je ve statistických ta-

bulkách. (Studentizované rozpětí je náhodná veličina  $Q = \frac{X_{(r)} - X_{(1)}}{S}$ .)

Existuje modifikace Tukeyovy metody pro nestejně rozsahy výběrů, nazývá se Tukeyova HSD metoda. V tomto případě má

testová statistika tvar  $\frac{|M_k - M_l|}{S \sqrt{\frac{1}{2} \left( \frac{1}{n_k} + \frac{1}{n_l} \right)}}$ . Rovnost středních hodnot  $\mu_k$  a  $\mu_l$  zamítne na hladině významnosti  $\alpha$ , když

$\frac{|M_k - M_l|}{S \sqrt{\frac{1}{2} \left( \frac{1}{n_k} + \frac{1}{n_l} \right)}} \geq q_{1-\alpha}(r, n-p)$ .

b) Nemají-li všechny výběry stejný rozsah, použijeme **Scheffého metodu**: rovnost středních hodnot  $\mu_k$  a  $\mu_l$  zamítneme na hladině významnosti  $\alpha$ , když

$$|\bar{M}_k - \bar{M}_l| \geq \sqrt{(r-1) \left( \frac{1}{n_k} + \frac{1}{n_l} \right) F_{1-\alpha}(r-1, n-r)}$$

Výhodou Scheffého testu je, že k jeho provedení nepotřebujeme speciální statistické tabulky s hodnotami kvantilů studentizovaného rozpětí, ale stačí běžné statistické tabulky s kvantily Fisherova – Snedecorova rozložení.

V případě vyváženého třídění, kdy lze aplikovat Tukeyovu i Scheffého metodu, použijeme tu, která je citlivější. Tukeyova metoda tedy bude výhodnější, když

$$q_{1-\alpha}^2(r, n-r) < 2(r-1)F_{1-\alpha}(r-1, n-r).$$

Metody mnohonásobného porovnávání mají obecně menší sílu než ANOVA.

Může nastat situace, kdy při zamítnutí  $H_0$  nenajdeme metodami mnohonásobného porovnávání významný rozdíl u žádné dvojice středních hodnot. K tomu dochází zvláště tehdy, když p-hodnota pro ANOVU je jen o málo nižší než zvolená hladina významnosti. Pak slabší test patřící do skupiny metod mnohonásobného porovnávání nemusí odhalit žádný rozdíl.

## Doporučený postup při provádění analýzy rozptylu:

- a) Ověření normality daných  $n$  náhodných výběrů (grafické metody - NP plot, Q-Q plot, histogram, testy hypotéz o normálním rozložení - Lilieforsova varianta Kolmogorovova – Smirnovova testu nebo Shapirov – Wilkov test).  
Doporučuje se kombinace obou způsobů. Závěry učiníme až na základě posouzení obou výsledků.  
Obecně lze říci, že analýza rozptylu není příliš citlivá na porušení předpokladu normality, zvláště při větších rozsazích výběrů (nad 20), což je důsledek působení centrální limitní věty. Mírné porušení normality tedy není na závadu, při větším porušení použijeme např. Kruskalův – Wallisův test jako neparametrickou obdobu analýzy rozptylu jednoduchého třídění.
- b) Po ověření normality se testuje homogenitu rozptylů, tj. předpoklad, že všechny náhodné výběry pocházejí z normálních rozložení s tímž rozplyem. Graficky ověřujeme shodu rozptylů pomocí krabicových diagramů, kdy sledujeme, zda je šířka krabic stejná. Numericky testujeme homogenitu rozptylů pomocí Levenova testu, Brownova – Forsytheova testu (oba jsou implementovány ve STATISTICE, Brownův – Forsytheův test v MINITABu) či Bartlettova testu (je k dispozici v MINITABu).  
Slabé porušení homogenity rozptylů nevedí, při větším se doporučuje mediánový test.
- c) Pokud jsou splněny předpoklady normality a homogenity rozptylů, můžeme přistoupit k testování shody středních hodnot. Předtím je samozřejmě vhodné vypočítat průměry a směrodatné odchylky či rozptyly v jednotlivých skupinách.
- d) Dojde-li na zvolené hladině významnosti k zamítnutí hypotézy o shodě středních hodnot, zajímá nás, které dvojice středních hodnot se od sebe liší. K řešení tohoto problému slouží post-hoc metody mnohonásobného porovnávání, např. Scheffého nebo Tukeyova metoda.

**Příklad:** U čtyř odrůd brambor (označených symboly A, B, C, D) se zjišťovala celková hmotnost brambor vyrostlých vždy z jednoho trsu. Výsledky (v kg):

odrůda	hmotnost
A	0,9 0,8 0,6 0,9
B	1,3 1,0 1,3
C	1,3 1,5 1,6 1,1 1,5
D	1,1 1,2 1,0

Na hladině významnosti 0,05 testujte hypotézu, že střední hodnota hmotnosti trsu brambor nezávisí na odrůdě. Zamítnete-li nulovou hypotézu, zjistěte, které dvojice odrůd se liší na hladině významnosti 0,05.

## Řešení:

Data považujeme za realizace čtyř nezávislých náhodných výběrů ze čtyř normálních rozložení se stejným rozptylem. Testujeme hypotézu, že všechny čtyři střední hodnoty jsou stejné.

Vypočítáme **výběrové průměry v jednotlivých výběrech**:  $M_1 = 0,8$ ,  $M_2 = 1,2$ ,  $M_3 = 1,4$ ,  $M_4 = 1,1$ ,

**celkový průměr**:  $M_{..} = 1,14$ ,

**výběrové rozptyly**:  $S_1^2 = 0,02$ ,  $S_2^2 = 0,03$ ,  $S_3^2 = 0,04$ ,  $S_4^2 = 0,01$ ,

**vážený průměr výběrových rozptylů**:  $S^2 = \frac{\sum_{i=1}^4 n_i S_i^2}{n} = \frac{3 \cdot 0,02 + 3 \cdot 0,03 + 3 \cdot 0,04 + 3 \cdot 0,01}{12} = 0,027$ ,

**reziduální součet čtverců**:  $S_E = \sum_{i=1}^4 (n_i - 1) S_i^2 = 3 \cdot 0,03 = 0,3$ ,

**skupinový součet čtverců**:  $S_A = \sum_{i=1}^4 n_i (M_i - M_{..})^2 = 3 \cdot [0,8 - 1,14]^2 + 3 \cdot [1,2 - 1,14]^2 + 3 \cdot [1,4 - 1,14]^2 + 3 \cdot [1,1 - 1,14]^2 = 0,816$

**celkový součet čtverců**:  $S_T = S_A + S_E = 0,816 + 0,3 = 1,116$ ,

**testová statistika**  $F_A = \frac{S_A / f_A}{S_E / f_E} = \frac{0,816 / 3}{0,3 / 9} = 9,97$ ,

Kritický obor  $W = \langle F_{0,95; 3; 9} = 2,1; F_{0,05; 3; 9} = 3,59 \rangle$ . Protože testová statistika se realizuje v kritickém oboru,  $H_0$  zamítáme na hladině významnosti 0,05.

Vypočteme **poměr determinace**:  $P^2 = \frac{S_A}{S_T} = \frac{0,816}{1,116} = 0,731$

Výsledky zapíšeme do tabulky ANOVA:

Zdroj variability	Součet čtverců	Stupně volnosti	podíl	$F_A$
skupiny	$S_A = 0,816$	3	$S_A/3 = 0,272$	$\frac{S_A/r}{S_E/p} = 9,97$
reziduální	$S_E = 0,3$	11	$S_E/11 = 0,02727$	-
celkový	$S_T = 1,116$	14	-	-

Nyní pomocí Scheffého metody zjistíme, které dvojice odrůd se liší na hladině významnosti 0,05.

Srovnávané odrůdy	Rozdíly $ M_k - M_l $	Pravá strana vzorce
A, B	0,4	0,41
A, C	0,67	0,36
A, D	0,3	0,41
B, C	0,2	0,40
B, D	0,1	0,44
C, D	0,3	0,40

Na hladině významnosti 0,05 se liší odrůdy A a C.

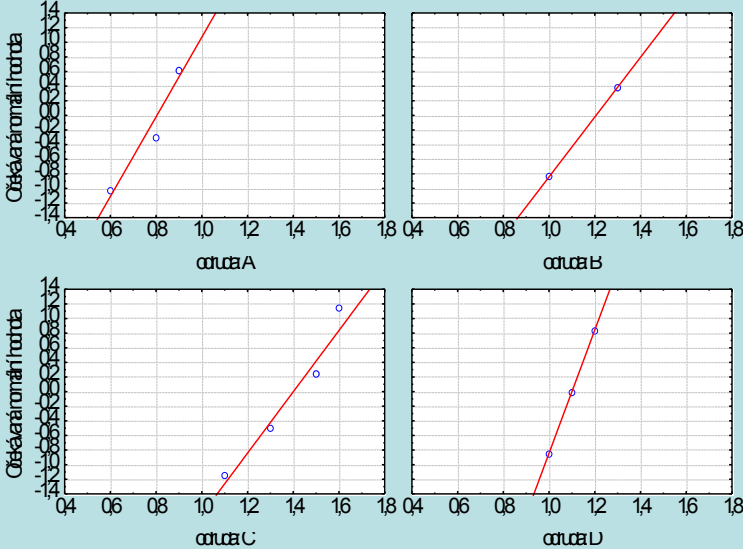


## Řešení pomocí systému STATISTICA

Otevřeme nový datový soubor o dvou proměnných X a odrůda a 15 případech. Do proměnné X zapíšeme zjištěné hmotnosti, do proměnné odrůda kódy pro dané odrůdy (1 pro A, 2 pro B, 3 pro C a 4 pro D).

	1 X	2 odruc
1	0	A
2	0	A
3	0	A
4	0	A
5	1	B
6	1	B
7	1	B
8	1	C
9	1	C
10	1	C
11	1	C
12	1	C
13	1	C
14	1	D
15	1	D

Ověříme normalitu daných čtyř náhodných výběrů pomocí N-P plotu:



Odchyly od normality jsou jen nepatrné.

Vypočteme výběrové průměry a výběrové rozptyly:

Statistiky – Základní statistiky a tabulky – Rozklad & jednofakt. ANOVA – OK – Proměnné – Závislé – X, Grupovací - odrůda – OK – Skupiny tabulek - zaškrtneme Rozptyly - Výpočet.

Rozkladová tabulka popisných s				
N=15 (V seznamu záv. prom. ne)				
odruc	X	X	X	X
	průmě	N	Sm.od	Rozpt
A	0,800	4	0,141	0,020
B	1,200	3	0,173	0,030
C	1,400	5	0,200	0,040
D	1,100	3	0,100	0,010
Vs.ski	1,140	15	0,282	0,079

Nyní ověříme předpoklad shody rozptylů.

Na záložce Skupiny tabulek zaškrtneme Levenův test – Výpočet.

Levenův test homogenity rozptylu (příklad 8301)								
Označ. efekty jsou význ. na hlad. $p < ,05000$								
Promě	SC	SV	PC	SC	SV	PC	F	p
	efekt	efekt	efekt	chyb	chyt	chyb		
X	0,018	3	0,006	0,065	1	0,005	1,047	0,410

Vidíme, že p-hodnota Levenova testu je 0,41, tedy větší než hladina významnosti 0,05. Hypotézu o shodě rozptylů nezamítáme na hladině významnosti 0,05.

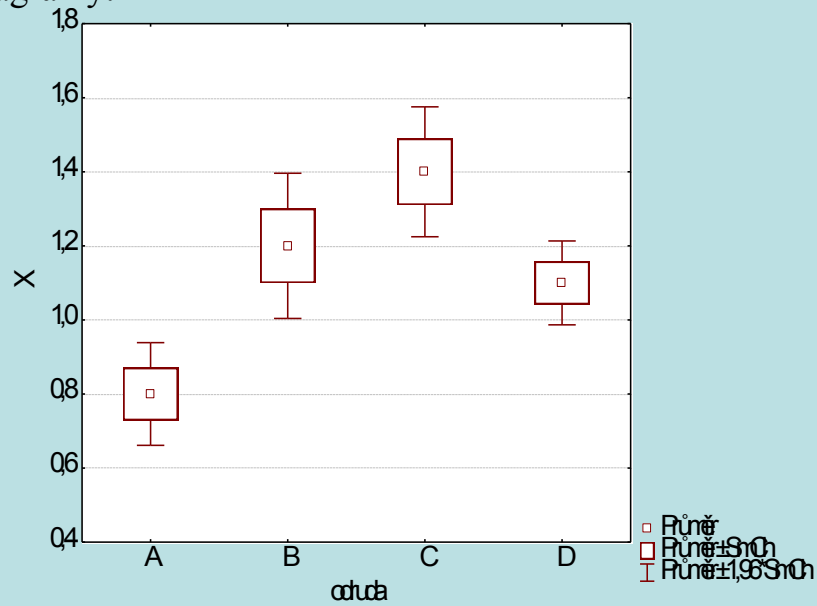
Přistoupíme k testu hypotézy o shodě středních hodnot.

Na záložce Skupiny tabulek zaškrtneme Analýza rozptylu – Výpočet.

Analýza rozptylu (příklad8301)								
Označ. efekty jsou význ. na hlad. p < ,05000								
Promě	SC	SV	PC	SC	SV	PC	F	p
	efekt	efekt	efekt	chyb	chyb	chyb		
X	0,816	0,272	0,300	1	0,027	9,973	0,001	

Jelikož p-hodnota = 0,001805 je menší než hladina významnosti 0,05, hypotézu o shodě středních hodnot zamítáme na hladině významnosti 0,05.

Výpočet doplníme krabicovými diagramy:



Nyní aplikujeme Scheffého metodu mnohonásobného porovnávání, abychom zjistili, které dvojice odrůd se liší na hladině významnosti 0,05. Na záložce Post – hoc zvolíme Schefféův test.

		Scheffého test; proměnná: X (přík)			
		Označ. rozdíly jsou významné i			
odrůda		{1}	{2}	{3}	{4}
		M=,80	M=1,2	M=1,4	M=1,1
A			0,059	0,001	0,190
B		0,059		0,464	0,905
C		0,001	0,464		0,163
D		0,190	0,905	0,163	

Tabulka obsahuje p-hodnoty pro vzájemné porovnání středních hodnot hmotnosti všech čtyř odrůd. Vidíme, že na hladině významnosti 0,05 se liší odrůdy A, C.

## Význam předpokladů v analýze rozptylu

- a) **Nezávislost jednotlivých náhodných výběrů** – velmi důležitý předpoklad, musí být splněn, jinak dostaneme nesmyslné výsledky.
- b) **Normalita** – ANOVA není příliš citlivá na porušení normality, zvláště pokud mají všechny výběry rozsah nad 20 (důsledek centrální limitní věty). Při výraznějším porušení normality se doporučuje Kruskalův – Wallisův test.
- c) **Shoda rozptylů** – mírné porušení nevádí, při větším se doporučuje Kruskalův – Wallisův test. Test shody rozptylů má smysl provádět až po ověření předpokladu normality.

## II. Příklad $r \geq 3$ nezávislých náhodných výběrů z alternativních rozložení

### Test homogenity binomických rozložení

Nechť máme  $r \geq 3$  nezávislých náhodných výběrů o rozsazích  $n_1, \dots, n_r$ , přičemž  $j$ -tý náhodný výběr pochází z alternativního rozložení  $A(p_j)$ ,  $j = 1, 2, \dots, r$ .

Testujeme hypotézu  $H_0: p_1 = \dots = p_r$  proti alternativní hypotéze  $H_1$ : aspoň jedna dvojice parametrů je různá.

Označme

$n = \sum_{j=1}^r n_j$  celkový rozsah všech  $r$  výběrů,

$M = \frac{1}{n} \sum_{j=1}^r n_j p_j$  vážený průměr výběrových průměrů.

Testové kritérium:  $Q = \frac{1}{M(1-M)} \sum_{j=1}^r \frac{n_j (p_j - M)^2}{n_j} \approx \chi^2_{r-1}$ , když  $H_0$  platí.

Kritický obor:  $W = [\chi^2_{r-1, 1-\alpha}, \infty)$

$H_0$  tedy zamítáme na asymptotické hladině významnosti  $\alpha$ , když  $Q \in W$ .

Podmínka dobré aproximace:  $n_j M^* > 5$  pro všechna  $j = 1, \dots, r$ .

Brandtův – Snedecorův výpočetní tvar:  $Q = \frac{1}{M(1-M)} \sum_{j=1}^r \frac{n_j M_j^2}{n_j} - \frac{n}{1-M}$ .



### Test homogenity založený na arkussinusové transformaci

Není-li splněna podmínka  $n_j M_* > 5$  pro všechna  $j = 1, \dots, r$ , doporučuje se následující postup: označme

$$A_j = \frac{r \cos j \pi / M}{\sqrt{M}}, j = 1, \dots, r,$$

$$B = \frac{1}{n} \sum_{j=1}^r n_j A_j.$$

Pak statistika  $Q = 4 \sum_{j=1}^r n_j A_j B^2 \approx \chi^2_{(r-1)}$ .

$H_0$  tedy zamítáme na asymptotické hladině významnosti  $\alpha$ , když  $Q \geq \chi^2_{1-\alpha}(r-1)$ .

### Mnohonásobné porovnávání

Zamítneme-li nulovou hypotézu na asymptotické hladině významnosti  $\alpha$ , chceme zjistit, které dvojice parametrů  $q_k, q_r$  se

liší. Platí-li nerovnost  $|A_k - A_r| \geq \frac{1}{8} \sqrt{\frac{1}{n_k + n_r}} q_{1-\alpha}(r, \infty)$ , pak na hladině významnosti  $\alpha$  zamítáme hypotézu o shodě parametrů  $q_k, q_r$ . (Hodnoty  $q_{1-\alpha}(r, \infty)$  najdeme v tabulkách.)

**Příklad:** Na gymnázium bylo přijato 142 studentů. Ti byli náhodně rozděleni do čtyř tříd A, B, C, D. V každé třídě byla matematika vyučována jinou metodou. Na konci školního roku psali všichni studenti stejnou písemnou práci a byl zaznamenán počet těch studentů, kteří vyřešili všechny zadané úkoly.

Třída	A	B	C	D
Počet studentů	35	36	37	34
Počet úspěšných studentů	5	8	17	15

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že rozdíly mezi třídami jsou způsobeny pouze náhodnými vlivy.

### Řešení:

Máme čtyři nezávislé náhodné výběry,  $j$ -tý pochází z rozložení  $A(n_j, p_j)$ ,  $j = 1, 2, 3, 4$ .

Testujeme hypotézu  $H_0: p_1 = p_2 = p_3 = p_4 = p_*$ .

$$n_1 = 35, n_2 = 36, n_3 = 37, n_4 = 34, n = 142$$

$$m_1 = 5/35, m_2 = 8/36, m_3 = 17/37, m_4 = 15/34, m_* = (5+8+17+15)/142 = 45/142.$$

Podmínky dobré aproximace:

$$3 \cdot \frac{45}{142} = 0,95 < 3, \quad 3 \cdot \frac{45}{142} = 0,95 < 4, \quad 3 \cdot \frac{45}{142} = 0,95 < 17, \quad 3 \cdot \frac{45}{142} = 0,95 < 37.$$

Testová statistika

$$Q = \frac{1}{n} \sum_{j=1}^k \frac{m_j^2}{n_j} - \frac{M^2}{n} = \frac{1}{142} \left[ \frac{5^2}{35} + \frac{8^2}{36} + \frac{17^2}{37} + \frac{15^2}{34} \right] - \frac{45^2}{142} = 28$$

Kritický obor:  $W = \chi_{0,95}^2, \infty, 7,81, \infty$ .

Protože testové kritérium se realizuje v kritickém oboru,  $H_0$  zamítáme na asymptotické hladině významnosti 0,05.

Nyní metodou mnohonásobného porovnávání zjistíme, které dvojice parametrů se od sebe liší na hladině významnosti 0,05.

Pomocí arkussinusové transformace vypočteme hodnoty  $A_j = \arcsin \sqrt{M_j}$ :

$$A_1 = 0,3876, \quad A_2 = 0,4909, \quad A_3 = 0,7448, \quad A_4 = 0,7264$$

Platí-li nerovnost  $|A_k - A_l| \geq \sqrt{\frac{1}{8} \left( \frac{1}{n_k} + \frac{1}{n_l} \right)} q_{1-\alpha/2, \infty}$ , pak na hladině významnosti  $\alpha$  zamítáme hypotézu o shodě parametrů  $\mu_k$  a  $\mu_l$ .

Kvantil studentizovaného rozpětí najdeme v tabulkách:  $q_{0,95}(4, \infty) = 3,63$

Srovnávané třídy	Rozdíly $ A_k - A_l $	Pravá strana vzorce
A, B	0,1033	0,30
A, C	0,3572	0,30
A, D	0,3388	0,31
B, C	0,2539	0,30
B, D	0,2356	0,31
C, D	0,0184	0,30

Na hladině významnosti 0,05 se liší třídy A, C a A, D.

## Řešení pomocí systému STATISTICA

Vytvoříme nový datový soubor se dvěma proměnnými a 142 případy.

Proměnná USPECH obsahuje hodnotu 1, pokud student vyřešil všechny zadané úkoly, jinak obsahuje hodnotu 0.

Proměnná TRIDA má hodnotu 1, pokud student pochází z třídy A, hodnotu 2 pro třídu B, hodnotu 3 pro třídu C a hodnotu 4 pro třídu D.

Nejprve zjistíme podíly úspěšných studentů v jednotlivých třídách.

Statistiky – Základní statistiky a tabulky – Rozklad – OK – Proměnné – Závislé – USPECH, Grupovací - TRIDA – OK – Skupiny tabulek - odškrtneme Směrovat. odchylka - Výpočet.

TRID	USPE Průmě	USPE N
A	0,142	39
B	0,222	31
C	0,459	31
D	0,441	31
Vs.SKI	0,316	142

Vidíme, že nejslabší výkony podávali studenti ze třídy A, úspěšných bylo pouze 14,3% studentů, ve třídě B 22,2%, ve třídě C 45,9% a ve třídě D 44,1%. Třídy C a D se z hlediska úspěchu v písemce z matematiky liší jen nepatrně

Dále provedeme testování hypotézy o shodě parametrů čtyř alternativních rozložení. Nejprve ověříme splnění podmínek dobré aproximace:  $n_j m^* > 5$  pro všechna  $j = 1, \dots, r$ . Vážený průměr  $m^*$  se nachází v posledním řádku výstupní tabulky procedury Rozklad. Jeho hodnotu okopírujeme do políček pro průměry tříd A, B, C, D, poslední řádek odstraníme a k tabulce přidáme jednu novou proměnnou, do jejíhož Dlouhého jména napíšeme  $=v2*v3$ .

TRID	JSPEJ Průmě	JSPE N	NPror =v2*v3
A	0,316	3	11,09
B	0,316	3	11,40
C	0,316	3	11,72
D	0,316	3	10,77

Vidíme, že podmínky dobré aproximace jsou splněny.

Statistiky – Základní statistiky/tabulky – Kontingenční tabulky - OK - Specif. tabulky – List 1 USPECH, List 2 TRIDA, OK– Možnosti – Statistiky dvourozměrných tabulek - zaškrtněte Pearson & M-L Chi-square – Detailní výsledky - Detailní 2-rozm. tabulky.

Statist.	Chi-kv:	sv	p
Pearsonuv c	12,28	df=	p=,00
M-V chi-kva	12,80	df=	p=,00

Testová statistika  $Q$  se realizuje hodnotou 12,2876, počet stupňů volnosti je 3, odpovídající p-hodnota = 0,00646, tedy na asymptotické hladině významnosti 0,05 hypotézu  $H_0$  zamítáme. S rizikem omylu nejvýše 0,05 jsme tedy prokázali, že rozdíly v podílech úspěšných studentů v jednotlivých třídách nelze vysvětlit náhodnými vlivy.

**Upozornění:** Systém STATISTICA neumožňuje provedení metody mnohonásobného porovnávání pro náhodné výběry z alternativního rozložení. Pro orientaci lze použít Scheffého metodu. V našem případě:

	$\{1\}$	$\{2\}$	$\{3\}$	$\{4\}$
TRIDA	$M=,14$	$M=,22$	$M=,45$	$M=,44$
A		0,907	0,034	0,060
B	0,907		0,173	0,253
C	0,034	0,173		0,998
D	0,060	0,253	0,998	

Na asymptotické hladině významnosti 0,05 se liší třídy A a C.

## Neparametrické testy o mediánech

### Osnova:

- jednovýběrové a párové testy
- dvouvýběrové testy
- neparametrické obdoby jednofaktorové analýzy rozptylu

**Motivace:** Při aplikaci t-testů či analýzy rozptylu by měly být splněny určité předpoklady:

- normalita dat (pro výběry větších rozsahů ( $n \geq 30$ ) nemá mírné porušení normality závažný dopad na výsledky)
- homogenita rozptylů
- intervalový či poměrový charakter dat

Pokud nejsou tyto předpoklady splněny, použijeme tzv. neparametrické testy, které nevyžadují předpoklad o konkrétním typu rozložení (např. normálním), stačí např. předpokládat, že distribuční funkce rozložení, z něhož náhodný výběr pochází, je spojitá.

Nevýhoda - ve srovnání s klasickými parametrickými testy jsou neparametrické testy slabší, tzn., že nepravdivou hypotézu zamítají s menší pravděpodobností než testy parametrické.

V této kapitole se omezíme na ty neparametrické testy, které se týkají mediánů.

**Jednovýběrové testy** (Jde o neparametrické obdoby jednovýběrového t-testu a párového t-testu.)

### Znaménkový test a jeho asymptotická varianta

Nechť  $X_1, \dots, X_n$  je náhodný výběr ze spojitého rozložení. Nechť  $x_{0,50}$  je mediánem tohoto rozložení a  $c$  je reálná konstanta.

Testujeme hypotézu  $H_0: x_{0,50} = c$  proti oboustranné alternativě  $H_1: x_{0,50} \neq c$  (resp. proti levostranné alternativě

$H_1: x_{0,50} < c$  resp. proti pravostranné alternativě  $H_1: x_{0,50} > c$ ).

Znaménkový test se nejčastěji používá jako párový test, kdy máme náhodný výběr ze spojitého dvourozměrného rozložení

$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \dots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$  a testujeme hypotézu o rozdílu mediánů, tj.  $H_0: x_{0,50} - y_{0,50} = 0$  proti  $H_1: x_{0,50} - y_{0,50} \neq 0$  (resp. proti jednostranným alternativám).

Přejdeme k rozdílům  $Z_1 = X_1 - Y_1, \dots, Z_n = X_n - Y_n$  a testujeme hypotézu o mediánu těchto rozdílů, tj.

$H_0: z_{0,50} = 0$ .

a) Utvoříme rozdíly  $D_1, \dots, D_n$  pro jednovýběrový test resp.  $D_1, \dots, D_n$  pro párový test,  $1 \leq i \leq n$ . (Jsou-li některé rozdíly nulové, pak za  $n$  bereme jen počet nenulových hodnot.)

b) Zavedeme statistiku  $S_{Z^+}$ , která udává počet těch rozdílů  $D_i$ , které jsou kladné.  $S_{Z^+}$  je součtem náhodných veličin s alternativním rozložením ( $i$ -tá veličina nabývá hodnoty 1, když  $i$ -tý rozdíl je kladný a hodnoty 0, když je záporný). Platí-li  $H_0$ , pak pravděpodobnost kladného i záporného rozdílu je stejná, tedy  $S_{Z^+} \sim B(n, \frac{1}{2})$ . Z vlastností binomického rozložení plyne, že  $ES_{Z^+} = \frac{n}{2}$ ,  $DS_{Z^+} = \frac{n}{4}$ .

c) Stanovíme kritický obor.

Pro oboustrannou alternativu:  $W = \{k_1, \dots, k_2\}$ , pro levostrannou alternativu:  $W = \{k_1\}$ , pro pravostrannou alternativu:  $W = \{k_2, n\}$ .

(Nezáporná celá čísla  $k_1, k_2$  pro oboustranný test i pro jednostranné testy lze najít v tabulkové příloze. Pozor – čísla  $k_1, k_2$  pro oboustrannou alternativu jsou jiná než pro jednostranné alternativy!)

d)  $H_0$  zamítáme na hladině významnosti  $\alpha$  když  $S_{Z^+} \in W$ .



## Asymptotická varianta testu

Pro velká  $n$  (prakticky  $n \geq 10$ ) lze využít asymptotické normality statistiky  $S_{Z^+}$ .

Testová statistika  $U_0 = \frac{S_{Z^+} - \frac{n}{2}}{\sqrt{np_{Z^+}}} = \frac{S_{Z^+} - \frac{n}{2}}{\sqrt{n/4}}$  má za platnosti  $H_0$  asymptoticky rozložení  $N(0,1)$ .

Kritický obor pro oboustranný test:  $W = (-\infty, -u_{1-\alpha/2}) \cup (u_{1-\alpha/2}, \infty)$ .

Kritický obor pro levostranný test:  $W = (-\infty, u_{1-\alpha})$ .

Kritický obor pro pravostranný test:  $W = (u_{1-\alpha}, \infty)$ .

Aproximace rozložením  $N(0,1)$  se zlepší, když použijeme tzv. **korekci na nespojitost**. Testová statistika pak má

tvar  $U_0 = \frac{S_{Z^+} - \frac{n}{2} + \frac{1}{2}}{\sqrt{n/4}}$ , přičemž  $\frac{1}{2}$  přičteme, když  $S_{Z^+} < \frac{n}{2}$  a odečteme v opačném případě.

### Příklad na jednovýběrový znaménkový test:

U 10 náhodně vybraných vzorků benzínu byly zjištěny následující hodnoty oktanového čísla:

98,2 96,8 96,3 99,8 96,9 98,6 95,6 97,1 97,7 98,0.

Na hladině významnosti 0,05 testujte hypotézu, že medián oktanového čísla je 98 proti oboustranné alternativě.

### Řešení:

rozdíly  $x_i - 98$ : 0,2 -1,2 -1,7 1,8 -1,1 0,6 -2,4 -0,9 -0,3 0,0

$S_Z^+ = 3$ , nenulových rozdílů je 9. Ve statistických tabulkách najdeme pro  $n = 9$  a  $\alpha = 0,05$  kritické hodnoty  $k_1 = 1$ ,  $k_2 = 8$ .

Protože kritický obor  $W = \{1, \dots, 8\}$  neobsahuje hodnotu 3, nemůžeme  $H_0$  zamítnout na hladině významnosti 0,05.

### Výpočet pomocí systému STATISTICA:

Vytvoříme nový datový soubor se dvěma proměnnými a 10 případy. Do proměnné X napíšeme hodnoty oktanového čísla a do proměnné konst uložíme číslo 98.

Statistiky – Neparametrická statistika – Porovnání dvou závislých vzorků – OK – 1. seznam proměnných X, 2. seznam proměnných konst – OK – Znaménkový test.

		Znaménkový test (oktanové číslo)		
		Označené testy jsou významné n		
Dvojice prom	Počet proce různý	$v < v$	Z	Uroveň
X & konst	9	66,66	0,666	0,504

Vidíme, že nenulových hodnot  $n = 9$ . Z nich záporných je 66,7%, tj. 6. Hodnota testové statistiky  $S_Z^+ = 9 - 6 = 3$ . Asymptotická testová statistika  $U_0$  (zde označená jako Z) se realizuje hodnotou 0,6667. Odpovídající asymptotická p-hodnota je 0,505, tedy na asymptotické hladině významnosti 0,05 nezamítáme hypotézu, že medián oktanového čísla je 98.

**Upozornění:** V tomto případě není splněna podmínka pro využití asymptotické normality statistiky  $S_Z^+$ , tj.  $n > 20$ . Je tedy vhodnější najít v tabulkách kritické hodnoty pro znaménkový test. Pro  $n = 9$  a  $\alpha = 0,05$  jsou kritické hodnoty  $k_1 = 1$ ,  $k_2 = 8$ . Protože kritický obor  $W = \{1, \dots, 8\}$  neobsahuje hodnotu 3, nezamítáme  $H_0$  na hladině významnosti 0,05. Dostáváme týž výsledek jako při použití asymptotického testu.

### Příklad na párový znaménkový test

U 9 náhodně vybraných manželských párů byl zjištěn průměrný roční příjem (v tisících Kč).

číslo páru	1	2	3	4	5	6	7	8	9
příjem manžela	216	336	384	432	456	528	552	600	1872
příjem manželky	336	240	192	336	384	288	960	312	576

Na hladině významnosti 0,05 testujte hypotézu, že mediány příjmů manželů a manželek jsou stejné.

#### Řešení:

Jedná se o párový test. Vypočteme rozdíly mezi příjmy manželů a manželek, čímž úlohu převedeme na jednovýběrový test. Testujeme  $H_0: z_{0,50} =$  proti oboustranné alternativě  $H_1: z_{0,50} \neq$ , kde  $z_{0,50}$  je medián rozložení, z něhož pochází rozdílový náhodný výběr  $Z_1, \dots, Z_9$ .

Vypočtené rozdíly  $X_i$ : -120 96 192 96 72 240 -408 288 1296

Testová statistika  $S_{Z^+} = 7$ .

Ve statistických tabulkách najdeme pro  $n =$  a  $\alpha =$  kritické hodnoty  $k_1 =$ ,  $k_2 =$ .

Protože kritický obor  $W = \{ \dots \}$  neobsahuje hodnotu 7, nemůžeme  $H_0$  zamítnout na hladině významnosti 0,05.

Neprokázaly se tedy významné rozdíly v mediánech příjmů manželů a manželek.

### Výpočet pomocí systému STATISTICA:

Vytvoříme nový datový soubor se dvěma proměnnými a 9 případy. Do proměnné X napíšeme příjmy manželů, do proměnné Y příjmy manželek.

Statistiky – Neparametrická statistika – Porovnání dvou závislých vzorků – OK – 1. seznam proměnných X, 2. seznam proměnných Y – OK – Znaménkový test.

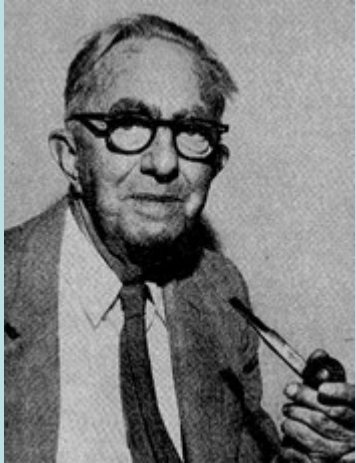
Dvojice prom.	Počet různý	procento $v < v$	Z	Uroveň
X & Y	9	22,22	1,333	0,182

Vidíme, že nenulových hodnot  $n = 9$ . Z nich záporných je 22%, tj. 2. Hodnota testové statistiky  $S_Z = \dots$ .

Asymptotická testová statistika  $U_0$  (zde označená jako Z) se realizuje hodnotou 1,3. Odpovídající asymptotická p-hodnota je 0,1824, tedy na asymptotické hladině významnosti 0,05 nezamítáme hypotézu, že mediány příjmů manželů a manželek jsou stejné.

**Upozornění:** V tomto případě není splněna podmínka pro využití asymptotické normality statistiky  $S_Z$ , tj.  $n \geq 10$ . Je tedy vhodnější najít v tabulkách kritické hodnoty pro znaménkový test. Pro  $n = 9$  a  $\alpha = 0,05$  jsou kritické hodnoty  $k_1 = 1$ ,  $k_2 = 8$ . Protože kritický obor  $W = \{1, \dots, 8\}$  neobsahuje hodnotu 7, nezamítáme  $H_0$  na hladině významnosti 0,05. Dostáváme týž výsledek jako při použití asymptotického testu.

## Jednovýběrový Wilcoxonův test a jeho asymptotická varianta



Frank Wilcoxon (1892 – 1965): Americký statistik a chemik

Nechť  $X_1, \dots, X_n$  je náhodný výběr ze spojitého rozložení s hustotou  $\varphi(x)$ , která je symetrická kolem mediánu  $x_{0,50}$ , tj.

$\varphi(x_{0,50} + x) = \varphi(x_{0,50} - x)$ . Nechť  $c$  je reálná konstanta.

Testujeme hypotézu  $H_0: x_{0,50} = c$

proti oboustranné alternativě  $H_1: x_{0,50} \neq c$  nebo

proti levostranné alternativě  $H_1: x_{0,50} < c$  nebo

proti pravostranné alternativě  $H_1: x_{0,50} > c$ .

### Postup provedení testu:

a) Utvoříme rozdíly  $D_i = X_i - c$ ,  $i = 1, \dots, n$ . (Jsou-li některé rozdíly nulové, pak za  $n$  bereme jen počet nenulových hodnot.)

b) Absolutní hodnoty  $|D_i|$  uspořádáme vzestupně podle velikosti a spočteme pořadí  $R_i$ .

c) Zavedeme statistiky

$S_W^+ = \sum_{D_i > 0} R_i^+$ , což je součet pořadí přes kladné hodnoty  $D_i$ ,

$S_W^- = \sum_{D_i < 0} R_i^-$ , což je součet pořadí přes záporné hodnoty  $D_i$ .

Přitom platí, že součet  $S_W^+ + S_W^- = n(n+1)/2$ .

Je-li  $H_0$  pravdivá, pak  $E(S_W^+) = n(n+1)/4$  a  $D(S_W^+) = n(n+1)(2n+1)/24$ .

d) Testová statistika =  $\min(S_W^+, S_W^-)$  pro oboustrannou alternativu,  
=  $S_W^+$  pro levostrannou alternativu,  
=  $S_W^-$  pro pravostrannou alternativu.

e)  $H_0$  zamítáme na hladině významnosti  $\alpha$ , když testová statistika je menší nebo rovna tabelované kritické hodnotě.

### Asymptotická varianta jednovýběrového Wilcoxonova testu:

Pro  $n \geq 30$  lze využít asymptotické normality statistiky  $S_w^+$ .

Platí-li  $H_0$ , pak  $U_0 = \frac{S_w^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(n+2)}{24}}} \approx N(0,1)$ .

Kritický obor:

pro oboustrannou alternativu  $W = (-\infty, u_{1-\alpha/2}) \cup (u_{\alpha/2}, \infty)$ ,

pro levostrannou alternativu  $W = (-\infty, u_{1-\alpha})$ ,

pro pravostrannou alternativu  $W = (u_{\alpha}, \infty)$ .

$H_0$  zamítáme na asymptotické hladině významnosti  $\alpha$ , když  $U_0 \in W$ .

### Předpoklady použití jednovýběrového Wilcoxonova testu:

- rozložení, z něhož daný náhodný výběr pochází, je spojité
- hustota tohoto rozložení je symetrická kolem mediánu
- sledovaná veličina  $X$  má aspoň ordinální charakter

(Není-li splněn předpoklad o symetrii hustoty kolem mediánu, lze použít např. znaménkový test.)



**Příklad:** U 12 náhodně vybraných zemí bylo zjištěno procento populace starší 60 let:

4,9 6,0 6,9 17,6 4,5 12,3 5,7 5,3 9,6 13,5 15,7 7,7.

Na hladině významnosti 0,05 testujte hypotézu, že medián procenta populace starší 60 let je 12 proti oboustranné alternativě.

**Řešení:**

Testujeme hypotézu  $H_0: x_{0,50} = 12$  proti oboustranné alternativě  $H_1: x_{0,50} \neq 12$ .

Vypočteme rozdíly pozorovaných hodnot od čísla 12: -7,1 -6,0 -5,1 5,6 -7,5 0,3 -6,3 -6,7 -2,4 1,5 3,7 -4,3.

Absolutní hodnoty těchto rozdílů uspořádáme vzestupně podle velikosti. Kladné rozdíly přitom označíme červeně:

usp.   $x_i - 12$	0,3	1,5	2,4	3,7	4,3	5,1	5,6	6	6,3	6,7	7,1	7,5
pořadí	1	2	3	4	5	6	7	8	9	10	11	12

$$S_w^+ = 1 + 2 + 4 + 7 = 14,$$

$$S_w^- = 3 + 5 + 6 + 8 + 9 + 10 + 11 + 12 = 64,$$

$n = 12$ ,  $\alpha = 0,05$ , tabelovaná kritická hodnota pro  $n = 12$  a  $\alpha = 0,05$  je 13,

testová statistika =  $\min(S_w^+, S_w^-) = \min(14, 64) = 14$ .

Protože  $14 > 13$ ,  $H_0$  nezamítáme na hladině významnosti 0,05. Znamená to, že na hladině významnosti 0,05 se nepodařilo prokázat, že aspoň v polovině zemí by se podíl populace nad 60 let odlišoval od 12 %.

## Výpočet pomocí systému STATISTICA:

Utvoříme nový datový soubor se dvěma proměnnými a 12 případy. Do proměnné procento napíšeme zjištěné hodnoty a do proměnné konst uložíme číslo 12.

Statistiky – Neparametrická statistika – Porovnání dvou závislých vzorků – OK – 1. seznam proměnných rozdíl, Druhý seznam proměnných konst – OK – Wilcoxonův párový test.

Wilcoxonův párový test (populace)				
Označené testy jsou významné				
Dvojice prom	Počet platný	T	Z	Uroveň
procento & k	12	14,00	1,961	0,049

Výstupní tabulka poskytne hodnotu testové statistiky  $SW^+$  (zde označena T), hodnotu asymptotické testové statistiky  $U_0$  a p-hodnotu pro  $U_0$ . V tomto případě je p-hodnota 0,049861, tedy nulová hypotéza se zamítá na asymptotické hladině významnosti 0,05. Tento výsledek je v rozporu s výsledkem, ke kterému jsme dospěli při přesném výpočtu. Je to způsobeno tím, že není splněna podmínka pro využití asymptotické normality statistiky  $SW^+$ , tj.  $n \geq 30$ .

## Párový Wilcoxonův test

Nechť  $(X_1, Y_1), \dots, (X_n, Y_n)$  je náhodný výběr ze spojitého dvourozměrného rozložení.

Testujeme  $H_0: x_{0,50} - y_{0,50} = c$  proti  $H_1: x_{0,50} - y_{0,50} \neq c$  (resp. proti jednostranným alternativám).

Utvoříme rozdíly  $Z_i = X_i - Y_i$ ,  $i = 1, \dots, n$  a testujeme hypotézu o mediánu  $z_{0,50}$ , tj.  $H_0: z_{0,50} = c$  proti  $H_1: z_{0,50} \neq c$ .

**Příklad:** K zjištění cenových rozdílů mezi určitými dvěma druhy zboží bylo náhodně vybráno 15 prodejen a byly zjištěny ceny zboží A a ceny zboží B: (11,10), (14,11), (11,9), (13,9), (11,9), (10,9), (12,10), (10,8), (12,11), (11,9), (13,10), (14,10), (14,12), (19,15), (14,12). Na hladině významnosti 0,05 je třeba testovat hypotézu, že medián cenových rozdílů činí 3 Kč.

**Řešení:** Testujeme  $H_0: z_{0,50} = 3$  proti oboustranné alternativě  $H_1: z_{0,50} \neq 3$ , kde  $z_{0,50}$  je medián rozložení, z něhož pochází rozdílový náhodný výběr  $Z_1 = X_1 - Y_1, \dots, Z_{15} = X_{15} - Y_{15}$ . Vypočteme rozdíly mezi cenou zboží A a cenou zboží B, čímž úlohu převedeme na jednovýběrový test. Výpočty uspořádáme do tabulky:

č. prodejny	cena zboží A	cena zboží B	rozdíl	rozdíl-medián	pořadí
1	11	10	1	2	12
2	14	11	3	0	-
3	11	9	2	1	5,5
4	13	9	4	1	<b>5,5</b>
5	11	9	2	1	5,5
6	10	9	1	2	12
7	12	10	2	1	5,5
8	10	8	2	1	5,5
9	12	11	1	2	12
10	11	9	2	1	5,5
11	13	10	3	0	-
12	14	10	4	1	<b>5,5</b>
13	14	12	2	1	5,5
14	19	15	4	1	<b>5,5</b>
15	14	12	2	1	5,5

(Tučně jsou vytištěna pořadí pro kladné hodnoty rozdíl - medián.)

$$S_w^+ = 5,5 + 5,5 + 5,5 = 16,5,$$

$$S_w^- = 12 + 5,5 + 5,5 + 12 + 5,5 + 5,5 + 12 + 5,5 + 5,5 + 5,5 = 74,5,$$

$n = 13$ ,  $\alpha = 0,05$ , tabelovaná kritická hodnota = 17, testová statistika =  $\min(S_w^+, S_w^-) = \min(16,5; 74,5) = 16,5$ . Protože  $16,5 \leq 17$ ,  $H_0$  zamítáme na hladině významnosti 0,05.

### Výpočet pomocí systému STATISTICA:

Vytvoříme nový datový soubor se čtyřmi proměnnými A, B, rozdíl, konst a 15 případy. Do proměnných A, B napíšeme ceny zboží A a B, do proměnné rozdíl uložíme rozdíl cen A a B a do proměnné konst uložíme číslo 3.

Statistiky – Neparametrická statistika – Porovnání dvou závislých vzorků – OK – 1. seznam proměnných rozdíl, 2. seznam proměnných konst – OK – Wilcoxonův párový test.

		Wilcoxonův párový test, (ceny zboží)		
		Označené testy jsou významné		
Dvojice proměnných	Počet platných	T	Z	Uroveň významnosti
rozdíl & konst	15	16,50	2,026	0,042

Testová statistika (zde označená jako T) nabývá hodnoty 16,5, asymptotická testová statistika (označená jako Z) nabývá hodnoty 2,026684, odpovídající asymptotická p-hodnota je 0,042696, tedy na asymptotické hladině významnosti 0,05 nulovou hypotézu zamítáme.

### Příklad (na asymptotickou variantu Wilcoxonova testu):

30 náhodně vybraných osob mělo nezávisle na sobě bez předchozího nácviku odhadnout, kdy od daného signálu uplyne právě 1 minuta. Byly získány následující výsledky (v sekundách):

53 48 45 55 63 51 66 56 50 58 61 51 64 63 59 47 46 58 52 56 61 57 48 62 54 49 51 46 53 58.

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že medián rozložení, z něhož daný náhodný výběr pochází, je 60 sekund proti oboustranné alternativě (nulová hypotéza vlastně tvrdí, že polovina osob délku jedné minuty podhodnotí a druhá nadhodnotí).

### Řešení:

Testujeme  $H_0: x_{0,50} = 60$  proti oboustranné alternativě  $H_1: x_{0,50} \neq 60$ .

Obvyklým způsobem stanovíme statistiku  $S_w^+ = 55$ .

Asymptotická testová statistika:

$$U_0 = \frac{S_w^+ - \frac{N+1}{2}}{\sqrt{\frac{N+1}{12}}} = \frac{55 - \frac{31}{2}}{\sqrt{\frac{31}{12}}} = \frac{55 - 15,5}{\sqrt{2,5833}} = \frac{39,5}{1,607} = 24,6$$

Kritický obor:

$$W = (-\infty, u_{1/2}) \cup (u_{1/2}, \infty) = (-\infty, u_{0,975}) \cup (u_{0,975}, \infty) = (-\infty, 196) \cup (196, \infty)$$

Testová statistika se realizuje v kritickém oboru, tedy  $H_0$  zamítáme na asymptotické hladině významnosti 0,05. S rizikem omylu nejvýše 5% jsme tedy prokázali, že pravděpodobnost nadhodnocení jedné minuty není stejná jako pravděpodobnost podhodnocení.

### Výpočet pomocí systému STATISTICA:

Utvoříme nový datový soubor se dvěma proměnnými a 30 případy. Do proměnné odhad napíšeme zjištěné hodnoty a do proměnné konst uložíme číslo 60.

Statistiky – Neparametrická statistika – Porovnání dvou závislých vzorků – OK – 1. seznam proměnných odhad, 2. seznam proměnných konst – OK – Wilcoxonův párový test.

		Wilcoxonův párový test, (odhad n Označené testy jsou významné n			
Dvojice prom odhad konst		Poče platný	T	Z	Urove
		3	55,00	3,650	0,000

Testová statistika (zde označená jako T) nabývá hodnoty 55, asymptotická testová statistika (označená jako Z) nabývá hodnoty 3,65088, odpovídající asymptotická p-hodnota je 0,000261, tedy na asymptotické hladině významnosti 0,05 nulovou hypotézu zamítáme.

**Dvouvýběrové testy** (Jedná se o neparametrickou obdobu dvouvýběrového t-testu)

### **Dvouvýběrový Wilcoxonův test a jeho asymptotická varianta**

Nechť  $X_1, \dots, X_n$  a  $Y_1, \dots, Y_m$  jsou dva nezávislé náhodné výběry ze dvou spojitých rozložení, jejichž distribuční funkce se mohou lišit pouze posunutím. Označme  $x_{0,50}$  medián prvního rozložení a  $y_{0,50}$  medián druhého rozložení. Na hladině významnosti 0,05 testujeme hypotézu, že distribuční funkce těchto rozložení jsou shodné neboli mediány jsou shodné proti alternativě, že jsou rozdílné, tj.

$H_0: x_{0,50} - y_{0,50} = 0$  proti  $H_1: x_{0,50} - y_{0,50} \neq 0$ .

#### **Postup provedení testu:**

- Všech  $n + m$  hodnot  $X_1, \dots, X_n$  a  $Y_1, \dots, Y_m$  uspořádáme vzestupně podle velikosti.
- Zjistíme součet pořadí hodnot  $X_1, \dots, X_n$  a označíme ho  $T_1$ .  
Součet pořadí hodnot  $Y_1, \dots, Y_m$  označíme  $T_2$ .
- Vypočteme statistiky  $U_1 = mn + n(n+1)/2 - T_1$ ,  $U_2 = mn + m(m+1)/2 - T_2$ .  
Přitom platí  $U_1 + U_2 = mn$ .
- Pokud  $\min(U_1, U_2) \leq$  tabelovaná kritická hodnota (pro dané rozsahy výběrů  $m$ ,  $n$  a dané  $\alpha$ ), pak nulovou hypotézu o totožnosti obou distribučních funkcí zamítáme na hladině významnosti  $\alpha$ . V tabulkách:  $n = \min\{m, n\}$  a  $m = \max\{m, n\}$ .



### Asymptotická varianta dvouvýběrového Wilcoxonova testu:

Pro velká  $n, m$  ( $n, m > 30$ ) lze využít asymptotické normality statistiky  $U_1$ .

Platí-li  $H_0$ , pak  $U_0 = \frac{U_1 - U_2}{\sqrt{\frac{U_1 + U_2}{n+m}}}$   $\approx N(0,1)$ , kde  $U_1 = \min(U_1, U_2)$ .

Kritický obor:

pro oboustrannou alternativu  $W = (-\infty, u_{1-\alpha/2}] \cup [u_{\alpha/2}, \infty)$ ,

pro levostrannou alternativu  $W = (-\infty, u_{1-\alpha})$ ,

pro pravostrannou alternativu  $W = [u_{\alpha}, \infty)$ .

$H_0$  zamítáme na asymptotické hladině významnosti  $\alpha$ , když  $U_0 \in W$ .

### Předpoklady použití dvouvýběrového Wilcoxonova testu:

- dané dva náhodné výběry jsou nezávislé
- rozložení, z nichž dané dva náhodné výběry pocházejí, jsou spojitá
- distribuční funkce těchto rozložení se mohou lišit pouze posunutím
- sledovaná veličina má aspoň ordinální charakter

(Není-li splněn předpoklad, že distribuční funkce se mohou lišit pouze posunutím, lze použít např. dvouvýběrový Kolmogorovův – Smirnovův test.)

### Příklad:

Bylo vybráno 10 polí stejné kvality. Na čtyřech z nich se zkoušel nový způsob hnojení, zbylých šest bylo ošetřeno starým způsobem. Pole byla oseta pšenicí a sledoval se její hektarový výnos. Je třeba zjistit, zda nový způsob hnojení má týž vliv na průměrné hektarové výnosy pšenice jako starý způsob hnojení.

hektarové výnosy při novém způsobu: 51 52 49 55

hektarové výnosy při starém způsobu: 45 54 48 44 53 50

Test proveďte na hladině významnosti 0,05.

### Řešení:

Na hladině významnosti 0,05 testujeme  $H_0: x_{0,50} - y_{0,50} = 0$  proti oboustranné alternativě  $H_1: x_{0,50} - y_{0,50} \neq 0$ .

usp. hodnoty	44	45	48	<b>49</b>	50	<b>51</b>	<b>52</b>	53	54	<b>55</b>
pořadí x-ových hodnot				4		6	7			10
pořadí y-ových hodnot	1	2	3		5			8	9	

$$T_1 = 4 + 6 + 7 + 10 = 27, T_2 = 1 + 2 + 3 + 5 + 8 + 9 = 28$$

$$U_1 = 4.6 + 4.5/2 - 27 = 7, U_2 = 4.6 + 6.7/2 - 28 = 17$$

Kritická hodnota pro  $\alpha = 0,05$ ,  $\min(4,6) = 4$ ,  $\max(4,6) = 6$  je 2. Protože  $\min(7,17) = 7 > 2$ , nemůžeme na hladině významnosti 0,05 zamítnout hypotézu, že nový způsob hnojení má na hektarové výnosy pšenice stejný vliv jako starý způsob.

## Výpočet pomocí systému STATISTICA:

Utvoříme nový datový soubor se dvěma proměnnými a 10 případy. Do proměnné vynos napíšeme zjištěné hodnoty a do proměnné hnojeni napíšeme 4x číslo 1 pro nový způsob hnojení a 6x číslo 2 pro starý způsob hnojení.

Statistiky – Neparametrická statistika – Porovnání dvou nezávislých vzorků – OK – Proměnné – Seznam závislých proměnných vynos, Nezáv. (grupov.) proměnná hnojeni – OK – M-W U test.

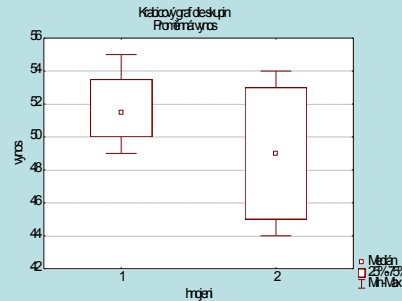
**Upozornění:** Ve STATISTICE je dvouvýběrový Wilcoxonův test uveden pod názvem Mannův – Whitneyův test.

Mann-Whitneyův U test (vynos)										
Dle proměnné hnojeni										
Označené testy jsou významné na hladině $p < ,05000$										
Promě	Sčt p	Sčt p	U	Z	Urove	Z	Urove	N pla	N pla	2*1st
vynos	skup.	skup.			uprave			skup.	skup.	přesné
	27,00	28,00	7,000	1,066	0,286	1,066	0,286	4	6	0,352

Ve výstupní tabulce jsou součty pořadí  $T_1$ ,  $T_2$ , hodnota testové statistiky

$\min(U_1, U_2)$  označená U, hodnota asymptotické testové statistiky  $U_0$  (označená Z), asymptotická p-hodnota pro  $U_0$  a přesná p-hodnota (ozn. 2\*1 str. přesné p – ta se používá pro rozsahy výběrů pod 30). V našem případě přesná p-hodnota = 0,352381, tedy  $H_0$  nezamítáme na hladině významnosti 0,05.

Výpočet je vhodné doplnit krabicovým diagramem.



Je zřejmé, že výnosy při novém způsobu hnojení jsou vesměs nižší než při starém způsobu a také vykazují mnohem větší variabilitu.

## Dvouvýběrový Kolmogorovův - Smirnovův test

Nechť  $X_1, \dots, X_n$  a  $Y_1, \dots, Y_m$  jsou dva nezávislé náhodné výběry ze dvou spojitých rozložení, jejichž distribuční funkce se mohou lišit nejenom posunutím, ale také tvarem.

Testujeme hypotézu  $H_0$ : distribuční funkce těchto rozložení jsou shodné (tj. všech  $n + m$  veličin pochází z téhož rozložení) proti alternativě  $H_1$ : distribuční funkce jsou rozdílné.

Nechť  $F_1(x)$  je výběrová distribuční funkce 1. výběru a  $F_2(y)$  je výběrová distribuční funkce 2. výběru.

Testová statistika  $D = \max_x |F_n(x) - F_m(x)|$ .

$H_0$  zamítáme na hladině významnosti  $\alpha$ , když  $D > D_{n,m,\alpha}$ , kde  $D_{n,m,\alpha}$  je tabelovaná kritická hodnota.

Pro větší rozsahy lze kritickou hodnotu aproximovat vzorcem  $\sqrt{\frac{n+m}{2nm} \ln \frac{2}{\alpha}}$ .

**Příklad:** Výrobce určitého výrobku se má rozhodnout mezi dvěma dodavateli polotovarů vyrábějících je různými technologiemi. Rozhodující je procentní obsah určité látky.

1. technologie: 1,52 1,57 1,71 1,34 1,68

2. technologie: 1,75 1,67 1,56 1,66 1,72 1,79 1,64 1,55

Na hladině významnosti 0,05 posuďte pomocí dvouvýběrového K-S testu, zda je oprávněný předpoklad, že obě technologie poskytují stejné procento účinné látky.

### Výpočet pomocí systému STATISTICA:

Utvoříme nový datový soubor se dvěma proměnnými a 13 případy. Do proměnné X napíšeme zjištěné hodnoty a do proměnné ID napíšeme 5x číslo 1 pro první technologii a 8x číslo 2 pro starý druhou technologii.

Statistiky – Neparametrická statistika – Porovnání dvou nezávislých vzorků – OK – Proměnné – Seznam závislých proměnných X, Nezáv. (grupov.) proměnná ID – OK – Kolmogorov-Smirnovův 2-výběrový test.

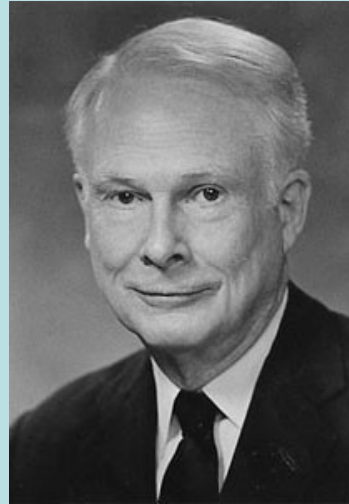
Promě.	Max z. rozdíl	Max k. rozdíl	Urove.	Prům. skup.	Prům. skup.	Sm.od. skup.	Sm.od. skup.	N pla. skup.	N pla. skup.
obsah	-0,400	0,025	$p > .$	1,564	1,667	0,147	0,085	5	8

Ve výstupní tabulce pro dvouvýběrový K-S test dostaneme maximální záporný a maximální kladný rozdíl mezi hodnotami obou výběrových distribučních funkcí, dolní omezení pro p-hodnotu ( $p > 0,1$ ), průměry, směrodatné odchylky a rozsahy obou výběrů. Jelikož p-hodnota převyšuje hladinu významnosti 0,05, na této hladině nelze nulovou hypotézu zamítnout.

## Kruskalův - Wallisův test



William Kruskal (1919 – 2005):  
Americký matematik



Wilson Allen Wallis (1912 – 1988):  
Americký matematik

Nechť je dáno  $r \geq 3$  nezávislých náhodných výběrů o rozsazích  $n_1, \dots, n_r$ . Předpokládáme, že tyto výběry pocházejí ze spojitých rozložení. Označme  $n = n_1 + \dots + n_r$ . Na asymptotické hladině významnosti  $\alpha$  chceme testovat hypotézu, že všechny tyto výběry pocházejí z téhož rozložení.

### Postup testu:

- a) Všechny  $n$  hodnot seřadíme do rostoucí posloupnosti.
- b) Určíme pořadí každé hodnoty v tomto sdruženém výběru.
- c) Označme  $T_j$  součet pořadí těch hodnot, které patří do  $j$ -tého výběru,  $j = 1, \dots, r$  (kontrola: musí platit  $T_1 + \dots + T_r = n(n+1)/2$ ).
- d) Testová statistika má tvar:  $Q = \frac{12}{n(n+1)} \sum_{j=1}^r \frac{T_j^2}{n_j} - 3(n+1)$ . Platí-li  $H_0$ , má statistika  $Q$  asymptoticky rozložení  $\chi^2(r-1)$ .
- e) Kritický obor:  $W = [c_1, r-1, \infty)$ .
- f)  $H_0$  zamítneme na asymptotické hladině významnosti  $\alpha$ , když  $Q \geq \chi_{1-\alpha}^2(r-1)$ .

**Příklad:** V roce 1980 byly získány tři nezávislé výběry obsahující údaje o průměrných ročních příjmech (v tisících dolarů) čtyř sociálních skupin ve třech různých oblastech USA.

jižní oblast: 6 10 15 29

pacifická oblast: 11 13 17 131

severovýchodní oblast: 7 14 28 25

Na hladině významnosti 0,05 testujte hypotézu, že příjmy v těchto oblastech se neliší.

**Řešení:**

Výpočty uspořádáme do tabulky

Usp. hodnoty	6	7	10	11	13	14	15	17	25	28	29	131
Pořadí 1.výběru	1		3				7				11	
Pořadí 2.výběru				4	5			8				12
Pořadí 3.výběru		2				6			9	10		

$$T_1 = 1 + 3 + 7 + 11 = 22,$$

$$T_2 = 4 + 5 + 8 + 12 = 29,$$

$$T_3 = 2 + 6 + 9 + 10 = 27,$$

$$Q = \frac{12}{n(n+1)} \sum_{j=1}^r \frac{T_j^2}{n_j} - \frac{3(n+1)}{n+1} = \frac{12}{12 \cdot 13} \left( \frac{22^2}{4} + \frac{29^2}{4} + \frac{27^2}{4} \right) - 13 = 5,$$

$$W = \chi_{1-\alpha}^2(r-1, \infty) = \chi_{0,95}^2(2, \infty) = 5,991$$

Protože  $Q < 5,991$ ,  $H_0$  nezamítáme na asymptotické hladině významnosti 0,05.

Rozdíly mezi průměrnými ročními příjmy v uvedených třech oblastech se neprokázaly.



## Mediánový test

Výchozí situace je stejná jako u K-W testu

### Postup testu:

- Všech  $n$  hodnot uspořádáme do rostoucí posloupnosti.
- Najdeme medián  $x_{0,50}$  těchto  $n$  hodnot.
- Označme  $P_j$  počet hodnot v  $j$ -tém výběru, které jsou větší nebo rovny mediánu  $x_{0,50}$ .
- Testová statistika má tvar  $Q_M = 4 \frac{\sum_{j=1}^r P_j^2}{n}$ . Platí-li  $H_0$ , má statistika  $Q_M$  asymptoticky rozložení  $\chi^2(r-1)$ .
- Kritický obor:  $W = [\chi_{1-\alpha}^2, \infty)$ .
- $H_0$  zamítneme na asymptotické hladině významnosti  $\alpha$ , když  $Q_M \geq \chi_{1-\alpha}^2(r-1)$ .

### Příklad:

Pro data o průměrných ročních příjmech proveďte mediánový test. Hladinu významnosti volte 0,05.

### Řešení:

Usp. hodnoty 6 7 10 11 13 14 15 17 25 28 29 131

Medián je průměr 6. a 7. uspořádané hodnoty:  $x_{0,50} = \frac{14 + 15}{2} = 14,5$ .

V prvním výběru existují 2 hodnoty, které jsou větší nebo rovny 14,5, stejně tak i ve druhém a třetím výběru, tedy  $P_1 = P_2 = P_3 = 2$ .

Testová statistika:  $Q_M = 4 \sum_{j=1}^r \frac{P_j^2}{n_j} - n = 4 \left( \frac{2^2}{10} + \frac{2^2}{10} + \frac{2^2}{10} \right) - 30 = 2$ .

Kritický obor:  $W = \chi^2_{1-\alpha, r-1, \infty} = \chi^2_{0,95, 2, \infty} = 5,991$ .

Protože  $Q_M < 5,991$ ,  $H_0$  nezamítáme na asymptotické hladině významnosti 0,05.

## Metody mnohonásobného porovnávání

Zamítneme-li hypotézu, že všechny náhodné výběry pocházejí z téhož rozložení, zajímá nás, které dvojice náhodných výběrů se liší na zvolené hladině významnosti. Testujeme  $H_0$ : k-tý a l-tý náhodný výběr pocházejí z téhož rozložení,  $k, l = 1, \dots, r, k \neq l$  proti  $H_1$ : aspoň jedna dvojice výběrů pochází z různých rozložení.

### a) Neményiho metoda (Peter Neményi 1927 – 2002: Americký matematik maďarského původu)

- Všechny výběry mají též rozsah  $p$  (třídění je vyvážené).
- Vypočteme  $|T_l - T_k|$ .
- V tabulkách najdeme kritickou hodnotu (pro dané  $p, r, \alpha$ ).
- Pokud  $|T_l - T_k| \geq$  tabelovaná kritická hodnota, pak na hladině významnosti  $\alpha$  zamítáme hypotézu, že l-tý a k-tý výběr pocházejí z téhož rozložení.

### b) Obecná metoda mnohonásobného porovnávání

- Vypočteme  $\left| \frac{T_l}{n_l} - \frac{T_k}{n_k} \right|$ .
- Ve speciálních statistických tabulkách najdeme kritickou hodnotu  $h_{KW}(\alpha)$ . Při větších rozsazích výběrů je možno ji nahradit kvantilem  $\chi_{1-\alpha}^2(r-1)$ .
- Jestliže  $\left| \frac{T_l}{n_l} - \frac{T_k}{n_k} \right| \geq \sqrt{\frac{1}{n_l} \left( \frac{1}{n_l} + \frac{1}{n_k} \right)} h_{KW}(\alpha)$ , pak na hladině významnosti  $\alpha$  zamítáme hypotézu, že l-tý a k-tý výběr pocházejí z téhož rozložení.

### **Příklad:**

Čtyři laboranti provedli analytické stanovení procenta niklu v oceli. Každý hodnotil pět vzorků.

Laborant A: 4,15 4,26 4,10 4,30 4,25

Laborant B: 4,38 4,40 4,29 4,39 4,45

Laborant C: 4,23 4,16 4,20 4,24 4,27

Laborant D: 4,41 4,31 4,42 4,37 4,43

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že všechny čtyři náhodné výběry pocházejí ze stejného rozložení. Pokud nulovou hypotézu zamítnete, zjistěte, které dvojice výběrů se liší.

### **Výpočet pomocí systému STATISTICA:**

Vytvoříme nový datový soubor o dvou proměnných a 20 případech. Do proměnné nikl napíšeme změřené hodnoty, do proměnné laborant napíšeme 5x1 pro 1. laboranta atd. až 5x4 pro 4. laboranta.

Statistiky – Neparametrická statistika – Porovnání více nezávislých vzorků - OK – Seznam závislých proměnných nikl, Nezáv. (grupovací) proměnná laborant – OK – Summary: Kruskal-Wallis ANOVA & Median test. Ve dvou výstupních tabulkách se objeví výsledky K-W testu a mediánového testu.

Kruskal-Wallisova ANOVA závislá na			
Nezávislá (grupovací) proměnná:			
Kruskal-Wallisův test: $H(3, N=20)$			
Závislá nikl	Ko	Počet platný	Součet pořadí
1	1	5	29,00
2	2	5	75,00
3	3	5	27,00
4	4	5	79,00

Mediánový test, celk. medián = 4,5					
Nezávislá (grupovací) proměnná:					
Chi-Kvadr. = 13,60000 sv = 3 p =					
Závislá: nikl	1	2	3	4	Celkem
<= Median: počet	4,00	1,00	5,00	0,00	10,00
oček.	2,50	2,50	2,50	2,50	
poz.-	1,50	-1,50	2,50	-2,50	
> Median: počet	1,00	4,00	0,00	5,00	10,00
oček.	2,50	2,50	2,50	2,50	
poz.-	-1,50	1,50	-2,50	2,50	
Celkem:	5,00	5,00	5,00	5,00	20,00

Oba testy zamítají hypotézu o shodě mediánů v daných čtyřech skupinách na asymptotické hladině významnosti 0,05.

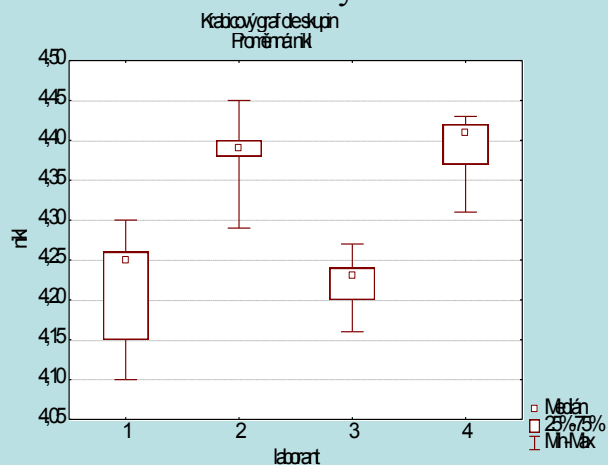
Nyní provedeme mnohonásobné porovnávání, abychom zjistili, které dvojice laborantů se liší. Zvolíme Vícenás. porovnání průměrného pořadí pro vš. skupiny.

Vícenásobné porovnání p-hodnot (0,05)  
 Nezávislá (grupovací) proměnná:  
 Kruskal-Wallisův test:  $H(3, N=20) =$

Závislá	1	2	3	4
ni	R: 5,8	R: 15,0	R: 5,4	R: 15,0
1		0,083	1,000	0,045
2	0,083		0,061	1,000
3	1,000	0,061		0,032
4	0,045	1,000	0,032	

Tabulka obsahuje p-hodnoty pro porovnání dvojic skupin. Vidíme, že na hladině významnosti 0,05 se liší laboranti A, D a laboranti C, D.

### Grafické znázornění výsledků



## Porovnání empirického a teoretického rozložení

### Osnova:

- testy dobré shody pro diskrétní a spojitě rozložení při úplně i neúplně specifikovaném problému
- jednoduchý test pro exponenciální a Poissonovo 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.

Opomíjení předpokladů o typu rozložení může v praxi vést i ke zcela zavádějícím výsledkům, proto je nutné věnovat tomuto problému patřičnou pozornost.

## Testy dobré shody pro diskrétní a spojité rozložení

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

- Je-li distribuční funkce spojitá, pak data rozdělíme do  $r$  třídících intervalů  $[u_j, u_{j+1})$ ,  $j = 1, \dots, r$ . Zjistíme absolutní četnost  $n_j$   $j$ -tého třídícího intervalu a 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  $p_j = P(u_j < X \leq u_{j+1}) = \Phi(u_{j+1}) - \Phi(u_j)$ .
- Má-li distribuční funkce nejvýše spočetně mnoho bodů nespojitosti, pak místo třídících intervalů použijeme varianty  $x_{[j]}$ ,  $j = 1, \dots, r$ . Pro variantu  $x_{[j]}$  zjistíme absolutní četnost  $n_j$  a vypočteme pravděpodobnost  $p_j$ , že náhodná veličina  $X$  s distribuční funkcí  $\Phi(x)$  se bude realizovat variantou  $x_{[j]}$ . Platí-li nulová hypotéza, pak

$$p_j = P(X = x_{[j]}) = \Phi(x_{[j]}) - \Phi(x_{[j-1]})$$

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.) Nulovou hypotézu zamítáme na asymptotické hladině významnosti  $\alpha$ , když testová statistika  $K \geq \chi^2_{1-\alpha}(r-1-p)$ . Aproximace se považuje za vyhovující, když teoretické četnosti  $np_j \geq 5$ ,  $j = 1, \dots, r$ .

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



### Příklad: Testování shody empirického a teoretického rozložení při úplně specifikovaném problému

Byl zjišťován počet poruch určitého zařízení za 100 hodin provozu ve 150 disjunktních 100 h intervalech. Výsledky měření:

Počet poruch za 100 hodin provozu	0	1	2	3	4 a víc
Absolutní četnost	52	48	36	10	4

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že náhodný výběr  $X_1, \dots, X_{150}$  pochází z rozložení  $Po(1,2)$ .

#### Řešení:

Pravděpodobnost, že náhodná veličina s rozložením  $Po(\lambda)$ , kde  $\lambda = 1,2$  bude nabývat hodnot 0, 1, ..., 4 a víc je

$$p_j = \frac{e^{-\lambda} \lambda^j}{j!}, j = 0, 1, 2, 3, 4, \dots$$

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

$j$	$n_j$	$p_j$	$np_j$	$(n_j - np_j)^2 / np_j$
0	52	0,301	150.0,301=45,15	1,039
1	48	0,361	150.0,361=54,15	0,698
2	36	0,217	150.0,217=32,55	0,366
3	10	0,087	150.0,087=13,05	0,713
4	4	0,034	150.0,034=5,1	0,237

Podmínky dobré aproximace jsou splněny, všechny teoretické četnosti jsou větší než 5.

$K = 1,039 + 0,698 + 0,713 + 0,237 = 3,053$ ,  $r = 5$ ,  $\chi^2_{0,95}(4) = 9,488$ . Protože  $3,053 < 9,488$ , nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.

## Výpočet pomocí systému STATISTICA:

Načteme datový soubor poruchy.sta. Proměnná POCET obsahuje počet poruch, proměnná CETNOST pak absolutní četnosti zjištěného počtu poruch.

Statistiky – Prokládání rozdělení – Diskrétní rozdělení – Poissonovo – OK – Proměnná POCET – klikneme na ikonu se závažím – Proměnná vah CETNOST – Stav Zapnuto – OK – záložka Parametry - Lambda 1,2 - Výpočet.

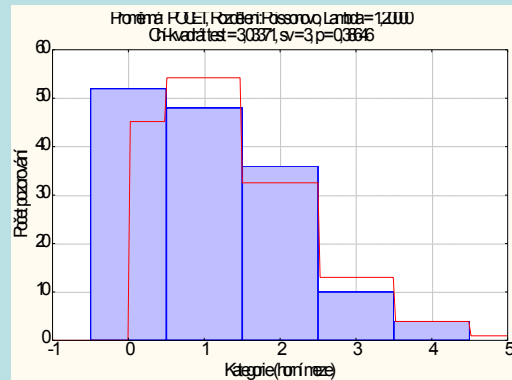
Proměnná: CETNOST, Rozdělení: Poissonovo, Lambda = 1,200 (poruchy.sta)								
Chi-kvadrát = 3,03371, sv = 3, p = 0,38646								
Kategorie	Pozorov. Cetnos	Kumul. Pozorov.	Procer. Pozorov.	Kumul. Pozorov.	Oček. Cetnc	Kumula. Očeká	Proce. Očeká	Kumul. Očeká
<= 0,000	5	5	34,66	34,66	45,17	45,17	30,11	30,11
1,00000	4	10	32,00	66,66	54,21	99,38	36,14	66,25
2,00000	3	13	24,00	90,66	32,52	131,90	21,68	87,93
3,00000	1	14	6,66	97,33	13,01	144,91	8,67	96,60
< Nekonec	4	15	2,66	100,0	5,06	150,0	3,37	100,0

V záhlaví výstupní tabulky je uvedena hodnota testového kritéria (3,03371), počet stupňů volnosti = 3 a p-hodnota (0,38646). Nulová hypotéza se tedy nezamítá na asymptotické hladině významnosti 0,05.

Počet stupňů volnosti 3 však neodpovídá tomu, že známe parametr  $\lambda$ , ve skutečnosti je počet stupňů volnosti 4. Proto pro výpočet p-hodnoty otevřeme nový datový soubor o jedné proměnné a jednom případě.

Do Dlouhého jména napíšeme =1-IChi2(3,03371;4). Dostaneme p-hodnotu 0,5522.

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



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

## Příklad: Testování shody empirického a teoretického rozložení při neúplně specifikovaném problému

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

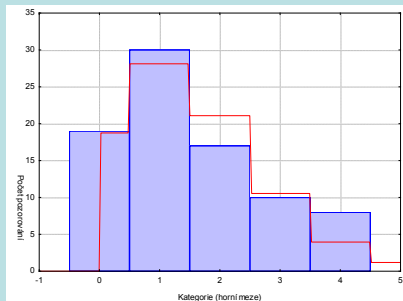
### Výpočet pomocí systému STATISTICA:

Načteme datový soubor branky.sta. 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) Chi-kvadrát = 2,07051, sv = 3, p = 0,55790									
Kategorie	Pozorov. Cetnos	Kumul. Pozorov.	Proce Pozorov.	Kumul. Pozorov.	Oček. Cetnc	Kumula Očeká	Proce Oček.	Kumul. Očeká	
<= 0,000	19	19	22,61	22,61	18,74	18,74	22,31	22,31	
1,00000	30	49	35,71	58,32	28,11	46,85	33,46	55,76	
2,00000	17	66	20,23	78,55	21,08	67,94	25,10	80,86	
3,00000	10	76	11,90	90,45	10,54	78,48	12,55	93,41	
< Nekonec	8	84	9,52	100,00	5,51	84,00	6,56	100,00	

V tomto případě je parametr  $\lambda$  Poissonova rozložení neznámý, je odhadnut pomocí výběrového průměru a odhad činí 1,5. Dále je v záhlaví výstupní tabulky uvedena hodnota testového kritéria (Chi kvadrát = 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 diskretních rozložení – Základní výsledky – Graf pozorovaného a očekávaného rozdělení.



**Poznámka k testu dobré shody:** Tento test 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:

číslo rostliny	1	2	3	4	5	6	7	8	9	10
počet žlutých semen	25	32	14	70	24	20	32	44	50	44
počet zelených semen	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.

$j$	$n_j$	$p_j$	$np_j$	$(n_j - np_j)^2 / np_j$
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

$$K = 0,148148 + 0,258547 + \dots + 0,134409 = 1,797495, r = 10, \chi^2_{0,95}(9) = 16,9.$$

Protože  $1,797495 < 16,9$ , nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.

## Výpočet pomocí systému STATISTICA:

Načteme datový soubor Mendel hrach.sta. Proměnná celkem obsahuje celkový počet semen, X obsahuje pozorovaný počet žlutých semen a Y vypočítané teoretické četnosti žlutých semen (v našem případě  $X \cdot 0,75$ ).

Statistiky – Neparametrická statistika – Pozorované versus očekávané  $\chi^2$  – OK - Pozorované četnosti X, Očekávané četnosti Y - OK – Výpočet. Dostaneme tabulku:

Přípa	Pozorovane vs. ocekavane cetno: Chi-Kvadr. = 1,797495 sv = 9 p = POZN.: Nestejné součty pozor. a			
	pozor X	očeká Y	P - C	(P-O) /O
C:	25,00	27,00	-2,00	0,148
C:	32,00	29,25	2,75	0,258
C:	14,00	14,25	-0,25	0,004
C:	70,00	72,75	-2,75	0,103
C:	24,00	27,75	-3,75	0,506
C:	20,00	19,50	0,50	0,012
C:	32,00	33,75	-1,75	0,090
C:	44,00	39,75	4,25	0,454
C:	50,00	48,00	2,00	0,083
C:	44,00	46,50	-2,50	0,134
Sčt	355,0	358,5	-3,50	1,797

Ve výstupní tabulce najdeme hodnotu testové statistiky (Chi-Kvadr = 1,797495), počet stupňů volnosti (sv = 9) a odpovídající p-hodnotu, kterou porovnáme se zvolenou hladinou významnosti. V našem případě je p-hodnota 0,99428, takže nulová hypotéza se nezamítá na asymptotické hladině významnosti 0,05.

**Příklad:** Při 60 hodech kostkou jsme dosáhli těchto výsledků: 9 x jednička, 11 x dvojka, 10 x trojka, 13 x čtyřka, 11 x pětka a 6 x šestka. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že kostka je homogenní.

**Řešení:**  $n = 60$

j	$n_j$	$p_j$	$np_j$	$(n_j - np_j)^2$	$(n_j - np_j)^2 / np_j$
1	9	1/6	10	1	1/10
2	11	1/6	10	1	1/10
3	10	1/6	10	0	0
4	13	1/6	10	9	9/10
5	11	1/6	10	1	1/10
6	6	1/6	10	16	16/10

$K = 2,8$ ,  $r = 6$ ,  $p = 0$ ,  $\chi^2_{0,95}(5) = 11,07$ . Protože  $K < 11,07$ ,  $H_0$  nezamítáme na asymptotické hladině významnosti 0,05.

### Výpočet pomocí systému STATISTICA:

Načteme datový soubor kostka.sta. Proměnná X obsahuje pozorované četnosti jednotlivých čísel 1, ..., 6 a proměnná Y obsahuje teoretické četnosti (v našem případě 10).

Statistiky – Neparametrická statistika – Pozorované versus očekávané  $\chi^2$  – OK - Pozorované četnosti X, Očekávané četnosti Y - OK – Výpočet. Dostaneme tabulku:

Příp.	Pozorované vs. očekávané četnosti			
	pozor. X	očeká. Y	P - C	(P-O)/O
C:	9,000	10,000	-1,000	0,100
C:	11,000	10,000	1,000	0,100
C:	10,000	10,000	0,000	0,000
C:	13,000	10,000	3,000	0,900
C:	11,000	10,000	1,000	0,100
C:	6,000	10,000	-4,000	1,600
Sčt	60,000	60,000	0,000	2,800

Ve výstupní tabulce najdeme hodnotu testové statistiky (Chi-Kvadr = 2,8), počet stupňů volnosti (sv = 5) a odpovídající p-hodnotu, kterou porovnáme se zvolenou hladinou významnosti. V našem případě je p-hodnota 0,730786, takže nulová hypotéza se nezamítá na asymptotické hladině významnosti 0,05.

**Příklad:** Ze záznamů autosalónu byl ve 100 náhodně vybraných dnech zjištěn počet prodaných aut.

Počet prodaných aut za den 0 1 2 3 4 5 a víc

Počet dnů 9 43 29 11 5 3

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že počet prodaných aut za den se řídí Poissonovým rozložením.

**Řešení:**

Parametr  $\lambda$  Poissonova rozložení neznáme, odhadneme ho pomocí výběrového průměru.

$$\bar{m} = \frac{1}{n} \sum_{j=0}^r n_j x_j = \frac{1}{100} (0 \cdot 9 + 1 \cdot 43 + 2 \cdot 29 + 3 \cdot 11 + 4 \cdot 5 + 5 \cdot 3) = 1,7$$

Pravděpodobnost, že náhodná veličina  $X \sim \text{Po}(1,7)$  bude

nabývat hodnot  $p_j, j = 0, 1, 2, 3, 4, 5$  a víc, je  $p_j = \frac{1,7^j}{j!} e^{-1,7}, j = 0, 1, 2, 3, 4, 5$

j	$n_j$	$p_j$	$np_j$	$(n_j - np_j)^2$	$(n_j - np_j)^2 / np_j$
0	9	0,1827	18,27	85,9329	4,7035
1	43	0,3106	31,06	142,5636	4,5899
2	29	0,264	26,4	6,76	0,2561
3	11	0,1496	14,96	15,6816	1,0482
4	5	0,0636	6,36	1,8496	0,2908
5 a víc	3	0,0296	2,96	0,0016	0,0005

Vidíme, že není splněna podmínka dobré aproximace. Sloučíme proto varianty 4 a 5.

j	$n_j$	$p_j$	$np_j$	$(n_j - np_j)^2$	$(n_j - np_j)^2 / np_j$
0	9	0,1827	18,27	85,9329	4,7035
1	43	0,3106	31,06	142,5636	4,5899
2	29	0,264	26,4	6,76	0,2561
3	11	0,1496	14,96	15,6816	1,0482
4 a víc	8	0,0932	9,32	1,7424	0,1869

$K = 10,7846, r = 5, p = 1, \chi^2_{0,95}(3) = 7,815$ . Protože  $K \geq 7,815$ ,  $H_0$  zamítáme na asymptotické hladině významnosti 0,05.

## Výpočet pomocí systému STATISTICA:

Načteme datový soubor autosalon.sta. Proměnná POCET obsahuje počet prodaných aut, proměnná CETNOST pak počet dnů, v nichž byl prodán zjištěný počet aut.

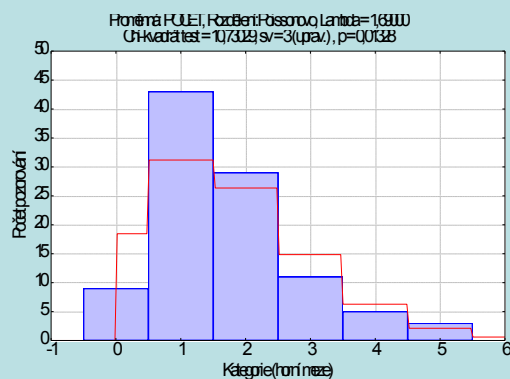
Statistiky – Prokládání rozdělení – Diskrétní rozdělení – Poissonovo – OK – Proměnná POCET – klikneme na ikonu se závažím – Proměnná vah CETNOST – Stav Zapnuto – OK – Výpočet.

Kategorie	Proměnná: POCET, Rozdělení: Poissonovo, Lambda = 1,69000 Chi-kvadrát = 10,73029, sv = 3 (uprav.), p = 0,01328								
	Pozorov. Cetnos	Kumul. Pozorov.	Procer. Pozorov.	Kumul. Pozorov.	Oček. Cetnc	Kumula Očeká	Proce Očeká	Kumul. Očeká	
<= 0,0000	5	5	9,00	9,00	18,45	18,45	18,45	18,45	
1,00000	4	9	43,00	52,00	31,18	49,63	31,18	49,63	
2,00000	2	11	29,00	81,00	26,35	75,98	26,35	75,98	
3,00000	1	12	11,00	92,00	14,84	90,82	14,84	90,82	
4,00000	5	17	5,00	97,00	6,27	97,10	6,27	97,10	
< Nekonec	3	20	3,00	100,0	2,89	100,0	2,89	100,0	

V záhlaví výstupní tabulky uvedena hodnota testového kritéria (10,73029), počet stupňů volnosti 3 a p-hodnota (0,01328). Nulová hypotéza se tedy zamítá na asymptotické hladině významnosti 0,05.

Vidíme, že nesouhlasí počet stupňů volnosti, měl by být 4. Proto p-hodnotu vypočteme zvlášť. Otevřeme nový datový soubor o jedné proměnné a jednom případě. Do Dlouhého jména napíšeme =1-ICHI2(10,73029;4). Dostaneme p-hodnotu 0,0298.

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



V tomto případě jsou patrné značné rozdíly mezi pozorovanými a teoretickými četnostmi.



### 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:  $W = (0, \chi^2_{\alpha/2}(n-1)] \cup [\chi^2_{1-\alpha/2}(n-1), \infty)$ .

Jestliže  $K \in W$ ,  $H_0$  zamítáme na asymptotické hladině významnosti  $\alpha$ .

**Příklad:** Byla zkoumána doba životnosti 45 součástek (v hodinách). Zjistili jsme, že průměrná doba životnosti činila  $m = 99,93$  h a rozptyl  $s^2 = 7328,91$  h<sup>2</sup>. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že daný náhodný výběr pochází z exponenciálního rozložení.

**Řešení:**

Testová statistika:  $K = \frac{(n-1)S^2}{M^2} = \frac{44 \cdot 7328,91}{99,93^2} = 292$

Kritický obor:  $W = (0, \chi^2_{\alpha/2}(n-1)] \cup [\chi^2_{1-\alpha/2}(n-1), \infty) = (0, \chi^2_{0,025}(44)] \cup [\chi^2_{0,975}(44), \infty) = (0, 25,75] \cup [64,02, \infty)$

Protože se testová statistika nerealizuje v kritickém oboru, hypotézu o exponenciálním rozložení 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  $K = \frac{\sum_{i=1}^n (X_i - M)^2}{M}$ , která se v případě platnosti  $H_0$  asymptoticky řídí rozložením  $\chi^2(n-1)$ .

Kritický obor:  $W = (0, \chi^2_{\alpha/2}(n-1)] \cup [\chi^2_{1-\alpha/2}(n-1), \infty)$ .

**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	5	6	7	8	9	10
Pozorovaná 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:

Nejprve musíme vypočítat realizaci výběrového průměru a výběrového rozptylu:

$$\bar{x} = \frac{1}{200} (0 \cdot 79 + 1 \cdot 188 + 2 \cdot 282 + 3 \cdot 275 + 4 \cdot 196 + 5 \cdot 114 + 6 \cdot 45 + 7 \cdot 10 + 8 \cdot 7 + 9 \cdot 3 + 10 \cdot 1) = 3,0$$

$$s^2 = \frac{1}{199} (0 \cdot 79 \cdot 3,0^2 + 1 \cdot 188 \cdot 3,0^2 + 2 \cdot 282 \cdot 3,0^2 + 3 \cdot 275 \cdot 3,0^2 + 4 \cdot 196 \cdot 3,0^2 + 5 \cdot 114 \cdot 3,0^2 + 6 \cdot 45 \cdot 3,0^2 + 7 \cdot 10 \cdot 3,0^2 + 8 \cdot 7 \cdot 3,0^2 + 9 \cdot 3 \cdot 3,0^2 + 10 \cdot 1 \cdot 3,0^2) = 7,085$$

$$K = \frac{\sum_{i=1}^n (X_i - \bar{x})^2}{\bar{x}} = \frac{1197,7085}{3,0} = 399,2361$$

$$\text{Kritický obor: } W = (0, \chi^2_{\alpha/2}(n-1)] \cup [\chi^2_{1-\alpha/2}(n-1), \infty) = (0, 110,93] \cup [296,986, \infty)$$

$H_0$  nezamítáme na asymptotické hladině významnosti 0,05.

**Příklad:** V systému hromadné obsluhy byla sledována doba obsluhy 70 zákazníků (v min). Výsledky jsou uvedeny v tabulce rozložení četností:

Doba obsluhy	Počet zákazníků
(0, 3]	14
(3, 6]	16
(6, 9]	10
(9, 12]	9
(12, 15]	8
(15, 18]	5
(18, 21]	3
(21, 24]	5

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

- test dobré shody,
- jednoduchý test exponenciálního rozložení

**Řešení:**

Testujeme  $H_0$ : náhodný výběr  $X_1, \dots, X_{70}$  pochází z  $Ex(\lambda)$  proti  $H_1$ : non  $H_0$ .

Ad a) Nejprve odhadneme parametr  $\lambda$  exponenciálního rozložení:  $\hat{\lambda} = \frac{1}{\bar{x}} = \frac{1}{\frac{1}{n} \sum_{j=1}^r u_j \cdot k_j} = \frac{1}{\frac{1}{70} (14 \cdot 1,5 + 16 \cdot 4,5 + 10 \cdot 7,5 + 9 \cdot 10,5 + 8 \cdot 13,5 + 5 \cdot 16,5 + 3 \cdot 19,5 + 5 \cdot 22,5)} = 0,1122$

Pravděpodobnost, že náhodná veličina s rozložením  $Ex(\lambda)$ , kde  $\lambda = 0,1122$  se bude realizovat v intervalu  $[u_j, u_{j+1})$  je  $p_j = \Phi(u_{j+1}) - \Phi(u_j)$ ,  $j = 1, \dots, r$ , kde  $\Phi(x) = 1 - e^{-\lambda x}$ .

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

$u_j, u_{j+}$	$x_{[j]}$	$n_j$	$p_j$	$np_j$
(0, 3]	1,5	14	0,2858	20,0033
(3,6]	4,5	16	0,2041	14,2871
(6,9]	7,5	10	0,1458	10,2044
(9,12]	10,5	9	0,1041	7,2884
(12,15]	13,5	8	0,0744	5,2056
(15,18]	16,5	5	0,0531	3,7181
(18,21]	19,5	3	0,0378	2,6556
(21,24]	22,5	5	0,0271	1,8967

Podmínky dobré aproximace nejsou splněny, sloučíme tedy intervaly (15,18], (18,21] a (21,24].

$u_j, u_{j+}$	$x_{[j]}$	$n_j$	$p_j$	$np_j$	$(n_j - np_j)^2 / np_j$
(0, 3]	1,5	14	0,2858	20,0033	1,8017
(3,6]	4,5	16	0,2041	14,2871	0,2054
(6,9]	7,5	10	0,1458	10,2044	0,0041
(9,12]	10,5	9	0,1041	7,2884	0,4020
(12,15]	13,5	8	0,0744	5,2056	1,5000
(15,24]	19,5	13	0,1181	8,2704	2,7047

Testová statistika  $K = 1,8017 + \dots + 2,7047 = 6,6178$ ,  $r = 6$ ,  $p = 1$ ,  $r - p - 1 = 4$ ,  $\chi^2_{0,95}(4) = 9,4877$ .

Testová statistika se nerealizuje v kritickém oboru  $W = 9,4877$ , na asymptotické hladině významnosti 0,05 nelze zamítnout hypotézu, že doba obsluhy se řídí exponenciálním rozložením.

Ad b) Jednoduchý test exponenciálního rozložení je založen na statistice  $K = \frac{n-1}{\sqrt{n}}$ , která se v případě platnosti  $H_0$  asymptoticky řídí rozložením  $\chi^2(n-1)$ .

Kritický obor:  $W = \{ \chi_{\alpha/2}^2(n-1) \cup \chi_{1-\alpha/2}^2(n-1), \infty \}$ .

Nejprve musíme vypočítat realizaci výběrového průměru a výběrového rozptylu:

$$\bar{m} = \frac{1}{n} (1,45 + 1,45 + \dots + 2,25) = 1,914$$

$$s^2 = \frac{1}{n} (9 \cdot 1,45^2 - 914^2 + 9 \cdot 1,45^2 - 914^2 + \dots + 225 - 914^2) = 1,144$$

$$K = \frac{n-1}{\sqrt{n}} = \frac{60 \cdot 1,1447}{1,43} = 5726$$

Kritický obor:  $W = \{ \chi_{\alpha/2}^2(n-1) \cup \chi_{1-\alpha/2}^2(n-1), \infty \} = \{ \chi_{0,025}^2(69) \cup \chi_{0,975}^2(69), \infty \} = \{ 0,479242 \cup 93,8565, \infty \}$ .

$H_0$  zamítáme na asymptotické hladině významnosti 0,05.

## Hodnocení kontingenčních tabulek

### Osnova:

- zavedení kontingenční tabulky
- testování hypotézy o nezávislosti a měření síly závislosti
- test homogeneity
- analýza čtyřpolních tabulek

### Motivace

Při zpracování dat se velmi často setkáme s úkolem zjistit, zda dvě náhodné veličiny nominálního typu jsou stochasticky nezávislé. Např. nás může zajímat, zda ve sledované populaci je barva očí a barva vlasů nezávislá.

Zpravidla chceme také zjistit intenzitu případné závislosti sledovaných dvou veličin. K tomuto účelu byly zkonstruovány různé koeficienty, které nabývají hodnot od 0 do 1. Čím je takový koeficient bližší 1, tím je závislost mezi danými dvěma veličinami silnější a čím je bližší 0, tím je slabší.

## Kontingenční tabulky

Nechť  $X, Y$  jsou dvě nominální náhodné veličiny (tj. obsahová interpretace je možná jenom u relace rovnosti). Nechť  $X$  nabývá variant  $x_{[1]}, \dots, x_{[r]}$  a  $Y$  nabývá variant  $y_{[1]}, \dots, y_{[s]}$ .

Označme:

$\pi_{jk} = P(X = x_{[j]} \wedge Y = y_{[k]})$  ... simultánní pravděpodobnost dvojice variant  $(x_{[j]}, y_{[k]})$

$\pi_{j.} = P(X = x_{[j]})$  ... marginální pravděpodobnost varianty  $x_{[j]}$

$\pi_{.k} = P(Y = y_{[k]})$  ... marginální pravděpodobnost varianty  $y_{[k]}$

Simultánní a marginální pravděpodobnosti zapíšeme do kontingenční tabulky:

	$y$	$y_{[1]}$	...	$y_{[s]}$	$\pi_{j.}$
$X$	$\pi_{jk}$				
$x_{[1]}$		$\pi_{11}$	...	$\pi_{1s}$	$\pi_{1.}$
...		...	...	...	...
$x_{[r]}$		$\pi_{r1}$	...	$\pi_{rs}$	$\pi_{r.}$
$\pi_{.k}$		$\pi_{.1}$	...	$\pi_{.s}$	1

Pořídíme dvourozměrný náhodný výběr  $(X_1, Y_1), \dots, (X_n, Y_n)$  rozsahu  $n$  z rozložení, kterým se řídí dvourozměrný diskretní náhodný vektor  $(X, Y)$ . Zjištěné absolutní simultánní četnosti  $n_{jk}$  dvojice variant  $(x_{[j]}, y_{[k]})$  uspořádáme do kontingenční tabulky:

	y	$y_{[1]}$	...	$y_{[s]}$	$n_{j\cdot}$
x	$n_{jk}$				
$x_{[1]}$	$n_{11}$	...	$n_{1s}$	$n_{1\cdot}$	
...	...	...	...	...	
$x_{[r]}$	$n_{r1}$	...	$n_{rs}$	$n_{r\cdot}$	
$n_{\cdot k}$	$n_{\cdot 1}$	...	$n_{\cdot s}$	$n$	

$n_{j\cdot} = n_{j1} + \dots + n_{js}$  je marginální absolutní četnost varianty  $x_{[j]}$

$n_{\cdot k} = n_{1k} + \dots + n_{rk}$  je marginální absolutní četnost varianty  $y_{[k]}$

Simultánní pravděpodobnost  $\pi_{jk}$  odhadneme pomocí simultánní relativní četnosti  $p_{jk} = \frac{n_{jk}}{n}$ , marginální pravděpodobnosti  $\pi_{j\cdot}$

a  $\pi_{\cdot k}$  odhadneme pomocí marginálních relativních četností  $p_{j\cdot} = \frac{n_{j\cdot}}{n}$  a  $p_{\cdot k} = \frac{n_{\cdot k}}{n}$ .



Pořídíme dvourozměrný náhodný výběr  $(X_1, Y_1), \dots, (X_n, Y_n)$  rozsahu  $n$  z rozložení, kterým se řídí dvourozměrný diskretní náhodný vektor  $(X, Y)$ . Zjištěné absolutní simultánní četnosti  $n_{jk}$  dvojice variant  $(x_{[j]}, y_{[k]})$  uspořádáme do kontingenční tabulky:

	y	$y_{[1]}$	...	$y_{[s]}$	$n_{j\cdot}$
x	$n_{jk}$				
$x_{[1]}$	$n_{11}$	...	$n_{1s}$	$n_{1\cdot}$	
...	...	...	...	...	
$x_{[r]}$	$n_{r1}$	...	$n_{rs}$	$n_{r\cdot}$	
$n_{\cdot k}$	$n_{\cdot 1}$	...	$n_{\cdot s}$	$n$	

$n_{j\cdot} = n_{j1} + \dots + n_{js}$  je marginální absolutní četnost varianty  $x_{[j]}$

$n_{\cdot k} = n_{1k} + \dots + n_{rk}$  je marginální absolutní četnost varianty  $y_{[k]}$

Simultánní pravděpodobnost  $\pi_{jk}$  odhadneme pomocí simultánní relativní četnosti  $p_{jk} = \frac{n_{jk}}{n}$ , marginální pravděpodobnosti  $\pi_{j\cdot}$

a  $\pi_{\cdot k}$  odhadneme pomocí marginálních relativních četností  $p_{j\cdot} = \frac{n_{j\cdot}}{n}$  a  $p_{\cdot k} = \frac{n_{\cdot k}}{n}$ .

## Testování hypotézy o nezávislosti

Testujeme nulovou hypotézu  $H_0$ : X, Y jsou stochasticky nezávislé náhodné veličiny proti alternativě  $H_1$ : X, Y nejsou stochasticky nezávislé náhodné veličiny.

Kdyby náhodné veličiny X, Y byly stochasticky nezávislé, pak by platil multiplikační vztah

$\forall j \in \{1, \dots, r\}, \forall k \in \{1, \dots, s\}: \pi_{jk} = \pi_j \cdot \pi_k$  neboli  $\frac{n_{jk}}{n} = \frac{n_j}{n} \cdot \frac{n_k}{n}$ , tj.  $n_{jk} = \frac{n_j \cdot n_k}{n}$ . Číslo  $\frac{n_j \cdot n_k}{n}$  se nazývá **teoretická četnost** dvojice variant  $(x_{[j]}, y_{[k]})$ .

Testová statistika: 
$$K = \sum_{j=1}^r \sum_{k=1}^s \left( \frac{n_{jk} - \frac{n_j \cdot n_k}{n}}{\frac{n_j \cdot n_k}{n}} \right)^2$$

Platí-li  $H_0$ , pak K se asymptoticky řídí rozložením  $\chi^2((r-1)(s-1))$ .

Kritický obor:  $W = [2_{1-\alpha}, (r-1)(s-1), \infty)$ .

Hypotézu o nezávislosti veličin X, Y tedy zamítáme na asymptotické hladině významnosti  $\alpha$ , když  $K \geq \chi^2_{1-\alpha}((r-1)(s-1))$ .

## Podmínky dobré aproximace

Rozložení statistiky K lze aproximovat rozložením  $\chi^2((r-1)(s-1))$ , pokud teoretické četnosti  $\frac{n_j \cdot n_k}{n}$  aspoň v 80% případů nabývají hodnoty větší nebo rovné 5 a ve zbylých 20% neklesnou pod 2. Není-li splněna podmínka dobré aproximace, doporučuje se slučování některých variant.

## Měření síly závislosti

**Cramérův koeficient:**  $V = \frac{\sqrt{K}}{\sqrt{m}}$ , kde  $m = \min\{r,s\}$ . Tento koeficient nabývá hodnot mezi 0 a 1. Čím blíže je k 1, tím je

závislost mezi X a Y těsnější, čím blíže je k 0, tím je tato závislost volnější.

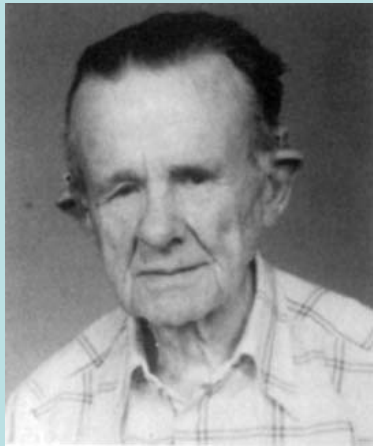
Význam hodnot Cramérova koeficientu:

mezi 0 až 0,1 ... zanedbatelná závislost,

mezi 0,1 až 0,3 ... slabá závislost,

mezi 0,3 až 0,7 ... střední závislost,

mezi 0,7 až 1 ... silná závislost.



Carl Harald Cramér (1893 – 1985): Švédský matematik

## Příklad

V sociologickém průzkumu byl z uchazečů o studium na vysokých školách pořízen náhodný výběr rozsahu 360. Mimo jiné se zjišťovala sociální skupina, ze které uchazeč pochází (veličina  $X$ ) a typ školy, na kterou se hlásí (veličina  $Y$ ). Výsledky jsou zaznamenány v kontingenční tabulce:

Sociální skupina	Typ školy			$n_{j.}$
	univerzitní	technický	ekonomický	
I	50	30	10	90
II	30	50	20	100
III	10	20	30	60
IV	50	10	50	110
$n_{.k}$	140	110	110	360

Na asymptotické hladině významnosti 0,05 testujte hypotézu o nezávislosti typu školy a sociální skupiny. Vypočtěte Cramérovův koeficient.

## Řešení:

Nejprve vypočteme všech 12 teoretických četností:

Sociální skupina	Typ školy			$n_{j.}$
	univerzitní	technický	ekonomický	
I	50	30	10	90
II	30	50	20	100
III	10	20	30	60
IV	50	10	50	110
$n_{.k}$	140	110	110	360

$$\begin{aligned} \frac{n_{11}n_{.1}}{n} &= \frac{90 \cdot 140}{360} = 35, & \frac{n_{12}n_{.2}}{n} &= \frac{90 \cdot 110}{360} = 27, & \frac{n_{13}n_{.3}}{n} &= \frac{90 \cdot 110}{360} = 27, \\ \frac{n_{21}n_{.1}}{n} &= \frac{100 \cdot 140}{360} = 38, & \frac{n_{22}n_{.2}}{n} &= \frac{100 \cdot 110}{360} = 29, & \frac{n_{23}n_{.3}}{n} &= \frac{100 \cdot 110}{360} = 29, \\ \frac{n_{31}n_{.1}}{n} &= \frac{60 \cdot 140}{360} = 23, & \frac{n_{32}n_{.2}}{n} &= \frac{60 \cdot 110}{360} = 18, & \frac{n_{33}n_{.3}}{n} &= \frac{60 \cdot 110}{360} = 18, \\ \frac{n_{41}n_{.1}}{n} &= \frac{110 \cdot 140}{360} = 42, & \frac{n_{42}n_{.2}}{n} &= \frac{110 \cdot 110}{360} = 33, & \frac{n_{43}n_{.3}}{n} &= \frac{110 \cdot 110}{360} = 33. \end{aligned}$$

Vidíme, že podmínky dobré aproximace jsou splněny, všechny teoretické četnosti převyšují číslo 5.

Dosadíme do vzorce pro testovou statistiku K:

$$K = \frac{50^2}{90} + \frac{30^2}{90} + \frac{10^2}{90} + \frac{30^2}{100} + \frac{50^2}{100} + \frac{20^2}{100} + \frac{10^2}{60} + \frac{20^2}{60} + \frac{30^2}{60} + \frac{50^2}{110} + \frac{10^2}{110} + \frac{50^2}{110} = 34.$$

Dále stanovíme kritický obor:

$$W = \chi_{1-0,05}^2(4-1)(3-1) = \chi_{0,95}^2(6) = 12,6$$

Protože  $K > W$ , hypotézu o nezávislosti typu školy a sociální skupiny zamítáme na asymptotické hladině významnosti 0,05.

Vypočteme Cramérův koeficient:  $V = \sqrt{\frac{764}{360}} = 0,46$ .

Hodnota Cramérova koeficientu svědčí o tom, že mezi veličinami X a Y existuje středně silná závislost.

## Výpočet pomocí systému STATISTICA:

Vytvoříme nový datový soubor o třech proměnných (X - sociální skupina, Y – typ školy, četnost) a 12 případech:

	1 X	2 Y	3 četn
1	I	univerzi	5
2	I	technick	3
3	I	ekonom	1
4	II	univerzi	3
5	II	technick	5
6	II	ekonom	2
7	III	univerzi	1
8	III	technick	2
9	III	ekonom	3
10	IV	univerzi	5
11	IV	technick	1
12	IV	ekonom	5

Statistiky – Základní statistiky/tabulky – OK – Specif. Tabulky – List 1 X, List 2 Y – OK, zapneme proměnnou vah četnost – OK, Výpočet – na záložce Možnosti zaškrtneme Očekávané četnosti. Dostaneme kontingenční tabulku teoretických četností:

Souhrnná tab.: Očekávané četnosti (Četnost označených buněk > 10)  
 Pearsonův chí-kv. : 76,8359, sv=6, p

X	Y univerzi	Y technic	Y ekonomi	Řádk součt
I	35,00	27,50	27,50	90,00
II	38,89	30,56	30,56	100,00
III	23,33	18,33	18,33	60,00
IV	42,78	33,64	33,64	110,00
vs.skl	140,00	110,00	110,00	360,00

Všechny teoretické četnosti jsou větší než 5, podmínky dobré aproximace jsou splněny. V záhlaví tabulky je uvedena hodnota testové statistiky  $K = 76,8359$ , počet stupňů volnosti 6 a odpovídající p-hodnota. Je velmi blízká 0, tedy na asymptotické hladině významnosti 0,05 zamítáme hypotézu o nezávislosti typu školy a sociální skupiny.

Hodnotu testové statistiky a Cramérův koeficient dostaneme také tak, že na na záložce Možnosti zaškrtneme Pearsonův & M-V chí kvadrát a Cramérovo V, na záložce Detailní výsledky vybereme Detailní 2 rozm. tabulky.

Statist.	Chi-kv.	sv	p
Pearsonův chí-k	76,83	df=	p=,00
M-V chí-kvadr.	84,53	df=	p=,00
F1	,4619		
Kontingenční ko	,4193		
Cramér. V	,3266		

## Test homogenity v tabulce typu 2 x s

Máme kontingenční tabulku, v níž veličina X má jen dvě varianty a veličina Y s variant:

	y	Y <sub>[1]</sub>	...	Y <sub>[s]</sub>	$\pi_{j.}$
X	$\pi_{jk}$				
X <sub>[1]</sub>		$\pi_{11}$	...	$\pi_{1s}$	$\pi_{1.}$
X <sub>[2]</sub>		$\pi_{21}$	...	$\pi_{2s}$	$\pi_{2.}$
$\pi_{.k}$		$\pi_{.1}$	...	$\pi_{.s}$	1

Pořídíme dvourozměrný náhodný výběr  $(X_1, Y_1), \dots, (X_n, Y_n)$  rozsahu n z rozložení, kterým se řídí dvourozměrný diskretní náhodný vektor  $(X, Y)$ . Zjištěné absolutní simultánní četnosti  $n_{jk}$  dvojice variant  $(x_{[j]}, y_{[k]})$  uspořádáme do kontingenční tabulky:

	y	Y <sub>[1]</sub>	...	Y <sub>[s]</sub>	$n_{j.}$
X	$\pi_{jk}$				
X <sub>[1]</sub>		$n_{11}$	...	$n_{1s}$	$n_{1.}$
X <sub>[2]</sub>		$n_{21}$	...	$n_{2s}$	$n_{2.}$
$\pi_{.k}$		$n_{.1}$	...	$n_{.s}$	n

Na asymptotické hladině významnosti  $\alpha$  testujeme hypotézu  $H_0: \pi_{1k} = \pi_{2k}, k = 1, 2, \dots, s$  proti alternativě  $H_1$ : aspoň jedna dvojice pravděpodobností se liší.

Na problém lze pohlížet tak, že máme s nezávislých náhodných výběrů z alternativních rozložení, přičemž první má rozsah  $n_1 = n_{11} + n_{21}$  a pochází z rozložení  $A(\underline{q})$ , ..., s-tý má rozsah  $n_s = n_{1s} + n_{2s}$  a pochází z rozložení  $A(\underline{q})$ . Testujeme hypotézu  $H_0: \underline{q} \dots =$  proti alternativě  $H_1$ : non  $H_0$ .



V kapitole o hodnocení náhodných výběrů z alternativních rozložení jsme použili testovou statistiku:

$$Q = \sum_{j=1}^s \frac{(f_j - \bar{f})^2}{\bar{f}} \approx \chi^2_{s-1}, \text{ když } H_0 \text{ platí.}$$

Kritický obor:  $W_{\alpha, s-1, \infty}$

$H_0$  tedy zamítáme na asymptotické hladině významnosti  $\alpha$ , když  $Q \in W_{\alpha, s-1, \infty}$ . Přitom  $M = \sum_{j=1}^s \frac{f_j}{n} \bar{x}_j$  je vážený průměr výběrových průměrů.

Nyní použijeme testovou statistiku  $K = \sum_{j=1}^s \sum_{k=1}^s \left( \frac{n_{jk} - \frac{n_j n_k}{n}}{\frac{n_j n_k}{n}} \right)^2$ , stejně jako u testu nezávislosti. Lze dokázat, že při výše

uvedeném označení jsou statistiky  $Q$  a  $K$  totožné. Tedy test homogenity lze provést stejně jako test nezávislosti.

Tato statistika se v případě platnosti nulové hypotézy asymptoticky řídí rozložením  $\chi^2(s-1)$ . Kritický obor:  $W_{\alpha, s-1, \infty}$ .

Nulovou hypotézu zamítáme na asymptotické hladině významnosti  $\alpha$ , když  $K \in W_{\alpha, s-1, \infty}$ .

**Příklad:** 104 náhodně vybraných matek bylo dotázáno, zda jejich kojeneček dostává dudlík. Zjišťoval se též nejvyšší stupeň dosaženého vzdělání matky.

Vzdělání matky	Počet matek	Počet dětí s dudlíkem
ZŠ	39	27
SŠ	47	34
VŠ	18	15

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že používání dudlíku nezávisí na vzdělání matky. (Jedná se o příklad 8.6.2. ze skript Základní statistické metody. Zde je uvedeno, že testová statistika Q se realizuje hodnotou 1,267, kritický obor je  $W_{=5992}_{\infty}$ , tedy nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.)

**Řešení:** Data zapíšeme do kontingenční tabulky 2 x 3.

	Matka ZŠ	Matka SŠ	Matka VŠ	$n_{.j}$
Dudlík ano	27	34	15	76
Dudlík ne	12	13	3	28
$n_{.k}$	39	47	18	104

Ověříme splnění podmínek dobré aproximace:

$$\frac{n_{11}n_{1.}}{n} = \frac{76 \cdot 39}{104} = 28,5, \quad \frac{n_{12}n_{1.}}{n} = \frac{76 \cdot 47}{104} = 34,5, \quad \frac{n_{13}n_{1.}}{n} = \frac{76 \cdot 18}{104} = 13,5, \quad \frac{n_{21}n_{2.}}{n} = \frac{28 \cdot 39}{104} = 10,5, \quad \frac{n_{22}n_{2.}}{n} = \frac{28 \cdot 47}{104} = 12,5, \quad \frac{n_{23}n_{2.}}{n} = \frac{28 \cdot 18}{104} = 4,8$$

Podmínky dobré aproximace jsou splněny, pouze v 1 případě ze 6 je teoretická četnost menší než 5.

Dosadíme do vzorce pro testovou statistiku K:

$$K = \frac{27^2}{29 \cdot 39} + \frac{34^2}{12 \cdot 47} + \frac{3^2}{12 \cdot 18} = 1,268$$

Kritický obor:  $W_{=2,1} s_{=1, \infty} \chi^2_{0,05} 2, \infty = 5992_{\infty}$

Na asymptotické hladině významnosti 0,05 se tedy neprokázalo, že používání dudlíku závisí na vzdělání matky.

## Čtyřpolní tabulky

Nechť  $r = s = 2$ . Pak hovoříme o **čtyřpolní kontingenční tabulce** a používáme označení:  $n_{11} = a$ ,  $n_{12} = b$ ,  $n_{21} = c$ ,  $n_{22} = d$ .

X	Y		$n_{j.}$
	$y_{[1]}$	$y_{[2]}$	
$x_{[1]}$	a	b	a+b
$x_{[2]}$	c	d	c+d
$n_{.k}$	a+c	b+d	n

## Test nezávislosti ve čtyřpolní tabulce

Testovou statistiku pro čtyřpolní kontingenční tabulku lze zjednodušit do tvaru:

$$K = \frac{n \cdot \text{nad}^2}{(a+b)(c+d)(a+c)(b+d)}$$

Platí-li hypotéza o nezávislosti veličin X, Y, pak K se asymptoticky řídí rozložením  $\chi^2(1)$ .

Kritický obor:  $W = [1, \infty)$

Nulovou hypotézu zamítáme na asymptotické hladině významnosti  $\alpha$ , když  $K \in W$ .

Povšimněte si, že za platnosti hypotézy o nezávislosti  $ad = bc$ .

Pro čtyřpolní tabulku navrhl R. A. Fisher přesný (exaktní) test nezávislosti známý jako **Fisherův faktoriálový test**.



Sir Ronald Aylmer Fisher (1890 – 1962): Britský statistik a genetik.

(Fisherův přesný test je popsán např. v knize K. Zvára: Biostatistika, Karolinum, Praha 1998. Princip spočívá v tom, že pomocí kombinatorických úvah se vypočítají pravděpodobnosti toho, že při daných marginálních četnostech dostaneme tabulky, které se od nulové hypotézy odchyľují aspoň tak, jako daná tabulka.)

**Upozornění:** STATISTICA poskytuje p-hodnotu pro Fisherův přesný test. Jestliže vyjde  $p \leq \alpha$ , pak hypotézu o nezávislosti zamítáme na hladině významnosti  $\alpha$ .

**Příklad:** V náhodném výběru 50 obézních dětí ve věku 6 – 14 let byla zjišťována obezita rodičů. Veličina X – obezita matky, veličina Y – obezita otce. Výsledky průzkumu jsou uvedeny v kontingenční tabulce:

X	Y		$n_{j.}$
	ano	ne	
ano	15	9	24
ne	7	19	26
$n_{.k}$	22	28	50

Pomocí Fisherova exaktního testu ověřte, zda lze na hladině významnosti 0,05 zamítnout hypotézu o nezávislosti náhodných veličin X a Y.

## Výpočet pomocí systému STATISTICA:

Vytvoříme datový soubor o třech proměnných X, Y (varianty 0 – neobézní, 1 – obézní) a četnost a čtyřech případech:

	1 X	2 Y	3 četno
1	obezn	obezn	1
2	obezn	neobe	9
3	neobe	obezn	7
4	neobe	neobe	1

Statistiky – Základní statistiky/tabulky – OK – Specif. Tabulky – List 1 X, List 2 Y – OK, zapneme proměnnou vah četnost – OK, Výpočet – na záložce Možnosti zaškrtneme Fisher exakt., Yates, McNemar (2x2). Dostaneme výstupní tabulku:

Statist.	Statist. : X(2) x Y(2) (d		
	Chi-kv	sv	p
Pearsonuv chi-kv	6,410	df=	p=,01
M-V chi-kvadr.	6,548	df=	p=,01
Yatesuv chi-kv.	5,048	df=	p=,02
Fisheruv přesny, 2-stranny			p=,01 p=,02
McNemaruv chi-k (B/C)	,2647 ,0625	df=	p=,60 p=,80

Vidíme, že p-hodnota pro Fisherův exaktní oboustranný test je 0,02163, tedy na hladině významnosti 0,05 zamítáme hypotézu, že obezita matky a otce spolu nesouvisí.

## Test homogenity ve čtyřpolní tabulce

Na asymptotické hladině významnosti  $\alpha$  testujeme hypotézu  $H_0: \pi_{1k} = \pi_{2k}, k = 1, 2$  proti alternativě  $H_1$ : aspoň jedna dvojice pravděpodobností se liší. Na problém lze pohlížet tak, že máme dva nezávislé výběry z alternativních rozložení, první má rozsah  $n_1 = a+c$  a pochází z rozložení  $A(\underline{q})$ , druhý má rozsah  $n_2 = b+d$  a pochází z rozložení  $A(\underline{q})$ . Testujeme hypotézu  $H_0: \underline{q} = 0$  proti oboustranné alternativě.

V kapitole o hodnocení náhodných výběrů z alternativních rozložení jsme použili testovou statistiku

$T_0 = \frac{M_1 - M_2}{\sqrt{M_1(1-M_1) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$ , která se za platnosti nulové hypotézy asymptoticky řídí rozložením  $N(0,1)$ . ( $M_*$  je vážený průměr výběrových průměrů.)

Nyní použijeme testovou statistiku  $K = \frac{n \hat{c}^2}{a_+ b_+ r_+ d_+}$ , stejně jako u testu nezávislosti. Tato statistika se v případě platnosti nulové hypotézy asymptoticky řídí rozložením  $\chi^2(1)$ . Kritický obor:  $W = [1, \infty)$ . Nulovou hypotézu zamítáme na asymptotické hladině významnosti  $\alpha$ , když  $K \geq W$ .

**Příklad:** Očkování proti chřipce se zúčastnilo 460 dospělých, z nichž 240 dostalo očkovací látku proti chřipce a 220 dostalo placebo. Na konci experimentu onemocnělo 100 lidí chřipkou. 20 z nich bylo z očkové skupiny a 80 z kontrolní skupiny. Na asymptotické hladině významnosti 0,01 testujte hypotézu, že výskyt chřipky v očkové a kontrolní skupině je shodný.

**Řešení:**

Údaje uspořádáme do čtyřpolní kontingenční tabulky, kde roli veličiny X hraje onemocnění chřipkou a roli veličiny Y existence očkování.

X onemocnění chřipkou	Y existence očkování		n <sub>j</sub>
	ano	ne	
ano	20	80	100
ne	220	140	360
n <sub>k</sub>	240	220	460

Vypočteme sloupcově podmíněné relativní četnosti:

X onemocnění chřipkou	Y existence očkování	
	ano	ne
ano	8,3%	36,4%
ne	91,7%	63,6%

Vidíme, že v očkové skupině onemocnělo chřipkou 8,3% lidí, v kontrolní skupině však 36,4%. Zjistíme, zda takto velký rozdíl je způsoben pouze náhodnými vlivy.



Ověříme splnění podmínek dobré aproximace, tedy nejprve vypočteme teoretické četnosti:

X onemocnění chřipkou	Y existence očkování		n <sub>j.</sub>
	ano	ne	
ano	20	80	100
ne	220	140	360
n <sub>k</sub>	240	220	460

$$\frac{n_1 n_1}{n} = \frac{100 \cdot 240}{460} = 51,7, \quad \frac{n_1 n_2}{n} = \frac{100 \cdot 220}{460} = 47,83$$

$$\frac{n_2 n_1}{n} = \frac{360 \cdot 240}{460} = 383, \quad \frac{n_2 n_2}{n} = \frac{360 \cdot 220}{460} = 727$$

Všechny teoretické četnosti jsou větší než 5, podmínky dobré aproximace jsou splněny.

Realizace testové statistiky:

$$K = \sum_{j=1}^k \frac{(n_{kj} - \frac{n_{k.} n_{.j}}{n})^2}{\frac{n_{k.} n_{.j}}{n}} = \frac{460(20 - \frac{240 \cdot 100}{460})^2}{\frac{240 \cdot 100}{460}} = 30$$

Kritický obor:  $W = \chi_{1-\alpha}^2$   $\chi_{0,99}^2$   $\chi_{0,99}^2$   $663,5$

Protože  $K \leq W$ ,  $H_0$  zamítáme na asymptotické hladině významnosti 0,01. S rizikem omylu nejvýše 0,01 jsme tedy prokázali, že výskyt chřipky v očkované a kontrolní skupině se liší.

Nyní provedeme výpočet pomocí statistiky  $T_0 = \frac{M_1 - M_2}{\sqrt{M_1(1-M_1) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$ , která se v případě platnosti nulové hypotézy

asymptoticky řídí rozložením  $N(0,1)$ .

Přitom očkovaných bylo 240, z nich onemocnělo 20, neočkovaných bylo 220, z nich onemocnělo 80.

V našem případě tedy  $n_1 = 240$ ,  $n_2 = 220$ ,  $m_1 = \frac{20}{240} = \frac{1}{12}$ ,  $m_2 = \frac{80}{220} = \frac{2}{11}$ ,  $m = \frac{20+80}{240+220} = \frac{100}{460} = \frac{5}{23}$ .

Ověření podmínek  $n_1 \hat{q} (1-\hat{q}) > 9$  a  $n_2 \hat{q} (1-\hat{q}) > 9$ : Parametry  $q$  a  $\hat{q}$  neznáme, nahradíme je odhady  $m_1$  a  $m_2$ , tedy

$$20 \cdot (1 - 20/240) = 18,333 > 9, \quad 80 \cdot (1 - 80/220) = 50,909 > 9.$$

Realizace testového kritéria:

$$t_0 = \frac{m_1 - m_2}{\sqrt{m \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{\frac{1}{12} - \frac{2}{11}}{\sqrt{\frac{5}{23} \left( \frac{1}{240} + \frac{1}{220} \right)}} = -2,80.$$

Kritický obor je  $W = (-\infty, u_{1/2}) \cup (u_{1/2}, \infty) = (-\infty, u_{0,995}) \cup (u_{0,995}, \infty) = (-\infty, -2,5758) \cup (2,5758, \infty)$ . Protože testové kritérium patří do kritického oboru,  $H_0$  zamítáme na asymptotické hladině významnosti 0,05.

## Podíl šancí ve čtyřpolní kontingenční tabulce

Ve čtyřpolních tabulkách používáme charakteristiku  $OR = \frac{a/c}{b/d}$ , která se nazývá výběrový **podíl šancí** (odds ratio). Považujeme ho za odhad neznámého teoretického podílu šancí  $OR = \frac{\pi_{11}}{\pi_{12}}$ . Můžeme si představit, že pokus se provádí za dvojích různých okolností a může skončit buď úspěchem nebo neúspěchem.

Výsledek pokusu	okolnosti		n <sub>j.</sub>
	I	II	
úspěch	a	b	a+b
neúspěch	c	d	c+d
n <sub>k.</sub>	a+c	b+d	n

Poměr počtu úspěchů k počtu neúspěchů (tzv. šance) za 1. okolností je  $\frac{a}{c}$ , za druhých okolností je  $\frac{b}{d}$ . Podíl šancí je tedy

$$OR = \frac{a/c}{b/d}$$

Jsou-li veličiny  $X, Y$  nezávislé, pak  $\pi_{11} = \pi_{10} \pi_{01}$ , tudíž teoretický podíl šancí  $\frac{\pi_{11}}{\pi_{12}} = \frac{\pi_{01}}{\pi_{02}}$ . Závislost veličin  $X, Y$  bude tím silnější, čím více se  $\frac{\pi_{11}}{\pi_{12}}$  bude lišit od 1. Avšak  $\frac{\pi_{11}}{\pi_{12}} \neq \frac{\pi_{01}}{\pi_{02}}$ , tedy hodnoty  $\frac{\pi_{11}}{\pi_{12}}$  jsou kolem 1 rozmístěny nesymetricky. Z tohoto důvodu raději používáme logaritmus teoretického či výběrového podílu šancí.

### Testování nezávislosti ve čtyřpolních tabulkách pomocí podílu šancí

Na asymptotické hladině významnosti  $\alpha$  testujeme hypotézu  $H_0: X, Y$  jsou stochasticky nezávislé náhodné veličiny (tj.  $I_{\alpha, \alpha}$ ) proti alternativě  $H_1: X, Y$  nejsou stochasticky nezávislé náhodné veličiny (tj.  $I_{\alpha, \alpha}$ ).

Testová statistika  $T_0 = \frac{\ln OR}{\sqrt{\frac{1}{a+b+c+d}}}$  se asymptoticky řídí rozložením  $N(0,1)$ , když nulová hypotéza platí.

Kritický obor:  $W = (-\infty, -u_{1-\alpha/2}) \cup (u_{1-\alpha/2}, \infty)$ .

Nulovou hypotézu tedy zamítáme na asymptotické hladině významnosti  $\alpha$ , když se testová statistika realizuje v kritickém oboru  $W$ .

Testování nezávislosti lze provést též pomocí  $100(1-\alpha)\%$  asymptotického intervalu spolehlivosti pro logaritmus podílu šancí  $\ln OR$ , který je dán vzorcem:

$$CI = \left( \ln OR - \frac{1}{\sqrt{\frac{1}{a+b+c+d}}} u_{1-\alpha/2}, \ln OR + \frac{1}{\sqrt{\frac{1}{a+b+c+d}}} u_{1-\alpha/2} \right)$$

Jestliže interval spolehlivosti neobsahuje 0, pak hypotézu o nezávislosti zamítneme na asymptotické hladině významnosti  $\alpha$ .

**Příklad (testování nezávislosti pomocí podílu šancí a pomocí statistiky K):**

U 135 uchazečů o studium na jistou fakultu byl hodnocen dojem, jakým zapůsobili na komisi u ústní přijímací zkoušky. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že přijetí na fakultu nezávisí na dojmu u přijímací zkoušky.

přijetí	dojem		n <sub>j</sub>
	dobry	špatny	
ano	17	11	28
ne	39	58	97
n <sub>k</sub>	56	69	125

**Řešení:**

**a) Testování pomocí podílu šancí:**

$OR = \frac{17}{39} \cdot \frac{58}{11} = 2,29$ . Podíl šancí nám říká, že uchazeč, který zapůsobil na komisi dobrým dojmem, má asi 2,3 x větší šanci na přijetí než uchazeč, který zapůsobil špatným dojmem.

Provedeme další pomocné výpočty:

$$\ln OR = \ln 2,29 = 0,832,$$

$$\sqrt{\frac{1}{a+b+c} + \frac{1}{d+e+f} + \frac{1}{g+h+i}} = \sqrt{\frac{1}{17+11+39} + \frac{1}{11+58+97}} = 0,139,$$

Dosadíme do vzorců pro meze asymptotického intervalu spolehlivosti pro podíl šancí:

$$\ln d = \ln OR - \frac{1}{\sqrt{\frac{1}{a+b+c} + \frac{1}{d+e+f} + \frac{1}{g+h+i}}} = 0,832 - \frac{1}{0,139} = -0,28$$

$$\ln h = \ln OR + \frac{1}{\sqrt{\frac{1}{a+b+c} + \frac{1}{d+e+f} + \frac{1}{g+h+i}}} = 0,832 + \frac{1}{0,139} = 1,692$$

Protože interval (-0,028; 1,692) obsahuje číslo 0, na asymptotické hladině významnosti 0,05 nezamítáme hypotézu o nezávislosti dojmu u přijímací zkoušky a přijetí na fakultu.

b) Testování pomocí statistiky K:

přijetí	dojem		n <sub>j</sub>
	dobrý	špatný	
ano	17	11	28
ne	39	58	97
n <sub>k</sub>	56	69	125

Ověříme splnění podmínek dobré aproximace:

$$\frac{n_1 n_{11}}{n} = \frac{28 \cdot 17}{125} = 3,808, \quad \frac{n_1 n_{12}}{n} = \frac{28 \cdot 11}{125} = 2,464,$$

$$\frac{n_2 n_{21}}{n} = \frac{97 \cdot 17}{125} = 13,408, \quad \frac{n_2 n_{22}}{n} = \frac{97 \cdot 11}{125} = 8,528$$

Podmínky dobré aproximace jsou splněny.

Dosadíme do zjednodušeného vzorce pro testovou statistiku K:

$$K = \frac{n \sum_{j=1}^2 \sum_{k=1}^2 \frac{f_{kj}^2}{n_{kj}} - \sum_{k=1}^2 n_{k.}^2}{n \sum_{k=1}^2 n_{k.} - \sum_{k=1}^2 n_{k.}^2} = \frac{12 \cdot 1758 - 39^2}{28 \cdot 125 - 97^2} = 0,595$$

Kritický obor:  $W = \chi_{0,95,1, \infty}^2 = 3,841$ .

Protože testová statistika se nerealizuje k kritickému oboru, nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05.

Vypočteme ještě Cramérův koeficient:  $V = \sqrt{\frac{K}{n-1}} = \sqrt{\frac{3,6953}{124}} = 0,171$

Vidíme, že mezi dojmem u přijímací zkoušky a přijetím na fakultu je pouze slabá závislost.

### **Poznámka k jednostranným alternativám:**

Nulová hypotéza tvrdí, že podíl šancí je roven 1, tj.  $H_0: o_p = 1$ .

Pokud víme, že za prvních okolností je šance na úspěch vyšší než za druhých okolností, pak proti nulové hypotéze postavíme pravostrannou alternativu

$H_1: o_p > 1$ .

Nulovou hypotézu zamítáme na asymptotické hladině významnosti  $\alpha$  ve prospěch pravostranné alternativy, když 100(1- $\alpha$ )% empirický asymptotický jednostranný interval spolehlivosti pro  $\ln o_p$  neobsahuje číslo 0.

Pokud víme, že za prvních okolností je šance na úspěch nižší než za druhých okolností, pak proti nulové hypotéze postavíme levostrannou alternativu

$H_1: o_p < 1$ .

Nulovou hypotézu zamítáme na asymptotické hladině významnosti  $\alpha$  ve prospěch levostranné alternativy, když 100(1- $\alpha$ )% empirický asymptotický jednostranný interval spolehlivosti pro  $\ln o_p$  neobsahuje číslo 0.

Pokud jsou šance na úspěch stejné za prvních i druhých okolností, pak proti nulové hypotéze postavíme oboustrannou alternativu

$H_1: o_p \neq 1$ .

Nulovou hypotézu zamítáme na asymptotické hladině významnosti  $\alpha$  ve prospěch oboustranné alternativy, když 100(1- $\alpha$ )% empirický asymptotický oboustranný interval spolehlivosti pro  $\ln o_p$  neobsahuje číslo 0.

**Příklad:** U 24 žáků 6. třídy základní školy bylo zjišťováno, zda jsou úspěšní v matematice (tj. mají na posledním vysvědčení známku 1 nebo 2 z matematiky) a zda hrají na nějaký hudební nástroj. Z 10 úspěšných matematiků 6 hrálo na nějaký hudební nástroj, kdežto ve skupině neúspěšných matematiků hrál pouze 1 žák na hudební nástroj. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že úspěch v matematice a hra na hudební nástroj jsou nezávislé veličiny. Proti nulové hypotéze postavte

- oboustrannou alternativu, tj. tvrzení, úspěch v matematice a hra na hudební nástroj spolu souvisí,
- pravostrannou alternativu, tj. tvrzení, že šance na úspěch v matematice jsou vyšší pro žáky, kteří hrají na nějaký hudební nástroj,
- levostrannou alternativu, tj. tvrzení, že šance na úspěch v matematice jsou nižší pro žáky, kteří hrají na nějaký hudební nástroj.

### Řešení:

Máme kontingenční tabulku

úspěch v M	hra na hudební nástroj		n <sub>j</sub>
	ano	ne	
ano	6	4	10
ne	1	13	14
n <sub>k</sub>	7	17	24

Vypočteme podíl šancí:  $O = \frac{6}{10} = 0,6$  a  $C = \frac{1}{14} = 0,0714$ . Podíl šancí nám říká, že žák, který hraje na nějaký hudební nástroj, má 19,5 x větší šanci na úspěch v matematice než žák, který nehraje na žádný hudební nástroj.



Ad a)

Pro testování nulové hypotézy proti oboustranné alternativě sestojíme oboustranný interval spolehlivosti:

Dolní a horní mez intervalu spolehlivosti pro  $\rho$  zjistíme pomocí STATISTIKY. Vytvoříme datový soubor o dvou proměnných DM a HM a jednom případě. Do Dlouhého jména proměnné DM napíšeme vzorec pro dolní mez:

$=\log(19,5)-\text{sqrt}(1/6+1/4+1/1+1/13)*\text{VNormal}(0,975;0;1)$

a analogicky do Dlouhého jména proměnné HM napíšeme vzorec pro horní mez:

$=\log(19,5)+\text{sqrt}(1/6+1/4+1/1+1/13)*\text{VNormal}(0,975;0;1)$

	1	2
	DM	HM
1	0,575	5,365

Vidíme, že  $0,575093 < \ln \rho < 5,365736$  s pravděpodobností aspoň 0,95. Protože tento interval neobsahuje 0, nulovou hypotézu zamítáme na asymptotické hladině významnosti 0,05 ve prospěch oboustranné alternativy. S rizikem omylu nejvýše 5% se tedy prokázalo, že úspěch v matematice souvisí s hrou na hudební nástroj.

Ad b)

Pro testování nulové hypotézy proti pravostranné alternativě sestrojíme levostranný interval spolehlivosti:

Do Dlouhého jména proměnné DM napíšeme vzorec pro dolní mez:

$$=\log(19,5)-\text{sqrt}(1/6+1/4+1/1+1/13)*\text{VNormal}(0,95;0;1)$$

	1 DM
1	0,960

Protože interval  $(0,960198; \infty)$  neobsahuje 0, nulovou hypotézu zamítáme na asymptotické hladině významnosti 0,05 ve prospěch pravostranné alternativy. S rizikem omylu nejvýše 5% se tedy prokázalo, že žáci, kteří hrají na nějaký hudební nástroj, mají vyšší šance na úspěch v matematice.

Ad c)

Pro testování nulové hypotézy proti levostranné alternativě sestrojíme pravostranný interval spolehlivosti:

Do Dlouhého jména proměnné HM napíšeme vzorec pro dolní mez:

$$=\log(19,5)+\text{sqrt}(1/6+1/4+1/1+1/13)*\text{VNormal}(0,95;0;1)$$

	1 HM
1	4,980

Protože interval  $(-\infty; 4,980631)$  obsahuje 0, nulovou hypotézu nezamítáme na asymptotické hladině významnosti 0,05 ve prospěch levostranné alternativy. Neprokázalo se tedy, že žáci, kteří hrají na nějaký hudební nástroj, mají nižší šance na úspěch v matematice.

## Jednoduchá korelační analýza

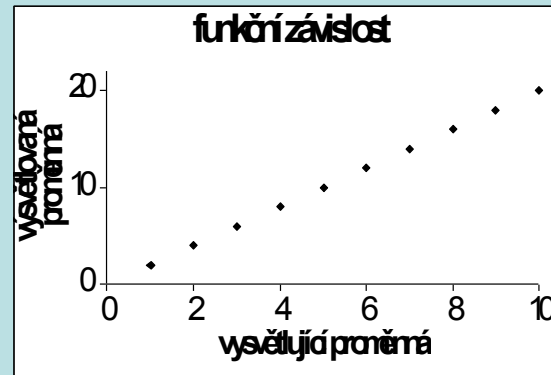
### Osnova:

- Spearmanův koeficient pořadové korelace
- testování pořadové nezávislosti
- Pearsonův koeficient korelace a výběrový koeficient korelace
- testování nezávislosti
- porovnání koeficientu korelace s danou konstantou
- porovnání dvou koeficientů korelace

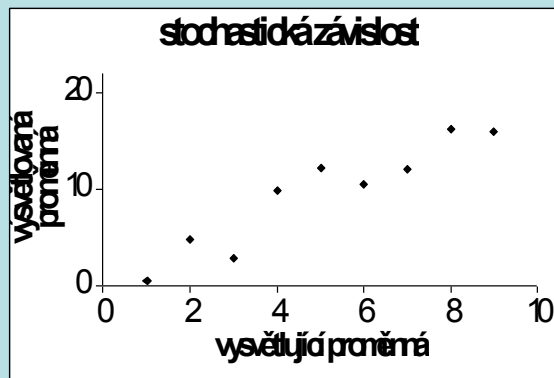
## Motivace

Uvažme náhodné veličiny  $X$ ,  $Y$ , které jsou aspoň ordinálního typu. Tyto náhodné veličiny mohou mít různý vztah:

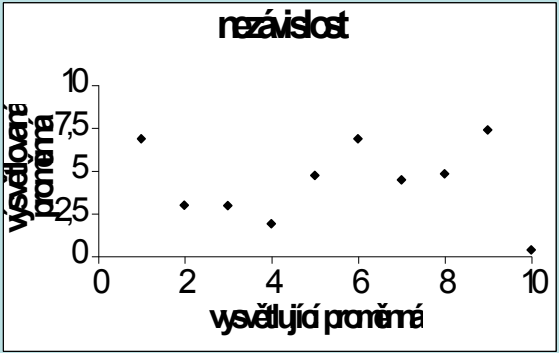
- **Deterministická (funkční) závislost:** jedna náhodná veličina je spjata s druhou náhodnou veličinou funkční závislostí vyjádřenou předpisem  $Y = g(X)$ , např.  $X$  – poloměr náhodně vybrané sériově vyráběné kuličky do kuličkových ložisek,  $Y = \frac{4}{3}\pi X^3$  – objem této kuličky. Každé realizaci náhodné veličiny  $X$  (vysvětlující proměnná) je přiřazena právě jedna realizace náhodné veličiny  $Y$  (vysvětlovaná proměnná).



- **Stochastická závislost:** jedna náhodná veličina ovlivňuje v různé míře druhou náhodnou veličinu, např. X – věk pracovníka v letech, Y – počet dnů absence za rok. Každé realizaci náhodné veličiny X může být přiřazeno více realizací náhodné veličiny Y. Závislost může být jednostranná i oboustranná.



- **Stochastická nezávislost**: náhodné veličiny se navzájem neovlivňují, např. házíme-li naráz dvěma kostkami a označíme X – počet ok padlých na jedné kostce, Y – počet ok padlých na druhé kostce, pak náhodné veličiny X, Y jsou stochasticky nezávislé.



X a Y jsou stochasticky nezávislé, když platí:  $\forall x \in \mathcal{X} : P(Y \in \mathcal{D} | X = x) = P(Y \in \mathcal{D})$

X a Y jsou nekorelované, když platí  $C(X, Y) = 0$  (tj. mezi X a Y není žádný lineární vztah).

Ze stochastické nezávislosti vyplývá nekorelovanost, avšak z nekorelovanosti nevyplývá stochastická nezávislost.

## Korelační analýza:

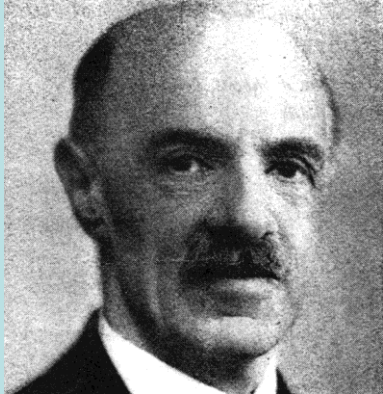
- zkoumá, zda existuje závislost mezi dvěma náhodnými veličinami  $X$ ,  $Y$ , které jsou buď ordinálního nebo intervalového či poměrového typu. **Důležité** – nelze se spokojit s formálním matematickým popisem závislosti, závislost musí být logicky zdůvodnitelná!
- pomocí Pearsonova či Spearmanova koeficientu korelace měří těsnost této závislosti
- pro náhodné veličiny intervalového a poměrového typu je založena na předpokladu, že dvourozměrný náhodný vektor

$\begin{pmatrix} X \\ Y \end{pmatrix}$  se řídí dvourozměrným normálním rozložením  $N_2\left(\begin{pmatrix} \mu_1 & \rho\sigma_1\sigma_2 \\ \mu_2 & \sigma_1\sigma_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}\right)$ , kde

$$\mu_1 = E(X), \mu_2 = E(Y), \sigma_1^2 = D(X), \sigma_2^2 = D(Y), \rho = R(X, Y)$$

- při výraznějším porušení předpokladu dvourozměrné normality doporučuje použití metod, které jsou určeny pro náhodné veličiny ordinálního typu

## Spearmanův koeficient pořadové korelace



Charles Edward Spearman (1863 – 1945): Britský psycholog a statistik, zakladatel faktorové analýzy

Nechť  $X, Y$  jsou náhodné veličiny ordinálního typu (tj. obsahová interpretace je možná jenom u relace rovnosti a relace uspořádání).

Pořídíme dvourozměrný náhodný výběr  $(X_1, Y_1), \dots, (X_n, Y_n)$  z rozložení, jímž se řídí náhodný vektor  $(X, Y)$ . Označíme  $R_i$  pořadí náhodné veličiny  $X_i$  a  $Q_i$  pořadí náhodné veličiny  $Y_i$ ,  $i = 1, \dots, n$ .

**Spearmanův koeficient pořadové korelace:** 
$$\rho_s = \frac{1}{n} \frac{\sum_{i=1}^n (R_i - Q_i)^2}{n^2 - 1}$$

Tento koeficient nabývá hodnot mezi  $-1$  a  $1$ . Čím je bližší  $1$ , tím je silnější přímá pořadová závislost mezi veličinami  $X$  a  $Y$ , čím je bližší  $-1$ , tím je silnější nepřímá pořadová závislost mezi veličinami  $X$  a  $Y$ . Teoretická hodnota Spearmanova koeficientu se značí  $\rho_s$ .



## Vlastnosti Spearmanova koeficientu pořadové korelace

Pro Spearmanův koeficient pořadové korelace platí  $-1 \leq r_s \leq 1$ . Čím je bližší 1, tím je silnější přímá pořadová závislost mezi veličinami X a Y, čím je bližší  $-1$ , tím je silnější nepřímá pořadová závislost mezi veličinami X a Y.

Je-li  $r_s \geq 0$  resp.  $r_s < 0$ , pak realizace  $(X_i, Y_i)_{i=1}^n$  daného náhodného výběru leží na nějaké rostoucí resp. klesající funkci.

Hodnoty  $r_s$  se nezmění, když provedeme vzestupnou transformaci původních dat.

Hodnoty  $r_s$  se vynásobí  $-1$ , když provedeme sestupnou transformaci původních dat.

Koeficient je symetrický.

Koeficient je rezistentní vůči odlehlým hodnotám.

Význam absolutní hodnoty Spearmanova koeficientu:

mezi 0 až 0,1 ... zanedbatelná pořadová závislost,

mezi 0,1 až 0,3 ... slabá pořadová závislost,

mezi 0,3 až 0,7 ... střední pořadová závislost,

mezi 0,7 až 1 ... silná pořadová závislost.

Spearmanův koeficient pořadové korelace se používá v situacích, kdy

- zkoumaná data mají ordinální charakter

- nelze předpokládat, že vztah mezi veličinami X, Y je lineární

- náhodný výběr nepochází z dvourozměrného normálního rozložení

## Testování nezávislosti ordinálních veličin

Na hladině významnosti  $\alpha$  testujeme hypotézu  $H_0$ : X, Y jsou pořadově nezávislé náhodné veličiny proti

- oboustranné alternativě  $H_1$ : X, Y jsou pořadově závislé náhodné veličiny
- levostranné alternativě  $H_1$ : mezi X a Y existuje nepřímá pořadová závislost
- pravostranné alternativě  $H_1$ : mezi X a Y existuje přímá pořadová závislost).

Jako testová statistika slouží Spearmanův koeficient pořadové korelace  $r_S$ .

Nulovou hypotézu zamítáme na hladině významnosti  $\alpha$  ve prospěch

- oboustranné alternativy, když  $|r_S| \geq r_{S,1-\alpha/2}(n)$
- levostranné alternativy, když  $r_S \leq -r_{S,1-\alpha}(n)$
- pravostranné alternativy, když  $r_S \geq r_{S,1-\alpha}(n)$ ,

kde  $r_{S,1-\alpha}(n)$  je kritická hodnota, kterou pro  $\alpha = 0,05$  nebo  $0,01$  a  $n \leq 30$  najdeme v tabulkách.

## Asymptotické varianty testu

Pro  $n > 20$  lze použít testovou statistiku  $T_0 = \frac{r\sqrt{n}}{s}$ , která se v případě platnosti nulové hypotézy asymptoticky řídí rozložením  $t(n-2)$ .

Kritický obor pro oboustrannou alternativu:  $W = (-\infty, -t_{1/2}(n-2)] \cup [t_{1/2}(n-2), \infty)$

Kritický obor pro levostrannou alternativu:

$$W = (-\infty, -\eta]$$

Kritický obor pro pravostrannou alternativu:

$$W = [t_{1/2}(n-2), \infty)$$

Hypotézu o pořadové nezávislosti náhodných veličin  $X, Y$  zamítáme na asymptotické hladině významnosti  $\alpha$ , když  $t_0 \in W$ .

**Upozornění:** Systém STATISTICA používá tuto variantu testu pořadové nezávislosti bez ohledu na rozsah náhodného výběru.

Pro  $n > 30$  lze použít testovou statistiku  $r_S\sqrt{n-1}$ . Platí-li  $H_0$ , pak  $r_S\sqrt{n-1} \approx N(0, 1)$ . Nulovou hypotézu tedy zamítáme na asymptotické hladině významnosti  $\alpha$  ve prospěch

oboustranné alternativy, když  $r_S\sqrt{n-1} = u_{1/2} \cup [u_{1/2}, \infty)$ ,

levostranné alternativy, když  $r_S\sqrt{n-1} = u_{1/2}$ ,

pravostranné alternativy, když  $r_S\sqrt{n-1} = u_{1/2}, \infty$

### Příklad na testování pořadové nezávislosti (jsou známa pořadí):

Dva lékaři hodnotili stav sedmi pacientů po téměř chirurgickém zákroku. Postupovali tak, že nejvyšší pořadí dostal nejtěžší případ.

Číslo pacienta	1	2	3	4	5	6	7
Hodnocení 1. lékaře	4	1	6	5	3	2	7
Hodnocení 2. lékaře	4	2	5	6	1	3	7

Vypočtete Spearmanův koeficient a na hladině významnosti 0,05 testujte hypotézu, že hodnocení obou lékařů jsou pořadově nezávislá.

#### Řešení:

Na hladině významnosti 0,05 testujeme  $H_0$ : X, Y jsou pořadově nezávislé náhodné veličiny proti oboustranné alternativě  $H_1$ : X, Y jsou pořadově závislé náhodné veličiny. V tomto příkladě přímo známe pořadí  $R_i$  (tj. hodnocení 1. lékaře) a pořadí  $Q_i$  (tj. hodnocení 2. lékaře). Vypočteme

$$r_s = \frac{6}{7} - \frac{4}{7} + \frac{4}{7} - \frac{2}{7} + \frac{2}{7} - \frac{1}{7} + \frac{1}{7} - \frac{1}{7} + \frac{1}{7} - \frac{1}{7} = 0,857$$

Kritická hodnota:  $r_{s,0,975}(7) = 0,745$ . Protože  $0,857 \geq 0,745$ , nulovou hypotézu zamítáme na hladině významnosti 0,05.

## Výpočet pomocí systému STATISTICA

Vytvoříme datový soubor o dvou proměnných X (hodnocení 1. lékaře), Y (hodnocení 2. lékaře) a sedmi případech. Do proměnných X a Y zapíšeme zjištěná hodnocení.

	1	2
X	4	2
1	4	4
2	1	2
3	6	5
4	5	6
5	3	1
6	2	3
7	1	1

Statistiky – Neparametrické statistiky – Korelace – OK – vybereme Vytvořit detailní report - Proměnné X, Y – OK – Spearmanův koef. R. Dostaneme tabulku

		Spearmanovy korelace (dva le ChD vynechány párově, Označ. korelace jsou významn		
Dvojice prom	Poc plat	Spear R	t(N-2	Urovel
X & Y	7	0,857	3,721	0,013

Spearmanův koeficient pořadové korelace nabývá hodnoty 0,857, testová statistika se realizuje hodnotou 3,721, odpovídající p-hodnota je 0,0137, tedy na asymptotické hladině významnosti 0,05 zamítáme hypotézu o pořadové nezávislosti hodnocení dvou lékařů ve prospěch oboustranné alternativy.

### Příklad na testování pořadové nezávislosti (pořadí musíme stanovit):

Jsou dány realizace náhodného výběru z dvourozměrného rozložení, kterým se řídí náhodný vektor (X,Y): (2,5 13,4), (3,4 15,2), (1,3 11,8), (5,8 13,1), (3,6 14,5). Na hladině významnosti 0,05 testujte hypotézu, že náhodné veličiny jsou pořadově nezávislé proti oboustranné alternativě.

### Řešení:

$x_i$	2,5	3,4	1,3	5,8	3,6
$y_i$	13,4	15,2	11,8	13,1	14,5
$R_i$	2	3	1	5	4
$Q_i$	3	5	1	2	4
$(R_i-Q_i)^2$	1	4	0	9	0

Testová statistika: 
$$T_S = \frac{6}{\sqrt{24}} = 0,3$$

Kritická hodnota: pro  $n = 5$  a  $\alpha = 0,05$  je kritická hodnota 0,9. Protože testová statistika se realizuje hodnotou 0,3, hypotézu o pořadové nezávislosti veličin X a Y nezamítáme na hladině významnosti 0,05.

### Výpočet pomocí systému STATISTICA

Postupujeme úplně stejně jako v předešlém případě. Výstupní tabulka má tvar:

Dvojice proměnných		Spearmanovy korelace (pořadí ChD vynechány párově, Označ. korelace jsou významné)			
		Počet platných	Spearman R	t(N-2)	Uroveň významnosti
X	& Y	5	0,300	0,544	0,623

Spearmanův koeficient pořadové korelace nabývá hodnoty 0,3, testová statistika se realizuje hodnotou 0,5447, odpovídající p-hodnota je 0,6238, tedy na asymptotické hladině významnosti 0,05 nezamítáme hypotézu o pořadové nezávislosti veličin X, Y.

## Pearsonův koeficient korelace



Karl Pearson (1857 – 1936): Britský statistik

Číslo

$$R_{XY} = \frac{\left( \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{D(X)} \cdot \sqrt{D(Y)}} \right) = \frac{C_{XY}}{\sqrt{D(X)} \sqrt{D(Y)}} \text{ pro } D(X), D(Y) > 0$$

nak

se nazývá Pearsonův koeficient korelace.

(Pro výpočet Pearsonova koeficientu korelace musíme znát simultánní distribuční funkci  $\Phi(x,y)$  v obecném případě resp. simultánní hustotu pravděpodobnosti  $\varphi(x,y)$  ve spojitém případě resp. simultánní pravděpodobnostní funkci  $\pi(x,y)$  v diskrétním případě.)

## Vlastnosti Pearsonova koeficientu korelace

a)  $R(a_1, Y) = R(X, a_2) = R(a_1, a_2) = 0$

b)  $R(a_1 + b_1X, a_2 + b_2Y) = \text{sgn}(b_1b_2) R(X, Y) = \begin{cases} R(X, Y) & \text{pro } b_1b_2 > 0 \\ -R(X, Y) & \text{pro } b_1b_2 < 0 \end{cases}$

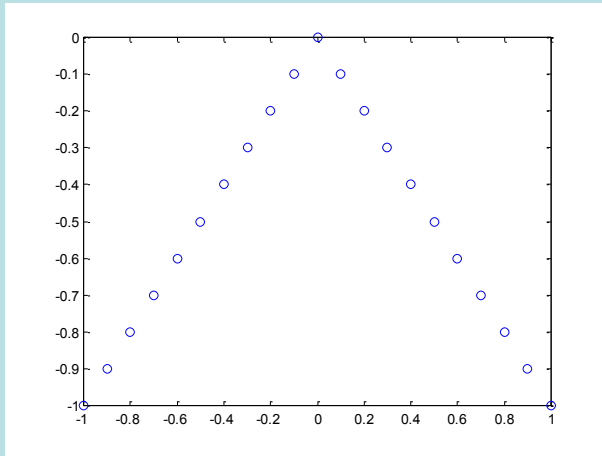
c)  $R(X, X) = 1$  pro  $D(X) \neq 0$ ,  $R(X, X) = 0$  jinak

d)  $R(X, Y) = R(Y, X)$

e)  $|R(X, Y)| \leq 1$  a rovnost nastane tehdy a jen tehdy, když mezi veličinami  $X, Y$  existuje s pravděpodobností 1 úplná lineární závislost, tj. existují konstanty  $a, b$  tak, že pravděpodobnost  $P(Y = a + bX) = 1$ . Přitom  $R(X, Y) = 1$ , když  $b > 0$  a  $R(X, Y) = -1$ , když  $b < 0$ . (Uvedená nerovnost se nazývá Cauchyova – Schwarzova – Buňakovského nerovnost.)

Z vlastností Pearsonova koeficientu korelace vyplývá, že se hodí pouze k měření těsnosti lineárního vztahu veličin  $X$  a  $Y$ . Při složitějších závislostech může dojít k paradoxní situaci, že Pearsonův koeficient korelace je nulový.

Ilustrace:





## Definice nekorelovanosti

Je-li  $R(X, Y) = 0$ , pak řekneme, že náhodné veličiny jsou **nekorelované**. (Znamená to, že mezi X a Y neexistuje žádná lineární závislost. Jsou-li náhodné veličiny X, Y stochasticky nezávislé, pak jsou samozřejmě i nekorelované.)

Je-li  $R(X, Y) > 0$ , pak řekneme, že náhodné veličiny jsou **kladně korelované**. (Znamená to, že s růstem hodnot veličiny X rostou hodnoty veličiny Y a s poklesem hodnot veličiny X klesají hodnoty veličiny Y.)

Je-li  $R(X, Y) < 0$ , pak řekneme, že náhodné veličiny **jsou záporně korelované**. (Znamená to, že s růstem hodnot veličiny X klesají hodnoty veličiny Y a s poklesem hodnot veličiny X rostou hodnoty veličiny Y.)

## Výběrový koeficient korelace

Nechť  $(X_1, Y_1), \dots, (X_n, Y_n)$  náhodný výběr rozsahu  $n$  z dvourozměrného rozložení daného distribuční funkcí  $\Phi(x, y)$ . Z tohoto dvourozměrného náhodného výběru můžeme stanovit:

výběrové průměry  $M_1 = \frac{1}{n} \sum_{i=1}^n X_i$ ,  $M_2 = \frac{1}{n} \sum_{i=1}^n Y_i$ ,

výběrové rozptyly  $S_1^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - M_1)^2$ ,  $S_2^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - M_2)^2$ ,

výběrovou kovarianci  $S_{12} = \frac{1}{n-1} \sum_{i=1}^n (X_i - M_1)(Y_i - M_2)$  a s jejich pomocí zavedeme

výběrový koeficient korelace  $R_{12} = \frac{\frac{1}{n-1} \sum_{i=1}^n (X_i - M_1)(Y_i - M_2)}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - M_1)^2} \cdot \sqrt{\frac{1}{n-1} \sum_{i=1}^n (Y_i - M_2)^2}} = \frac{S_{12}}{S_1 S_2}$  pro  $S_1 S_2 > 0$ . Vlastnosti Pearsonova koeficientu korelace se  
0jinak

přenášejí i na výběrový koeficient korelace.

(Spearmanův koeficient pořadové korelace odpovídá Pearsonovu koeficientu korelace aplikovanému na pořadí.)

### Příklad: Výpočet realizace výběrového koeficientu korelace

U 65 zaměstnanců jisté firmy byla zjišťována délka praxe v letech (veličina X) a výška prémie v Kč (veličina Y). Dvouzměrné rozložení četností je dáno kontingenční tabulkou:

x	y						
	1250	1750	2250	2750	3250	3750	4250
12,5	5	3	0	0	0	0	0
17,5	2	4	4	0	0	0	0
22,5	0	1	6	7	4	0	0
27,5	0	0	1	3	7	1	0
32,5	0	0	0	1	10	5	1

Vypočítejte realizaci  $r_{12}$  výběrového koeficientu korelace  $R_{12}$  a interpretujte jeho hodnotu. Pro úsporu času máte uvedeny následující součty:

$$\sum_{j=1}^5 n_{j.} X_j = 1562, \quad \sum_{j=1}^7 n_{.k} Y_k = 1727, \quad \sum_{j=1}^5 n_{j.}^2 = 104, \quad \sum_{k=1}^7 n_{.k}^2 = 198562,$$

$$\sum_{j=1}^5 \sum_{k=1}^7 n_{jk} X_j Y_k = 44468$$

## Řešení:

Známe tyto součty:  $\sum_{i=1}^5 x_{ij} = 1565$ ,  $\sum_{j=1}^7 x_{ij} = 727$ ,  $\sum_{i=1}^5 x_{ij}^2 = 104$ ,  $\sum_{j=1}^7 x_{ij}^2 = 98562$ ,  $\sum_{j=1}^7 \sum_{k=1}^7 n_{jk} x_{ij} y_{kj} = 44468$

Vypočteme

$$\text{průměrnou délku praxe: } m_1 = \frac{1565}{65} = 24,1$$

$$\text{průměrnou výšku prémie: } m_2 = \frac{172750}{65} = 2657,7$$

$$\text{rozptyl délky praxe: } s_1^2 = \left( \frac{1}{65} \left( 104 - \frac{1565^2}{65} \right) \right) = 2,4$$

$$\text{rozptyl výše prémie: } s_2^2 = \left( \frac{1}{65} \left( 98562 - \frac{172750^2}{65} \right) \right) = 163,4$$

$$\text{kovariance délky praxe a výše prémie: } s_{12} = \left( \frac{1}{65} \left( 44468 - \frac{1565 \cdot 172750}{65} \right) \right) = 59,4$$

$$\text{koeficient korelace délky praxe a výše prémie: } r_{12} = \frac{59,4}{\sqrt{2,4 \cdot 163,4}} = 0,370$$

Hodnota koeficientu korelace svědčí o tom, že mezi délkou praxe a výškou prémie existuje dosti silná přímá lineární závislost – čím delší praxe, tím vyšší prémie.

## Pearsonův koeficient korelace dvourozměrného normálního rozložení

Jak bylo uvedeno v motivaci, korelační analýza předpokládá, že daný náhodný výběr pochází z dvourozměrného normálního rozložení. Proč je tento předpoklad tak důležitý? Odpověď poskytne následující věta.

Nechť náhodný vektor  $(X, Y)$  má dvourozměrné normální rozložení s hustotou

$$\varphi_{X,Y}(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)} \left[ \frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right]}, \text{ přičemž } \mu_1 = E(X), \mu_2 = E(Y), \sigma_1^2 = D(X), \sigma_2^2 = D(Y), \rho = R(X,Y).$$

Marginální hustoty jsou:  $\varphi_X(x) = \int_{-\infty}^{\infty} \varphi_{X,Y}(x,y) dy = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}$ ,  $\varphi_Y(y) = \int_{-\infty}^{\infty} \varphi_{X,Y}(x,y) dx = \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}}$ .

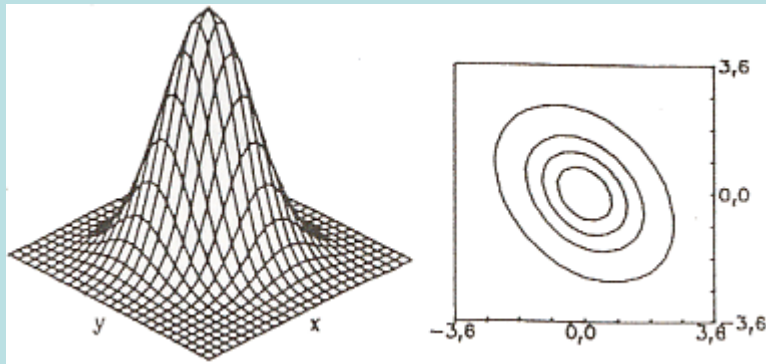
Je-li  $\rho = 0$ , pak pro  $\forall x \in \mathbb{R} : \varphi_{X,Y}(x,y) = \varphi_X(x)\varphi_Y(y)$ , tedy náhodné veličiny  $X, Y$  jsou stochasticky nezávislé. Jinými slovy: **stochastická nezávislost složek  $X, Y$  normálně rozloženého náhodného vektoru je ekvivalentní jejich nekorelovanosti**. Pro jiná dvourozměrná rozložení to neplatí!

**Upozornění:** nadále budeme předpokládat, že  $(X_1, Y_1), \dots, (X_n, Y_n)$  je náhodný výběr rozsahu  $n$  z dvourozměrného

normálního rozložení  $N_2 \left( \begin{pmatrix} \mu_1 & \rho\sigma_1\sigma_2 \\ \mu_2 & \sigma_2^2 \end{pmatrix} \right)$ .

Předpoklad dvourozměrné normality lze orientačně ověřit pomocí dvourozměrného tečkového diagramu: tečky by měly zhruba rovnoměrně vyplnit vnitřek elipsovitého obrazce. Vrstevnice hustoty dvourozměrného normálního rozložení jsou totiž elipsy:

Graf hustoty a vrstevnice dvourozměrného normálního rozložení s parametry  $\mu_1 = 0, \mu_2 = 0, \sigma_1^2 = 1, \sigma_2^2 = 1, \rho = -0,75$ :



Do dvourozměrného tečkového diagramu můžeme ještě zakreslit  $100(1-\alpha)\%$  elipsu konstantní hustoty pravděpodobnosti. Bude-li více než  $100\alpha\%$  teček ležet vně této elipsy, svědčí to o porušení dvourozměrné normality. Bude-li mít hlavní osa elipsy kladnou resp. zápornou směrnici, znamená to, že mezi veličinami  $X$  a  $Y$  existuje určitý stupeň přímé resp. nepřímé lineární závislosti.

## Testování hypotézy o nezávislosti

Na hladině významnosti  $\alpha$  testujeme  $H_0$ : X, Y jsou stochasticky nezávislé náhodné veličiny (tj.  $\rho = 0$ ) proti

- oboustranné alternativě  $H_1$ : X, Y nejsou stochasticky nezávislé náhodné veličiny (tj.  $\rho \neq 0$ )
- levostranné alternativě  $H_1$ : X, Y jsou záporně korelované náhodné veličiny (tj.  $\rho < 0$ )
- pravostranné alternativě  $H_1$ : X, Y jsou kladně korelované náhodné veličiny (tj.  $\rho > 0$ ).

Testová statistika má tvar:  $T_0 = \frac{R \cdot \sqrt{n}}{\sqrt{1 - R^2}}$ .

Platí-li nulová hypotéza, pak  $T_0 \sim t(n-2)$ .

Kritický obor pro test  $H_0$  proti

- oboustranné alternativě:  $W = (-\infty, -t_{1-\alpha/2, n-2}] \cup [t_{1-\alpha/2, n-2}, \infty)$ ,
- levostranné alternativě:  $W = (-\infty, -t_{1-\alpha, n-2})$ ,
- pravostranné alternativě:  $W = [t_{1-\alpha, n-2}, \infty)$ .

$H_0$  zamítáme na hladině významnosti  $\alpha$ , když  $t_0 \in W$ .

### Příklad: Testování hypotézy o nezávislosti proti oboustranné alternativě

V dílně pracuje 15 dělníků. Byl u nich zjištěn počet směn odpracovaných za měsíc (náhodná veličina X) a počet zhotovených výrobků (náhodná veličina Y):

X 20 21 18 17 20 18 19 21 20 14 16 19 21 15 15  
Y 92 93 83 80 91 85 82 98 90 60 73 86 96 64 81.

Předpokládejte, že data pocházejí z dvourozměrného normálního rozložení. Vypočítejte výběrový koeficient korelace mezi X a Y a na hladině 0,01 testujte hypotézu o nezávislosti X a Y proti oboustranné alternativě.

#### Řešení:

Vypočteme realizace

$$\text{výběrových průměrů: } m_1 = \frac{1}{n} \sum_{i=1}^n x_i = 18,267, m_2 = \frac{1}{n} \sum_{i=1}^n y_i = 83,6,$$

$$\text{výběrových rozptylů: } s_1^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - m_1)^2 = 5,6381, s_2^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - m_2)^2 = 121,4,$$

$$\text{výběrové kovariance: } s_{12} = \frac{1}{n-1} \sum_{i=1}^n (x_i - m_1)(y_i - m_2) = 24,2571,$$

$$\text{výběrového koeficientu korelace: } r_{12} = \frac{s_{12}}{s_1 s_2} = 0,927.$$

$$\text{Realizace testové statistiky: } t_0 = \frac{r \sqrt{n}}{\sqrt{1-r^2}} = 8,912,$$

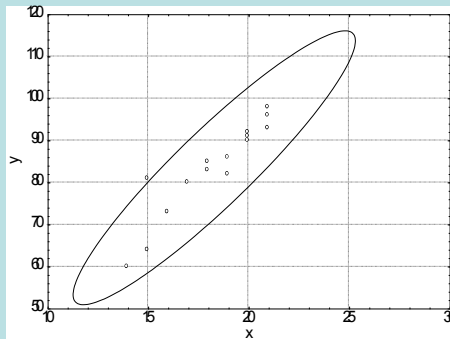
$$\text{kritický obor } W = (-\infty, -t_{0,995} | t_{0,995}, \infty) = (-\infty, -3,012 | 3,012, \infty).$$

Protože  $t_0 \in W$ , hypotézu o nezávislosti veličin X a Y zamítáme na hladině významnosti 0,01. S rizikem omylu nejvýše 1% jsme tedy prokázali, že mezi počtem směn odpracovaných za měsíc a počtem zhotovených výrobků existuje závislost.



## Výpočet pomocí systému STATISTICA

Vytvoříme datový soubor o dvou proměnných X, Y a 15 případech. Dvourozměrnou normalitu dat ověříme pomocí dvou-rozměrného tečkového diagramu: Grafy – Bodové grafy – Proměnné X, Y – OK – odškrtneme Typ proložení Lineární – na záložce Details zaškrtneme Elipsa Normální - OK.



Statistiky – Základní statistiky/tabulky – Korelační matice – OK – 1 seznam proměn. – X, Y – OK – na záložce Možnosti vybereme Zobrazit detailní tabulku výsledků – Výpočet.

Korelace (smeny a výrobky.sta) Označ. korelace jsou významné na hlad. $p < ,05000$ (Celé případy vynechány u ChD)											
Prom. \ prom.	Prum	Sm.Od	r(X,Y)	r2	t	p	N	Kons zav.:	Smerl zav.:	Kons zav.:	Smerl zav.:
X	18,26	2,374									
X	18,26	2,374	1,000	1,000			1	0,000	1,000	0,000	1,000
X	18,26	2,374									
Y	83,60	11,01	0,927	0,859	8,923	0,000	1	5,010	4,302	1,562	0,199
Y	83,60	11,01									
X	18,26	2,374	0,927	0,859	8,923	0,000	1	1,562	0,199	5,010	4,302
Y	83,60	11,01									
Y	83,60	11,01	1,000	1,000			1	0,000	1,000	0,000	1,000

Výběrový koeficient korelace se realizoval hodnotou 0,92718, testová statistika nabyla hodnoty 8,924, odpovídající p-hodnota je 0,000001, tedy na hladině významnosti 0,01 zamítáme hypotézu o nezávislosti veličin X, Y.

### Příklad: Testování hypotézy o nezávislosti proti levostranné alternativě

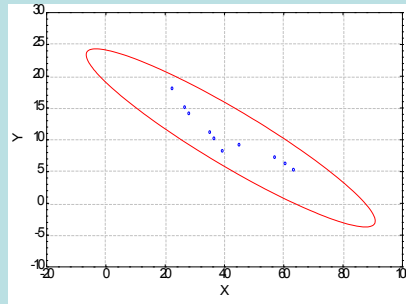
Pracovník personálního oddělení určité firmy zkoumá, zda existuje vztah mezi věkem zaměstnance (náhodná veličina X) a počtem dní absence za rok (náhodná veličina Y). Proto náhodně vybral údaje o 10 zaměstnancích:

X	27	61	37	23	46	58	29	36	64	40
Y	15	6	10	18	9	7	14	11	5	8

Na hladině významnosti 0,05 testujte hypotézu, že X a Y jsou nezávislé náhodné veličiny proti alternativě, že X, Y jsou záporně korelované náhodné veličiny.

### Řešení:

Předpoklad o dvourozměrné normalitě dat ověříme orientačně pomocí dvourozměrného tečkového diagramu.



Vzhled diagramu svědčí o tom, že předpoklad je oprávněný.

Na hladině významnosti 0,05 testujeme  $H_0: \rho = 0$  proti  $H_1: \rho < 0$ . Vypočítáme  $r_{12} = -0,9325$ , tedy mezi věkem pracovníka a počtem dnů pracovní neschopnosti existuje silná nepřímá lineární závislost.

Realizace testové statistiky:  $t_0 = \frac{r_{12} \sqrt{n}}{\sqrt{1 - r_{12}^2}} = \frac{-0,9325 \sqrt{10}}{\sqrt{1 - 0,9325^2}} = -3,59$

kritický obor  $W = (-\infty, -t_{0,05}(9)) = (-\infty, -1,833)$

Jelikož  $t_0 \in W$ , zamítáme na hladině významnosti 0,05 hypotézu o nezávislosti veličin X a Y ve prospěch levostranné alternativy. S rizikem omylu nejvýše 5% jsme prokázali, že mezi věkem pracovníka a počtem dnů absence za rok existuje nepřímá lineární závislost.

## Výpočet pomocí systému STATISTICA

Můžeme využít toho, že již známe  $r_{12}$ . Statistiky – Pravděpodobnostní kalkulátor – Korelace – vyplníme  $n = 10$ ,  $r = -0,9325$ , odškrtneme Dvojité, zaškrtneme Výpočet p z r – Výpočet. V okénku p se objeví hodnota 0,000041, tedy na hladině významnosti 0,05 zamítáme hypotézu o nezávislosti veličin X a Y ve prospěch levostranné alternativy.

Rozdělení Pearson. moment. korelačního koeficientu

N: 10

r: -0,9325

P: ,000042

Fisher. z: -1,677221

Oboustranné

Výpočet p z r

Výpočet r z p

Výpočet r ze z

Dg protokolu

Výpočet

Konec

### Příklad: Testování hypotézy o nezávislosti proti pravostranné alternativě

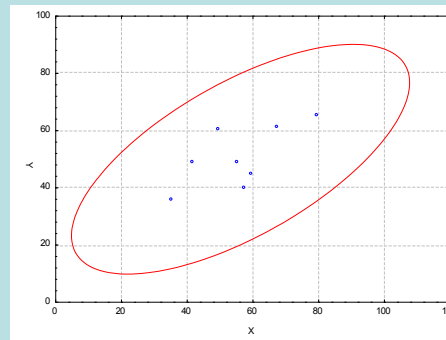
Máme k dispozici výsledky testů ze dvou předmětů zjištěné u osmi náhodně vybraných studentů určitého oboru.

Číslo studenta	1	2	3	4	5	6	7	8
Počet bodů v 1. testu	80	50	36	58	42	60	56	68
Počet bodů ve 2. testu	65	60	35	39	48	44	48	61

Na hladině významnosti 0,05 testujte hypotézu, že výsledky obou testů nejsou kladně korelované.

#### Řešení:

Nejprve se musíme přesvědčit, že uvedené výsledky lze považovat za realizace náhodného výběru z dvourozměrného normálního rozložení. Lze tak učinit orientačně pomocí dvourozměrného tečkového diagramu. Tečky by měly vytvořit elipsovitý obrazec.



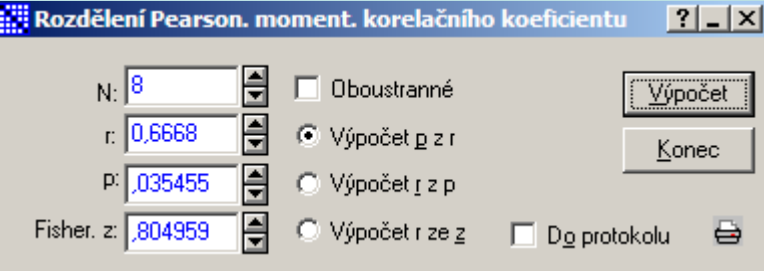
Obrázek svědčí o tom, že předpoklad dvourozměrné normality je oprávněný a že mezi počty bodů z 1. a 2. testu bude existovat určitý stupeň přímé lineární závislosti.

Na hladině významnosti 0,05 testujeme  $H_0: \rho = 0$  proti pravostranné alternativě  $H_1: \rho > 0$ .

Výpočtem zjistíme:  $r_{12} = 0,6668$ ,  $t_0 = 2,1917$ . Stanovíme kritický obor:  $W = t_{0,95}(6; \infty) = 1,9432$ . Jelikož  $t_0 \in W$ , zamítáme na hladině významnosti 0,05 hypotézu o nezávislosti veličin X a Y ve prospěch pravostranné alternativy. S rizikem omylu nejvýše 5% jsme prokázali, že mezi výsledky 1. a 2. testu existuje přímá lineární závislost.

## Výpočet pomocí systému STATISTICA

Můžeme využít toho, že již známe  $r_{12}$ . Statistika – Pravděpodobnostní kalkulátor – Korelace – vyplníme  $n = 8$ ,  $r = 0,6668$ , odškrtneme Dvojitě, zaškrtneme Výpočet p z r – Výpočet. V okénku p se objeví hodnota 0,035455, tedy na hladině významnosti 0,05 zamítáme hypotézu o nezávislosti veličin X a Y ve prospěch pravostranné alternativy.



The screenshot shows a window titled "Rozdělení Pearson. moment. korelačního koeficientu". It contains several input fields and control elements:

- N: 8
- r: 0,6668
- P: ,035455
- Fisher. z: ,804959
- Oboustranné
- Výpočet p z r
- Výpočet r z p
- Výpočet r ze z
- Dg protokolu
- Buttons: Výpočet, Konec
- Print icon

## Postup při nesplnění předpokladu dvourozměrné normality

Máme k dispozici realizace náhodného výběru rozsahu 12 z dvourozměrného rozložení:

X	1	3	4	5	6	8	10	11	13	14	16	17
Y	13	15	18	16	23	31	39	56	45	43	37	0

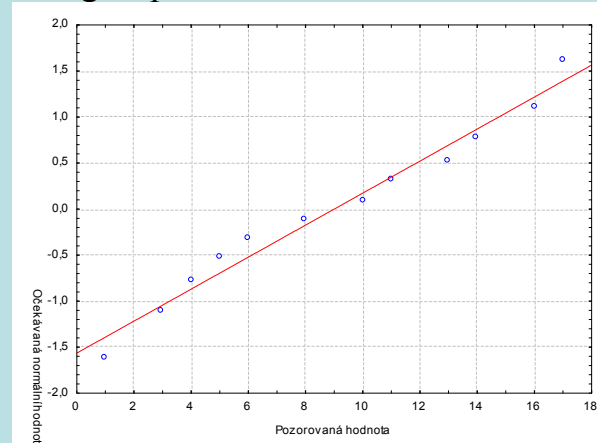
Na hladině významnosti 0,05 testujte hypotézu, že náhodné veličiny X, Y jsou nezávislé proti oboustranné alternativě.

### Řešení:

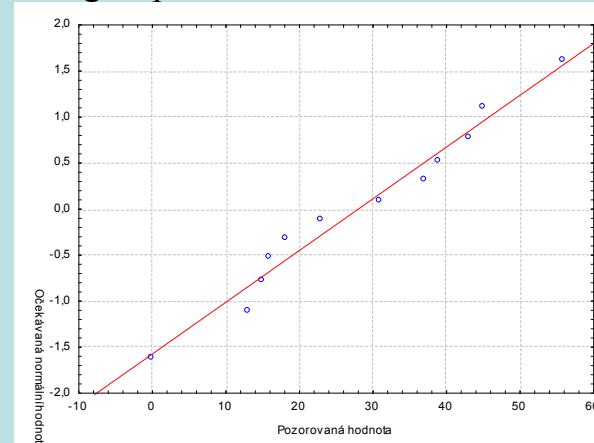
Na hladině významnosti 0,05 testujeme  $H_0: \rho = 0$  proti oboustranné alternativě  $H_1: \rho \neq 0$ . Pokud neověříme předpoklad dvourozměrné normality, obvyklým způsobem vypočteme realizaci výběrového koeficientu korelace  $r_{12} = 0,3729$  a realizaci testové statistiky  $t_0 = 1,271$ . Stanovíme kritický obor:  $W = \dots, t_{0,975} \cdot \frac{10}{\sqrt{10}} \cup t_{0,975} \cdot \frac{10}{\sqrt{10}}, 2,2281, 2,2281$ . Protože  $t_0 \notin$ , nezamítáme na hladině významnosti 0,05 hypotézu o nezávislosti náhodných veličin X a Y.

Nyní budeme testovat hypotézu o normalitě náhodné veličiny X a náhodné veličiny Y. Grafické ověření pomocí N-P grafů:

N-P graf pro veličinu X



N-P graf pro veličinu Y



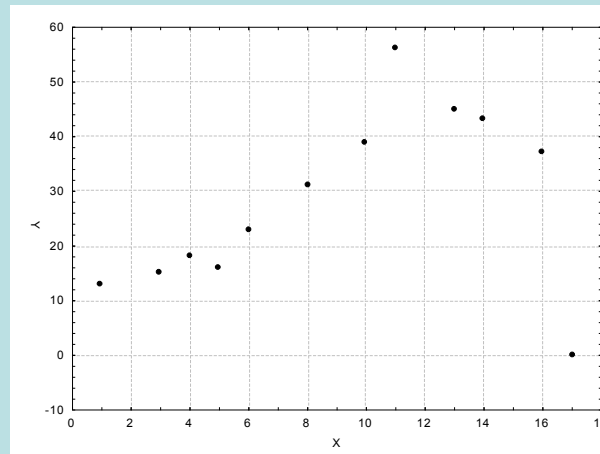
Vzhled grafů svědčí ve prospěch normality.

Testování pomocí Lilieforsovy varianty K - S testu a S – W testu:

Promě	Testy normality			
	N	max  Lilliefc	W	p
X	17	0,130	p > 0,956	0,736
Y	17	0,145	p > 0,968	0,899

V obou případech hypotézu o normalitě nezamítáme na hladině významnosti 0,05.

Ověření dvourozměrné normality pomocí dvourozměrného tečkového diagramu:



Dvourozměrná normalita je silně porušena, tečky nevyplňují vnitřek elipsovitého obrazce.

Přejdeme tedy k testování hypotézy o pořadové nezávislosti.

Testujeme hypotézu  $H_0$ : X, Y jsou pořadově nezávislé náhodné veličiny proti oboustranné alternativě  $H_1$ : X, Y jsou pořadově závislé náhodné veličiny.

Vypočítáme Spearmanův koeficient pořadové korelace.

X	1	3	4	5	6	8	10	11	13	14	16	17
Y	13	15	18	16	23	31	39	56	45	43	37	0
$R_i$	1	2	3	4	5	6	7	8	9	10	11	12
$Q_i$	2	3	5	4	6	7	9	12	11	10	8	1

$$r_s = \frac{6}{n(n^2-1)} \left[ \sum_{i=1}^n R_i Q_i - \frac{1}{n} \sum_{i=1}^n R_i \sum_{i=1}^n Q_i \right]$$

$$= \frac{6}{12(12^2-1)} \left[ 41 + \dots + 121 - \frac{1}{12} (1336) \right]$$

Stanovíme kritický obor:  $W = \dots \cup \dots \cup \dots \cup \dots \cup \dots$

Testová statistika se nerealizuje v kritickém oboru, nulovou hypotézu nezamítáme na hladině významnosti 0,05.



## Porovnání koeficientu korelace s danou konstantou

Nechť  $c$  je reálná konstanta. Testujeme  $H_0: \rho = c$  proti  $H_1: \rho \neq c$ . (Tento test se provádí např. tehdy, když experimentátor porovnává vlastnosti svých dat s vlastnostmi uváděnými v literatuře.) Test je založen na statistice

$U = \frac{1}{\sqrt{n}} \left( \frac{1}{\sqrt{1-c^2}} \left( \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) \right) - c \right) \sqrt{n}$ , která má za platnosti  $H_0$  pro  $n \geq 10$  asymptoticky rozložení  $N(0,1)$ , přičemž  $Z = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{1}{\sqrt{1-c^2}} (X_i - \bar{X})(Y_i - \bar{Y}) - c$

je tzv. **Fisherova Z-transformace**. Kritický obor pro test  $H_0$  proti oboustranné alternativě tedy je

$W = (-\infty, u_{1-\alpha/2}) \cup (u_{\alpha/2}, \infty)$ .  $H_0$  zamítáme na asymptotické hladině významnosti  $\alpha$ , když  $U \in W$ .

**Příklad:** U 600 vzorků rudy byl stanoven obsah železa dvěma analytickými metodami s výběrovým koeficientem korelace 0,85. V literatuře se uvádí, že koeficient korelace těchto dvou metod má být 0,9. Na asymptotické hladině významnosti 0,05 testujte hypotézu

$H_0: \rho = 0,9$  proti  $H_1: \rho \neq 0,9$ .

**Řešení:**  $Z = \frac{1}{\sqrt{600}} \left( \frac{1}{\sqrt{1-0,85^2}} \left( \frac{1}{600} \sum_{i=1}^{600} (X_i - \bar{X})(Y_i - \bar{Y}) \right) - 0,9 \right) \sqrt{600} = 1,97$ ,  $u_{0,975} = 1,96$ ,  $W = (-\infty, 1,96) \cup (1,96, \infty)$

Protože  $U \in W$ ,  $H_0$  zamítáme na asymptotické hladině významnosti 0,05.

## Výpočet pomocí systému STATISTICA (pouze přibližný):

Statistiky – Základní statistiky a tabulky – Testy rozdílů: r, %, průměry – OK – vybereme Rozdíl mezi dvěma korelačními koeficienty. Do políčka r1 napíšeme 0,85, do políčka N1 napíšeme 600, do políčka r2 napíšeme 0,9, do políčka N2 napíšeme 32767 (větší hodnotu systém neumožní) - Výpočet. Dostaneme p-hodnotu 0,0000, tedy zamítáme nulovou hypotézu na asymptotické hladině významnosti 0,05.

Testy rozdílů: r, %, průměry: Tabulka 11

Poslat/tisknout výsledky každ. výpočtu do okna protokolu Storno

Rozdíl mezi dvěma korelačními koeficienty

r1:  N1:   Jednostr. Výpočet  
r2:  N2:  p:   Oboustr.

Rozdíl mezi dvěma průměry (normální rozdělení)

Pr1:  SmOd1:  N1:  p:  Výpočet  
Pr2:  SmOd2:  N2:   Jednostr.  
 Výběrový průměr vs. střední hodnota  Oboustr.

Rozdíl mezi dvěma poměry

P 1:  N1:   Jednostr. Výpočet  
P 2:  N2:  p:   Oboustr.

**Upozornění:** Pokud bychom chtěli pomocí systému STATISTICA provést přesnější test s využitím statistiky U, můžeme vypočítat Fisherovu Z- transformaci pomocí Pravděpodobnostního kalkulátoru – Korelace, kde zadáme realizaci výběrového koeficientu korelace, rozsah výběru. Zajímá nás Fisher z.

## Porovnání dvou korelačních koeficientů

Nechť jsou dány dva nezávislé náhodné výběry o rozsazích  $n$  a  $n^*$  z dvourozměrných normálních rozložení s korelačními koeficienty  $\rho$  a  $\rho^*$ . Testujeme  $H_0: \rho = \rho^*$  proti  $H_1: \rho \neq \rho^*$ .

Označme  $R_{12}$  výběrový korelační koeficient 1. výběru a  $R_{12}^*$  výběrový korelační koeficient 2. výběru.

$$\text{Položme } Z = \frac{1}{\sqrt{n}} \ln \frac{1+R_{12}}{1-R_{12}} \text{ a } Z^* = \frac{1}{\sqrt{n^*}} \ln \frac{1+R_{12}^*}{1-R_{12}^*}.$$

Platí-li  $H_0$ , pak testová statistika  $U = \frac{Z - Z^*}{\sqrt{1/n + 1/n^*}}$  má asymptoticky rozložení  $N(0,1)$ .

Kritický obor pro test  $H_0$  proti oboustranné alternativě tedy je  $W = (-\infty, -u_{1-\alpha/2}) \cup (u_{1-\alpha/2}, \infty)$ .

$H_0$  zamítáme na asymptotické hladině významnosti  $\alpha$ , když  $U \in W$ .

**Příklad:** Lékařský výzkum se zabýval sledováním koncentrací látek A a B v moči pacientů trpících určitou ledvinovou chorobou. U 100 zdravých jedinců činil výběrový korelační koeficient mezi koncentracemi obou látek 0,65 a u 142 osob trpících zmíněnou chorobou byl 0,37. Na asymptotické hladině významnosti 0,05 testujte hypotézu, že korelační koeficienty v obou skupinách se neliší.

**Řešení:**  $Z = \frac{1}{\sqrt{100}} \ln \frac{1+0,65}{1-0,65} = 1,775$ ,  $Z^* = \frac{1}{\sqrt{142}} \ln \frac{1+0,37}{1-0,37} = 0,388$ ,  $U = \frac{1,775 - 0,388}{\sqrt{1/100 + 1/142}} = 1,924$ ,  $u_{0,975} = 1,96$ ,  $W = (-\infty, -1,96) \cup (1,96, \infty)$ .

Protože  $U \in W$ ,  $H_0$  zamítáme na asymptotické hladině významnosti 0,05.

## Výpočet pomocí systému STATISTICA:

Statistiky – Základní statistiky a tabulky – Testy rozdílů: r, %, průměry – OK – vybereme Rozdíl mezi dvěma korelačními koeficienty. Do políčka r1 napíšeme 0,65, do políčka N1 napíšeme 100, do políčka r2 napíšeme 0,37, do políčka N2 napíšeme 142 - Výpočet. Dostaneme p-hodnotu 0,0038, tedy zamítáme nulovou hypotézu na asymptotické hladině významnosti 0,05.

The screenshot shows the 'Testy rozdílů: r, %, průměry: smeny a vyrobky.sta' dialog box. It is divided into three sections for different types of tests. The first section, 'Rozdíl mezi dvěma korelačními koeficienty', is active and shows the following values: r1: .65, N1: 100, r2: .37, N2: 142, and a calculated p-value of .0038. The 'Oboustr.' (two-tailed) option is selected. The second section, 'Rozdíl mezi dvěma průměry (normální rozdělení)', shows Pr1: 0, SmOd1: 1, N1: 10, Pr2: 0, SmOd2: 1, N2: 10, and a p-value of 1,0000. The 'Oboustr.' option is selected. The third section, 'Rozdíl mezi dvěma poměry', shows P 1: .50000, N1: 10, P 2: .50000, N2: 10, and a p-value of 1,0000. The 'Oboustr.' option is selected. A 'Storno' button is located at the top right, and 'Výpočet' buttons are present in each section.

Testy rozdílů: r, %, průměry: smeny a vyrobky.sta

Poslat/tisknout výsledky každ. výpočtu do okna protokolu

Storno

Rozdíl mezi dvěma korelačními koeficienty

r1: .65 N1: 100 p: .0038

r2: .37 N2: 142

Jednostr.  Oboustr.

Výpočet

Rozdíl mezi dvěma průměry (normální rozdělení)

Pr1: 0 SmOd1: 1 N1: 10 p: 1,0000

Pr2: 0 SmOd2: 1 N2: 10

Jednostr.  Oboustr.

Výběrový průměr vs. střední hodnota

Výpočet

Rozdíl mezi dvěma poměry

P 1: .50000 N1: 10 p: 1,0000

P 2: .50000 N2: 10

Jednostr.  Oboustr.

Výpočet

## Interval spolehlivosti pro korelační koeficient

Jestliže dvourozměrný náhodný výběr rozsahu  $n$  pochází z dvourozměrného normálního rozložení, jehož korelační koeficient se příliš neliší od nuly (je splněna podmínka  $|\rho| < 0,5$ ) a rozsah výběru je dostatečně velký ( $n \geq 100$ ), lze odvodit, že  $100(1-\alpha)\%$  interval spolehlivosti pro  $\rho$  má meze  $R_{12} \pm \frac{1 - \frac{1}{n}}{\sqrt{n}}$ .

Nejsou-li uvedené podmínky splněny, pak nelze tento vzorec použít, protože rozložení výběrového korelačního koeficientu je příliš zešikmené. V takovém případě využijeme toho, že náhodná veličina  $Z = \frac{1}{2} \ln \frac{1 + R_{12}}{1 - R_{12}}$  má i při malém rozsahu výběru

přibližně normální rozložení se střední hodnotou  $EZ = \frac{1}{2} \ln \frac{1 + \rho}{1 - \rho}$  (2. sčítanec lze při větším  $n$  zanedbat) a rozptylem  $DZ = \frac{1}{n}$ . Standardizací veličiny  $Z$  dostaneme veličinu  $U = \frac{Z - \frac{1}{2} \ln \frac{1 + \rho}{1 - \rho}}{\sqrt{\frac{1}{n}}}$ , která má asymptoticky rozložení  $N(0,1)$ . Tudiž

$100(1-\alpha)\%$  asymptotický interval spolehlivosti pro  $\frac{1}{2} \ln \frac{1 + \rho}{1 - \rho}$  bude mít meze  $Z \pm \frac{u_{\alpha/2}}{\sqrt{n}}$ . Interval spolehlivosti pro  $\rho$  pak dostaneme zpětnou transformací.

**Poznámka:** Jelikož  $Z = \operatorname{arctgh} R_{12}$ , dostáváme  $R_{12} = \operatorname{tgh} Z$  a meze intervalu spolehlivosti pro  $\rho$  můžeme psát ve tvaru

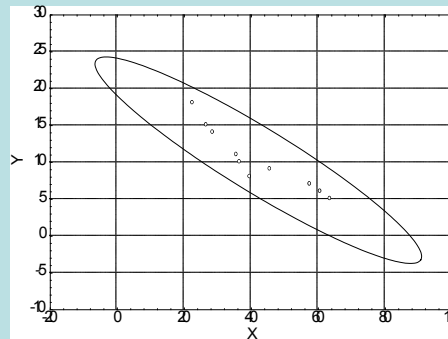
$\operatorname{tgh} \left( Z \pm \frac{u_{\alpha/2}}{\sqrt{n}} \right)$ , přičemž  $\operatorname{tgh} x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ .

**Příklad:** Pracovník personálního oddělení určité firmy zkoumá, zda existuje vztah mezi počtem dní absence za rok (veličina Y) a věkem pracovníka (veličina X). Proto náhodně vybral údaje o 10 pracovnících.

Č.prac.	1	2	3	4	5	6	7	8	9	10
X	27	61	37	23	46	58	29	36	64	40
Y	15	6	10	18	9	7	14	11	5	8

Za předpokladu, že uvedené údaje tvoří číselné realizace náhodného výběru rozsahu 10 z dvourozměrného normálního rozložení, vypočtete výběrový korelační koeficient a na hladině významnosti 0,05 testujte hypotézu, že X a Y jsou nezávislé náhodné veličiny. Sestrojte 95% asymptotický interval spolehlivosti pro skutečný korelační koeficient  $\rho$ .

**Řešení:** Předpoklad o dvourozměrné normalitě dat ověříme orientačně pomocí dvourozměrného tečkového diagramu.



Vzhled diagramu svědčí o tom, že předpoklad je oprávněný.

Testujeme  $H_0: \rho = 0$  proti  $H_1: \rho \neq 0$ . Vypočítáme  $R_{12} = -0,9325$ , tedy mezi věkem pracovníka a počtem dnů pracovní neschopnosti existuje silná nepřímá lineární závislost. Testová statistika:  $T = -7,3053$ , kvantil  $t_{0,975}(8) = 2,306$ , kritický obor  $W = (-\infty, -2,306) \cup (2,306, \infty)$ . Jelikož  $T \in W$ , zamítáme na hladině významnosti 0,05 hypotézu o nezávislosti veličin X a Y.

### Výpočet pomocí systému STATISTICA:

Ve STATISTICE vypočteme meze  $100(1-\alpha)\%$  asymptotického intervalu spolehlivosti pro koeficient korelace  $\rho$  tak, že otevřeme nový datový soubor se dvěma proměnnými (pojmenujeme je DM a HM) a jedním případem.

Do Dlouhého jména proměnné DM zapíšeme příkaz

$$= \text{TanH}(0,5 * \log((1-0,9325)/(1+0,9325)) - \text{VNormal}(0,975;0;1)/\text{sqrt}(7))$$

a do Dlouhého jména proměnné HM zapíšeme příkaz

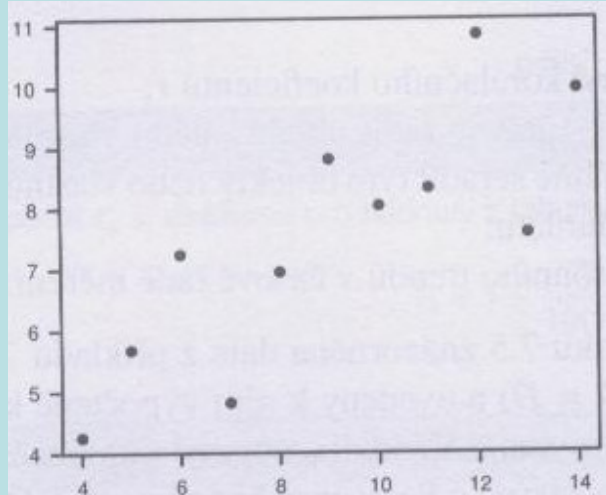
$$= \text{TanH}(0,5 * \log((1-0,9325)/(1+0,9325)) + \text{VNormal}(0,975;0;1)/\text{sqrt}(7))$$

	1	2
	DM	HM
1	-0,98	-0,73

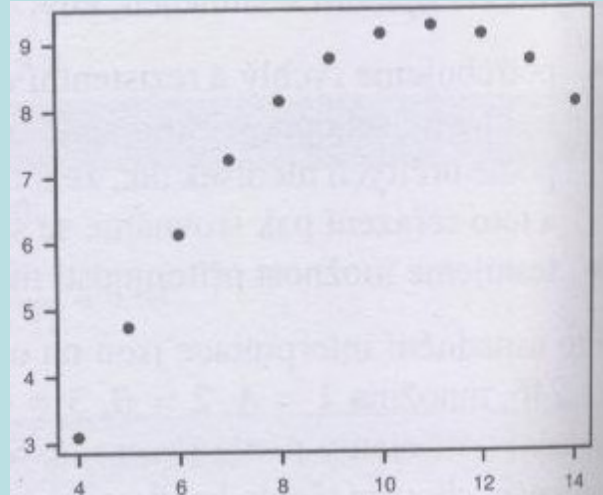
95% asymptotický interval spolehlivosti pro koeficient korelace  $\rho$  má tedy meze  $-0,98425$  a  $-0,73358$ . (Protože nepokrývá hodnotu 0, zamítáme hypotézu o nezávislosti veličin X, Y na asymptotické hladině významnosti 0,05.)

# Ilustrace vlastností Pearsonova a Spearmanova koeficientu korelace

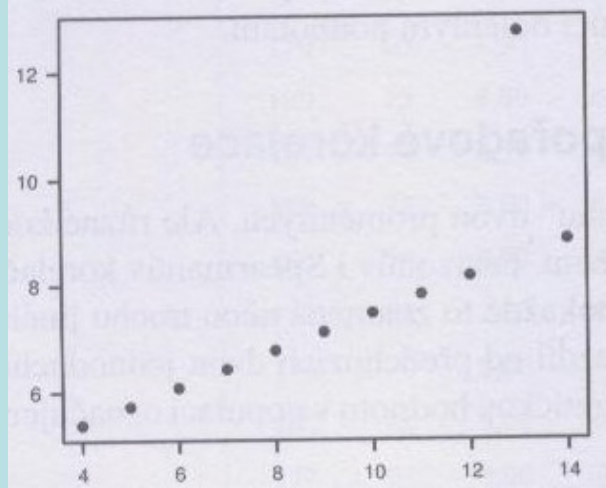
$r_{12} = 0,82, r_s = 0,82$



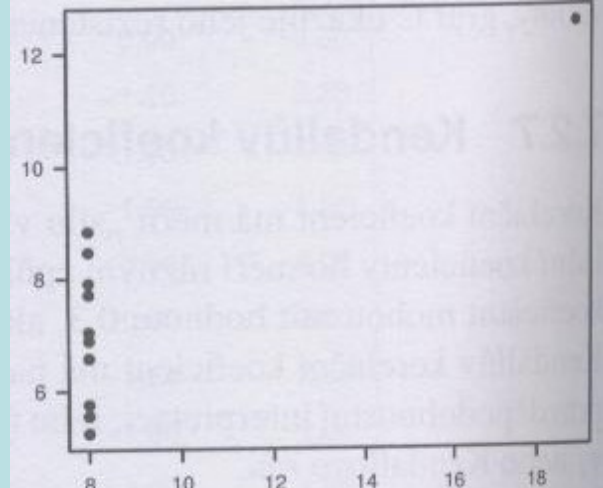
$r_{12} = 0,82, r_s = 0,69$



$r_{12} = 0,82, r_s = 0,99$



$r_{12} = 0,82, r_s = 0,5$

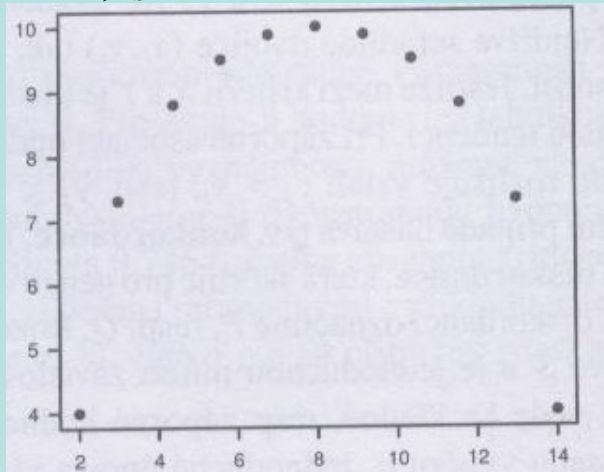


$r_{12} = 0, r_s = 0$

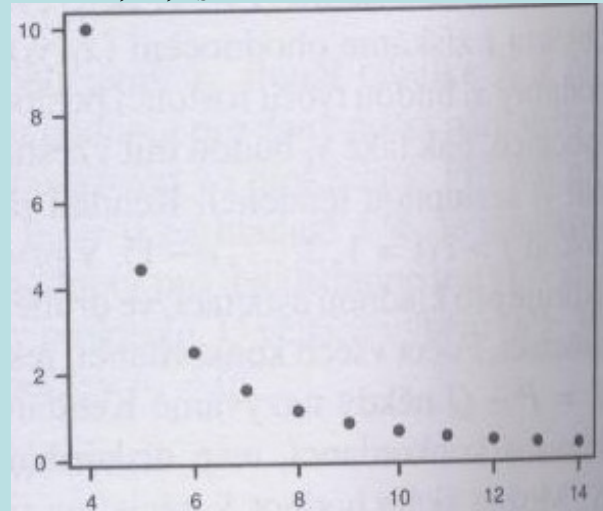
$r_{12} = -0,77, r_s = -1$



$$r_{12} = 0, r_s = 0$$



$$r_{12} = -0,77, r_s = -1$$



3. obrázek ukazuje odolnost Spearmanova koeficientu vůči odlehlým hodnotám.

6. obrázek dokumentuje schopnost Spearmanova koeficientu měřit monotónní vztahy.

## Využití modulu „Analýza síly testu“ v systému STATISTICA

Testujeme-li na hladině významnosti  $\alpha$  nulovou hypotézu (v našem případě  $H_0: \rho = 0$ ) proti alternativní hypotéze (v našem případě  $H_1: \rho \neq 0$ ), můžeme se dopustit jedné ze dvou chyb: chyba 1. druhu spočívá v tom, že  $H_0$  zamítneme, ač ve skutečnosti platí a chyba 2. druhu spočívá v tom, že  $H_0$  nezamítneme, ač ve skutečnosti neplatí.

Pravděpodobnost chyby 1. druhu se značí  $\alpha$  a nazývá se **hladina významnosti testu**.

Pravděpodobnost chyby 2. druhu se značí  $\beta$ .

Číslo  $1 - \beta$  se nazývá **síla testu** a vyjadřuje pravděpodobnost, s jakou test vypoví, že  $H_0$  neplatí.

Modul „Analýza síly testu“ nám umožní vyřešit tři úkoly:

- pro daný korelační koeficient  $\rho$  a danou hladinu významnosti  $\alpha$  stanovit, jaký musí být rozsah výběru  $n$ , aby síla testu byla aspoň rovna danému číslu  $1 - \beta$
- pro dané  $\rho$ ,  $\alpha$ ,  $n$  vypočítat sílu testu  $1 - \beta$
- pro daný výběrový koeficient korelace  $r$  a dané  $\alpha$  určit meze  $100(1 - \alpha)\%$  intervalu spolehlivosti pro  $\rho$ .

### Ad a) Stanovení rozsahu výběru

Předpokládáme, že náhodný výběr  $(X_1, Y_1), \dots, (X_n, Y_n)$  pochází z dvourozměrného normálního rozložení s koeficientem korelace  $\rho = 0,3$ . Jak velký musí být rozsah tohoto výběru, aby test  $H_0: \rho = 0$  proti  $H_1: \rho \neq 0$  měl sílu 0,8, je-li hladina významnosti  $\alpha = 0,05$ ?

Statistiky – Analýza síly testu – Výpočet velikosti vzorku – Jedna korelace, t-test – OK – Ró: 0,3, Alfa: 0,05, Požadovaná síla: 0,8 – OK – Vypočítat N.

Zjistíme, že minimální velikost výběru je 84.

### Ad b) Výpočet síly testu

Předpokládáme, že náhodný výběr  $(X_1, Y_1), \dots, (X_{25}, Y_{25})$  pochází z dvourozměrného normálního rozložení s koeficientem korelace  $\rho$ , který je neznámý. Výběrový koeficient korelace nabył hodnoty -0,56. Na hladině významnosti  $\alpha = 0,05$  testujeme  $H_0: \rho = 0$  proti  $H_1: \rho \neq 0$ . Jaká je síla testu?

Statistiky – Analýza síly testu – Výpočet síly testu - Jedna korelace, t-test – OK – Ró: -0,56, N: 25, Alfa: 0,05 – OK – Výpočetní algoritmus: zaškrtneme t-statistika – Vypočítat sílu.

Zjistíme, že síla testu je 0,8582.

### Ad c) Nalezení intervalu spolehlivosti

Předpokládáme, že náhodný výběr  $(X_1, Y_1), \dots, (X_{25}, Y_{25})$  pochází z dvourozměrného normálního rozložení s koeficientem korelace  $\rho$ , který je neznámý. Výběrový koeficient korelace nabył hodnoty -0,56. Najděte 95% interval spolehlivosti pro  $\rho$ .

Statistiky – Analýza síly testu – Odhad intervalu - Jedna korelace, t-test – OK – Pozorované R: -0,56, N: 25, Spolehlivost: 0,95 – Výpočetní algoritmus: zaškrtneme Fisherovo Z (původní) – Vypočítat.

Zjistíme, že Dolní mez = -0,7821, Horní mez = -0,2117.

## Jednoduchá lineární regrese

### Osnova:

- specifikace klasického modelu lineární regrese a jeho maticový zápis
- intervaly spolehlivosti pro regresní parametry
- celkový F-test
- dílčí t-testy
- kritéria pro posouzení vhodnosti zvolené regresní funkce
- detailní rozbor modelu regresní přímky

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

Nutnost vyřešení dvou problémů:

- a) jaký typ funkce se použije k popisu dané závislosti;
- b) 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. Připadá-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  $\beta_0, \beta_1, \dots, \beta_p$ .

**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

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

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

známých funkcích  $f_1(X), \dots, f_p(X)$ , které již neobsahují neznámé parametry, tj.  $f(X; \beta_0, \dots, \beta_p) = \beta_0 + \sum_{j=1}^p \beta_j f_j(X)$ , 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:  $Y_i = f(X_i; \beta_0, \dots, \beta_p) + \varepsilon_i$ .

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

- $E_{i|\varepsilon} = 0$  (odchylky nejsou systematické)
- $D_{i|\varepsilon} = \sigma^2$  (všechna pozorování jsou prováděna s touž přesností)
- $C_{i|\varepsilon, j} = 0$  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

$\sum_{i=1}^n (y_i - \hat{y}_i)^2$  nabývá svého minima pro  $\beta_j = b_j, j = 0, 1, \dots, p$ )

$\hat{y}_i = \hat{m}(x_i; b_0, \dots, b_p)$  - empirická regresní funkce

$\hat{y}_i = \hat{m}(x_i; b_0, \dots, b_p) = \sum_{j=0}^p b_j x_{ij}$  - 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

$s^2 = \frac{S_E}{n-2}$  - 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 ( $S_T = S_E + S_R$ )

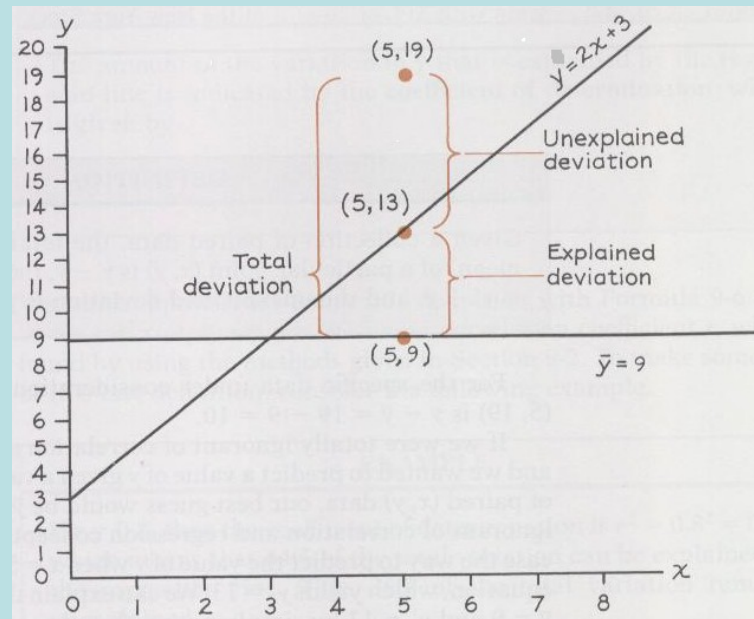
## Význam jednotlivých typů součtů čtverců

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

$\mathbf{y} = \mathbf{1}_n \beta_0 + \mathbf{X} \boldsymbol{\beta}$ , kde

$\mathbf{y} = (y_1, \dots, y_n)'$  - vektor pozorování závisle proměnné veličiny  $Y$ ,

$\mathbf{X} = \begin{pmatrix} 1 & f_1(x_1) & \dots & f_p(x_1) \\ \dots & \dots & \dots & \dots \\ 1 & f_1(x_n) & \dots & f_p(x_n) \end{pmatrix}$  - regresní matice

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

$\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$  - vektor regresních parametrů,

$\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)'$  - 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  $(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'$ ;

- 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

Sestrojte regresní matici  $X$  pro lineární regresní model

a)  $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$ , provedeme-li 4 měření,

b)  $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}$ , provedeme-li 5 měření.

**Řešení:**

$$\text{ad a) } X = \begin{pmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ 1 & X_{31} & X_{32} \\ 1 & X_{41} & X_{42} \end{pmatrix}, \text{ ad b) } X = \begin{pmatrix} 1 & X_{11} & X_{12} & X_{13} \\ 1 & X_{21} & X_{22} & X_{23} \\ 1 & X_{31} & X_{32} & X_{33} \\ 1 & X_{41} & X_{42} & X_{43} \\ 1 & X_{51} & X_{52} & X_{53} \end{pmatrix}$$

## Intervaly spolehlivosti pro regresní parametry

$S_{b_j} = \sqrt{v_{jj}}$  - směrodatná chyba odhadu  $b_j$ , kde  $v_{jj}$  je  $j$ -tý diagonální prvek matice  $(X'X)^{-1}$ .

Pro  $j = 0, 1, \dots, p$  statistika  $T_j = \frac{b_j - \beta_j}{S_{b_j}} \sim t_{n-p-1}$ , tedy  $100(1-\alpha)\%$  interval spolehlivosti pro  $\beta_j$  má meze:

$$b_j \pm t_{\alpha/2, n-p-1} \cdot S_{b_j}.$$

(S intervaly spolehlivosti souvisí relativní chyby odhadů regresních parametrů. Získají se tak, že se vypočítá absolutní hodnota podílu poloviční šířky intervalu spolehlivosti a hodnoty odhadu. Relativní chyba odhadu by neměla přesáhnout 10 %.)

### Příklad:

V tabulce jsou výnosy technické cukrovky v tunách na ha od roku 2000 do roku 2007.

i	rok	cukrovka technická
1	2000	45,83
2	2001	45,41
3	2002	49,45
4	2003	45,20
5	2004	50,34
6	2005	53,31
7	2006	51,48
8	2007	53,25

Předpokládejte, že závislost výnosu cukrovky na roku lze vyjádřit regresní přímkou  $y = \beta_0 + \beta_1 x$ .

- MNČ najděte odhady neznámých regresních parametrů  $\beta_0$ ,  $\beta_1$ .
- Sestrojte 95% intervaly spolehlivosti pro regresní parametry  $\beta_0$ ,  $\beta_1$ .
- Najděte relativní chyby odhadů regresních parametrů  $\beta_0$ ,  $\beta_1$ .

## Řešení:

Vytvoříme datový soubor se dvěma proměnnými rok, Y a osmi případy.

Získání odhadů  $b_0$ ,  $b_1$ :

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

Výsledky regrese se závislou proměnnou :						
R= ,84604287 R <sup>2</sup> = ,71578853 Upravené R <sup>2</sup>						
F(1,6)=15,111 p<,00810 Směrod. chyba odh						
N=8	b*	Sm.chy z b*	b	Sm.chy z b	t(6)	p-hoc
Abs.ci			-2312	607,4	-3,80	0,008
rok	0,846	0,217	1,1	0,30	3,88	0,008

Výpočet mezi intervalu spolehlivosti a relativních chyb odhadů:

K výstupní tabulce přidáme tři nové proměnné DM, HM a chyba.

Do Dlouhého jméne proměnné DM napíšeme

$$=v_3-v_4 \cdot V_{\text{Student}}(0,975;6)$$

Do Dlouhého jméne proměnné HM napíšeme

$$=v_3+v_4 \cdot V_{\text{Student}}(0,975;6)$$

Do Dlouhého jména proměnné chyba napíšeme

$$=100 \cdot \text{abs}(0,5 \cdot (v_8-v_7)/v_3)$$

Výsledky regrese se závislou proměnnou : Y (cukrovka, tec									
R= ,84604287 R <sup>2</sup> = ,71578853 Upravené R <sup>2</sup> = ,66841995									
F(1,6)=15,111 p<,00810 Směrod. chyba odhadu : 1,9651									
N=8	b*	Sm.chy z b*	b	Sm.chy z b	t(6)	p-hoc	DM =v <sub>3</sub> -v <sub>4</sub>	HM =v <sub>3</sub> +v <sub>4</sub>	chyb. =100*
Abs.ci			-2312	607,4	-3,80	0,008	-3798	-825	64,28
rok	0,846	0,217	1,1	0,30	3,88	0,008	0,436	1,920	62,94

S pravděpodobností 95% se bude úsek  $\beta_0$  regresní přímky nacházet v intervalu (-3798,71; -825,738). Odhad  $b_0$  úseku  $\beta_0$  je zatížen relativní chybou 64,3%.

S pravděpodobností 95% se bude směrnice  $\beta_1$  regresní přímky nacházet v intervalu (-3798,71; -825,738). Odhad  $b_1$  úseku  $\beta_1$  je zatížen relativní chybou 62,9%.

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

Na hladině významnosti  $\alpha$  testujeme

$H_0: \beta_0, \dots, \beta_p = 0, \dots, 0'$  proti  $H_1: \beta_0, \dots, \beta_p \neq 0, \dots, 0'$ .

(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.

Kritický obor:  $W_{F, p, n-p-1, \alpha}$ .

$F \geq W_{F, p, n-p-1, \alpha}$   $H_0$  zamítáme na hladině významnosti  $\alpha$ .

Výsledky F-testu zapisujeme do tabulky analýzy rozptylu:

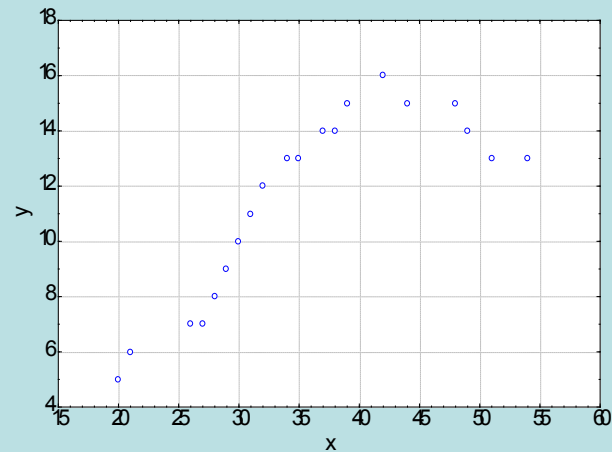
zdroj variability	součet čtverců	stupně volnosti	podíl	statistika F
model	$S_R$	$p$	$S_R/p$	$\frac{S_R/p}{S_E/(n-p-1)}$
reziduální	$S_E$	$n-p-1$	$S_E/(n-p-1)$	-
celkový	$S_T$	$n-1$	-	-

### 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

$$Y = 0,1X + 2X^2 + \dots$$

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 rok, nezávisle proměnné Y - OK – OK – Výpočet: Výsledky regrese.

Výsledky regrese se závislou proměnnou :						
R= ,95519276 R2= ,91239322 Upravené F						
F(2,17)=88,524 p<,00000 Směrod. chyba c						
N=20	b*	Sm.chy z b*	b	Sm.chy z b	t(17)	p-hoc
Abs.ci			-20,7	3,373	-6,15	0,000
X	4,526	0,548	1,56	0,189	8,256	0,000
XKV	-3,73	0,548	-0,0	0,002	-6,81	0,000

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–více násobná regrese – Detailní výsledky – ANOVA

Analýza rozptylu (prodejna					
Efekt	Souc čtverc	sv	Prům čtverc	F	p-hoc
Regre	199,8	2	99,90	88,52	0,000
Rezid	19,18	1	1,128		
Celk.	219,0				

## 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:  $T_j = \frac{b_j}{\text{Sm. chy. z } b_j}$  má rozložení  $t(n-p-1)$ , pokud  $H_0$  platí.

Kritický obor:  $W = (-\infty, -t_{\alpha/2, n-p-1}) \cup (t_{\alpha/2, n-p-1}, \infty)$ .  
 $T_j \in W \Rightarrow H_0$  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 :						
R= ,95519276 R2= ,91239322 Upravené F						
F(2,17)=88,524 p<,00000 Směrod. chyba c						
N=20	b*	Sm.chy z b*	b	Sm.chy z b	t(17)	p-hoc
Abs.ci			-20,7	3,373	-6,15	0,000
X	4,524	0,548	1,56	0,189	8,254	0,000
XKV	-3,73	0,548	-0,01	0,002	-6,81	0,000

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

$$ID = \frac{\sum_{i=1}^n \hat{y}_i^2}{\sum_{i=1}^n y_i^2} - \text{index determinace } (0 < ID < 1)$$

- 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:

$$ID_{adj}^2 = \frac{1 - \frac{ID}{n-1}}{1 - \frac{p}{n}} - \text{adjustovaný index determinace}$$

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

Výsledky regrese se závislou proměnnou :						
R= ,95519276 R2= ,91239322 Upravené F						
F(2,17)=88,524 p<,00000 Směrod. chyba c						
N=20	b*	Sm.chy z b*	b	Sm.chy z b	t(17)	p-hoc
Abs.ci			-20,7	3,373	-6,15	0,000
X	4,526	0,548	1,56	0,189	8,256	0,000
Xkv	-3,73	0,548	-0,01	0,002	-6,81	0,000

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  $F = \frac{S_R/p}{S_p/p}$  pro test významnosti modelu jako celku vyšší.

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

Výsledky regrese se závislou proměnnou :						
R= ,95519276 R2= ,91239322 Upravené F						
F(2,17)=88,524 p<,00000 Směrod. chyba c						
N=20	b*	Sm.chy z b*	b	Sm.chy z b	t(17)	p-hoc
Abs.ci			-20,7	3,373	-6,15	0,000
X	4,526	0,548	1,56	0,189	8,256	0,000
Xkv	-3,73	0,548	-0,01	0,002	-6,81	0,000

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

### c) Reziduální součet čtverců a reziduální rozptyl

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

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:  $s^2 = \frac{S_E}{n - k}$

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)					
Efekt	Součet čtverců	sv	Prům. čtverců	F	p-hod.
Regre	199,82	2	99,90	88,52	0,000
Rezid	19,18	1	1,128		
Celk.	219,0				

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

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

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

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 Statistiky – 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  $(\frac{1}{2}, \frac{3}{2})$  (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:

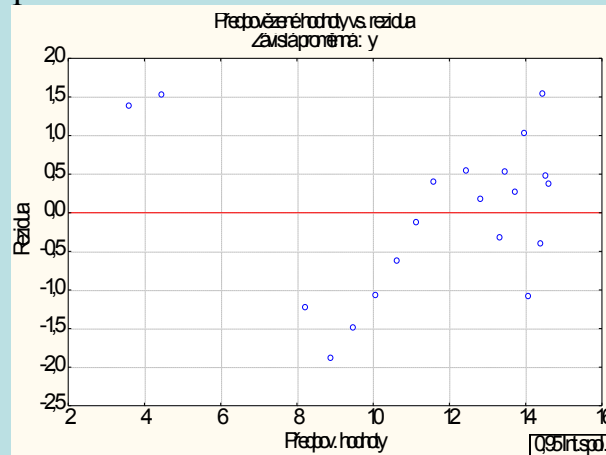
Statistiky – Vícenásobná regrese – proměnná Závislá: y, nezávislá x, xkv – OK – na záložce Residua/předpoklady/předpovědi vybereme Reziduální analýza - Detaily – Durbin-Watsonova statistika:

	Durbin Watson	Seriá korela
Odhad	0,702	0,599

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:

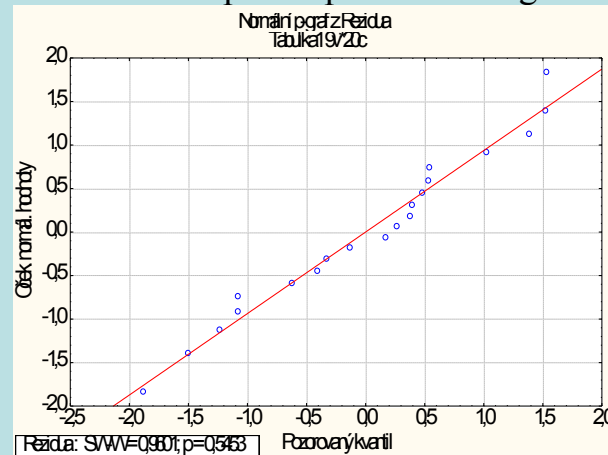
Pro proměnnou Rezidua z tabulky uložené pomocí Reziduální analýzy provedeme jednovýběrový t-test: Statistiky - Základní statistiky/tabulky – t-test, samost. vzorek – OK – proměnné Rezidua – OK.

Proměň	Průměr	Sm.od	N	Sm.chy	Referenční konstanta	t	Sv	p
Rezidua	-0,000	1,004	21	0,224	0,0	-0,000	19	1,000

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.

## Model regresní přímky

Máme regresní model  $Y = \beta_0 + \beta_1 X + \epsilon$ , kde

$\beta_0$  - teoretická regresní přímka (deterministická složka modelu).

(Parametr  $\beta_1$  interpretujeme jako teoretickou hodnotu Y při  $x = 0$  a  $\beta_1$  udává změnu Y, když X se změní o jednotku.)

Složka  $\epsilon$  - náhodná složka modelu.

### Předpoklady použití regresní přímky:

- Závislost Y na X má lineární charakter.
- Pro celý rozsah uvažovaných hodnot nezávisle proměnné X je reziduální rozptyl  $s^2$  konstantní (hovoříme o homoskedasticitě a znamená to, že variabilita hodnot závisle proměnné veličiny Y kolem regresní přímky je stejná pro všechny uvažované hodnoty nezávisle proměnné veličiny X).
- Hodnoty závisle proměnné veličiny Y mají normální rozložení pro dané hodnoty  $x_i$  a jsou stochasticky nezávislé (to souvisí s uspořádáním experimentu).

**Poznámka:** Menší odchylky od normality a homoskedasticity je možno tolerovat.



## System normálních rovnic pro regresní přímku

Uvažujeme regresní model  $Y = \beta_0 + \beta_1 X + \epsilon$ .

System normálních rovnic pro odhad regresních parametrů  $\beta_0$  a  $\beta_1$  získáme derivováním výrazu

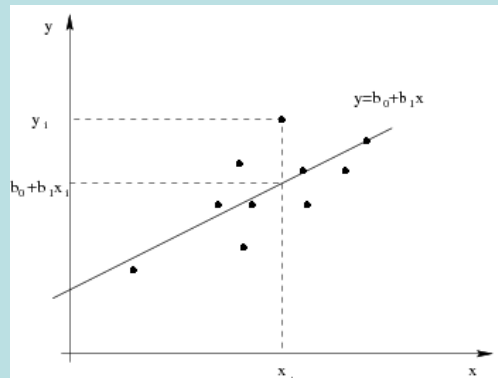
$Q(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$  parciálně podle  $\beta_0$  a  $\beta_1$ :

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} = \frac{2}{n} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) (-1) = 0, \quad \frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1} = \frac{2}{n} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) (-x_i) = 0$$

Řešením tohoto systému získáme odhady  $b_0 = \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n y_i x_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$ ,  $b_1 = \frac{r \sum_{i=1}^n y_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$

Po jednoduchých úpravách dospějeme ke tvaru  $b_1 = \frac{S_{12}}{S_1^2}$ , kde  $S_{12}$  je kovariance hodnot  $(x_i, y_i)$ ,  $i = 1, \dots, n$  a  $S_1^2$  je rozptyl

hodnot  $X_1, \dots, X_n$ . Dále dostáváme  $b_0 = \bar{y} - b_1 \bar{x}$ , tedy regresní přímku můžeme vyjádřit ve tvaru  $y = \bar{y} + \frac{S_{12}}{S_1^2} (x - \bar{x})$ .



## Index determinace regresní přímky

Kvalitu regresních modelů posuzujeme mj. pomocí indexu determinace:  $ID = \frac{S_R}{S_T}$ , kde

$S_R = \sum_{i=1}^n (\hat{y}_i - m)^2$  je regresní součet čtverců a  $S_T = \sum_{i=1}^n (y_i - m)^2$  je celkový součet čtverců.

Pro regresní přímku má regresní součet čtverců tvar:

$$S_R = \sum_{i=1}^n (\hat{y}_i - m)^2 = \sum_{i=1}^n \left( \frac{S}{T} (x_i - \bar{x}) + m - m \right)^2 = \frac{S^2}{T^2} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{S^2}{T^2} S_1^2.$$

Celkový součet čtverců  $S_T = \sum_{i=1}^n (y_i - m)^2 = S_2^2$ , tedy index determinace

$$ID = \frac{S_R}{S_T} = \frac{n \frac{S^2}{T^2} S_1^2}{S_2^2} = \frac{S^2}{T^2} \frac{S_1^2}{S_2^2} = r^2.$$

Vidíme tedy, že v případě regresní přímky **index determinace je roven kvadrátu koeficientu korelace**.

Index determinace nabývá hodnot z intervalu  $\langle 0, 1 \rangle$ . Často se vyjadřuje v procentech a informuje nás o tom, jakou část variability hodnot závisle proměnné veličiny Y vyčerpává regresní model.

## Sdružené regresní přímky

Předpokládáme, že obě veličiny  $Y$  a  $X$  jsou náhodné a veličina  $X$  nezávisí na náhodné složce  $\epsilon$ . Pak jde o případ oboustranné závislosti.

Závislost  $Y$  na  $X$  vystihuje regresní model  $Y = \alpha_0 + \alpha_1 X + \epsilon$ ,

závislost  $X$  na  $Y$  vystihuje regresní model  $X = \beta_0 + \beta_1 Y + \epsilon$ .

Odhady  $a_0, a_1$  regresních parametrů  $\alpha_0, \alpha_1$  v modelu  $Y = \alpha_0 + \alpha_1 X + \epsilon$  získáme opět MNČ ve tvaru

$$a_1 = \frac{s_{xy}}{s_x^2}, a_0 = \bar{y} - a_1 \bar{x} \quad b_1 = \frac{s_{xy}}{s_y^2}, b_0 = \bar{x} - b_1 \bar{y}$$

Empirická regresní přímka závislosti  $X$  na  $Y$  má tedy rovnici:

$$x = \bar{x} + \frac{s_{xy}}{s_y^2} (y - \bar{y})$$

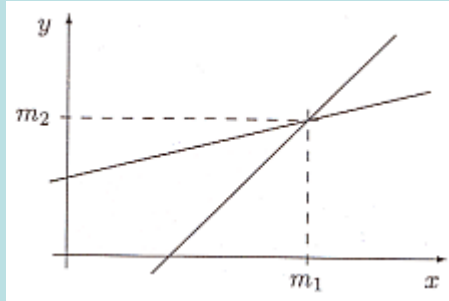
Obě empirické regresní přímky  $y = b_0 + b_1 x$ ,  $x = a_0 + a_1 y$  se nazývají **sdružené regresní přímky** a odhady regresních parametrů  $b_1, a_1$  se nazývají **odhady párově sdružených regresních parametrů**.

Je zřejmé, že  $b_1 a_1 = 1$ . Rovnice sdružených regresních přímek můžeme tedy psát ve tvaru:

$$y = \bar{y} + \frac{s_y}{s_x} (x - \bar{x}) \quad y = \bar{y} + \frac{s_y}{s_x} (x - \bar{x})$$

## Vlastnosti sdružených regresních přímek

a) Sdružené regresní přímky se protínají v bodě o souřadnicích  $m_1, m_2$  (tj. v těžišti dvourozměrného tečkového diagramu).



b) Je-li  $r_{12} = 0$  (tj. náhodné veličiny X, Y jsou nekorelované), pak sdružené regresní přímky mají rovnice  $y = \dots$ ,  $x = \dots$  (tj. jsou to kolmice rovnoběžné se souřadnými osami).

c) Je-li  $r_{12}^2 = 1$  (tj. mezi náhodnými veličinami X, Y existuje úplná lineární závislost), pak sdružené regresní přímky splnou a  $a_1 = \dots$ .

d) Je-li  $0 < r_{12}^2 < 1$ , pak sdružené regresní přímky se liší a svírají úhel, který je tím menší, čím je těsnější lineární závislost veličin X, Y.

e) Označíme-li  $\alpha$  úhel, který svírají sdružené regresní přímky, pak z předešlých úvah plyne:

$\cos \alpha = 0$  mezi X a Y neexistuje žádná lineární závislost;

$\cos \alpha = 1$  mezi X a Y existuje úplná přímá lineární závislost;

$\cos \alpha = -1$  mezi X a Y existuje úplná nepřímá lineární závislost.

### Příklad:

Z fiktivního základního souboru všech vzorků oceli odpovídajících „všem myslitelným tavnám“ bylo do laboratoře dodáno 60 vzorků a zjištěny a hodnoty proměnné  $X$  – mez plasticity a  $Y$  – mez pevnosti. Datový soubor má tvar:

154	178	83	98	73	76
133	164	106	111	77	85
58	75	92	104	47	61
145	161	85	103	68	85
94	107	112	118	137	142
113	141	98	102	44	68
86	97	103	108	92	116
121	127	99	119	141	157
119	138	104	128	155	189
112	125	107	118	136	155
85	97	98	140	82	81
41	72	97	115	136	163
96	113	105	101	72	79
45	89	71	93	66	81
99	109	39	69	42	61
51	95	122	147	113	123
101	114	33	52	42	85
160	169	78	117	133	147
87	101	114	137	153	179
88	139	125	149	85	91

- Určete regresní přímku meze pevnosti na mez plasticity.
- Zakreslete regresní přímku do dvourozměrného tečkového diagramu.
- Najděte regresní odhad meze pevnosti pro mez plasticity = 60.
- Vypočtete index determinace a interpretujte ho.
- Najděte reziduální součet čtverců a odhad rozptylu náhodných odchylek.
- Určete regresní přímku meze plasticity na mez pevnosti.
- Zakreslete regresní přímku do dvourozměrného tečkového diagramu.
- Obě regresní přímky zakreslete do téhož dvourozměrného tečkového diagramu.

## Řešení v systému STATISTICA:

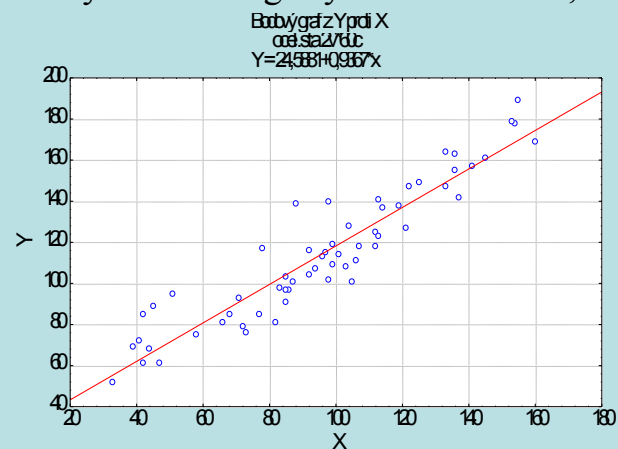
Ad a) Odhad parametrů 1. regresní přímky:

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

Výsledky regrese se závislou proměnnou						
R= ,93454811 R2= ,87338017 Upravené						
F(1,58)=400,06 p<0,0000 Směrod. chyba						
N=60	Beta	Sm.chy beta	B	Sm.chy B	t(58, Urovel	
Abs.ci			24,58	4,740	5,18,	0,000
X	0,934	0,046	0,934	0,046	20,00	0,000

Ad b) Zakreslení regresních přímky do dvourozměrného tečkového diagramu:

Grafy – Bodové grafy – Proměnné X, Y – OK – OK.



Ad c) Výpočet predikované hodnoty: Pro výpočet predikované hodnoty zvolíme Rezidua/předpoklady/předpovědi - Předpovědi závisle proměnné X: 60 OK. Ve výstupní tabulce je hledaná hodnota označena jako Předpověď: 80,79

Predpovězené hodn proměnné: Y			
Proměň	b-vař	Hodn	b-vař * Hod
X	0,936	60,00	56,20
Abs. čle			24,58
Predpoř			80,78
-95,0%L			76,25
+95,0%			85,32

Regresní odhad meze pevnosti pro mez plasticity 60 je tedy 80,8.

Ad d) Index determinace najdeme ve výstupní tabulce regrese pod označením R2:

Výsledky regrese se závislou proměnnou: Y (ocel.sta)						
R= 934581 R2= 8733017 Upravené R2= 8711970 F(1,58)=400,06 p<0,000 Šířka chybového odhadu: 11,768						
N=60	Beta	Šířka beta	B	Šířka B	t(58)	Uroveň
Abscisa			24,5884	4,74027	5,1870	0,0000
X	0,93458	0,04672	0,9368	0,04683	20,0018	0,0000

Vidíme, že variabilita meze pevnosti je regresní přímkou vyčerpána z 87,3 %.

Ad e) Reziduální součet čtverců a odhad rozptylu najdeme v tabulce ANOVA: Vrátime se do Výsledky – Vícenásobná regrese – na záložce Detailní výsledky zvolíme ANOVA (Celk. vhodnost modelu)

Analýza rozptylu (ocel.sta)					
Efekt	Součet čtverců	sv	Prům čtverců	F	p-hoc
Regre	55400	1	55400	400,0	0,000
Rezid	8031	58	138,		
Celk.	63432				

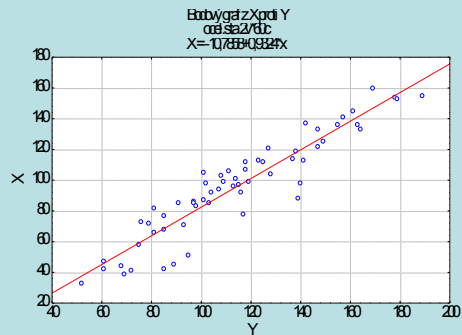
Vidíme, že reziduální součet čtverců je 8031,8 a reziduální rozptyl nabývá hodnoty 138,48.

Ad f) Výsledky pro 2. regresní přímku:

Výsledky regrese závislosti Y na X (v opačném směru)						
R <sup>2</sup> = 0,945811 R <sup>2</sup> = 0,873380 Upravené R <sup>2</sup> = 0,871970						
F(1,58) = 400,06 p < 0,0000 Střední chyba chyb: 11,741						
N=60	Beta	Směryba beta	B	Směryba B	t(58)	Uroveň
Abscisa			-10,7858	554425	-1,9454	0,0667
Y	0,9324	0,04672	0,932	0,04667	24,0716	0,0000

Vidíme, že  $x = -10,7858 + 0,9324y$ .

Ad g) Dvourozměrný tečkový diagram se zakreslenou 2. regresní přímkou

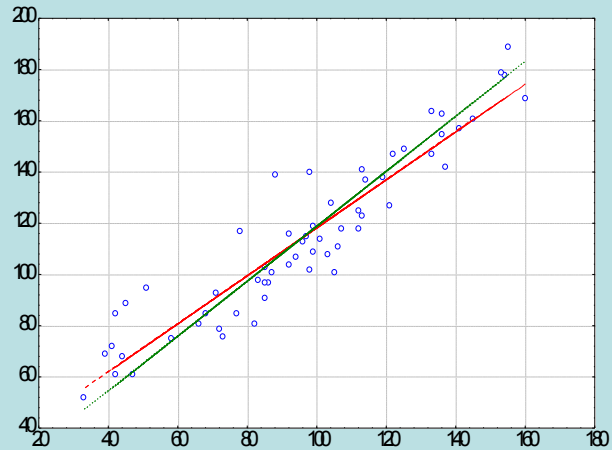




Ad h) Nakreslení sdružených regresních přímk do jednoho diagramu:

K datovému souboru ocel.sta přidáme dvě nové proměnné y1 a y2. Do proměnné y1 uložíme predikované hodnoty meze pevnosti na mezi plasticity (do Dlouhého jména proměnné y1 napíšeme  $=24,58814 + 0,93668 * x$  a do Dlouhého jména proměnné y2 napíšeme  $=(x+10,7858)/0,9324$

Grafy – Bodové grafy – zaškrtneme Vícenásobný – Proměnné X: X, Y: Y, y1, y2 – OK. Ve vytvořeném grafu pak vypneme zobrazování značek pro y1, y2 a naopak zapneme Spojnici.



Kritické hodnoty Durbinova-Watsonova testu pro autokorelaci 1. řádu pro  $\alpha = 0,05$ , rozsah výběru  $n$  a počet regresorů  $p$  (bez konstant)

n	p=1		p=2		p=3		p=4		p=5	
	d <sub>L</sub>	d <sub>U</sub>	d <sub>L</sub>	d <sub>U</sub>	d <sub>L</sub>	d <sub>U</sub>	d <sub>L</sub>	d <sub>U</sub>	d <sub>L</sub>	d <sub>U</sub>
15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83
40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79
60	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73	1,41	1,77
80	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74	1,51	1,77
100	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,57	1,78

## Úvod do analýzy časových řad

### Osnova:

- pojem časové řady
- druhy časových řad a jejich grafické znázornění
- statické a dynamické charakteristiky časové řady
- aditivní model časové řady
- odhad trendu časové řady pomocí klouzavých průměrů
- regresní analýza trendu

**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. Vypočtěte očištěné údaje.

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

$$\text{pro leden } y_1^{(0)} = 4000 \cdot \frac{365}{31} = 4741,94$$

$$\text{pro únor } y_2^{(0)} = 13000 \cdot \frac{365}{28} = 16964,29$$

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	240	3	2354,
2	210	2	2318,
3	240	3	2361,
4	242	3	2478,
5	289	3	2839,
6	339	3	3400,
7	350	3	3448,
8	350	3	3448,
9	320	3	3269,
10	300	3	3005,
11	269	3	2731,
12	260	3	2551,

## 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.

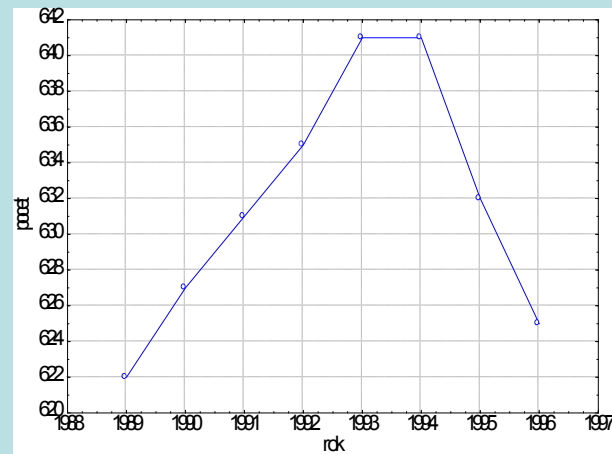
1989	1990	1991	1992	1993	1994	1995	1996
622	627	631	635	641	641	632	625

Znázorněte tuto časovou řadu graficky.

## Řešení pomocí systému STATISTICA:

Vytvoříme datový soubor o dvou proměnných nazvaných rok a počet 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.

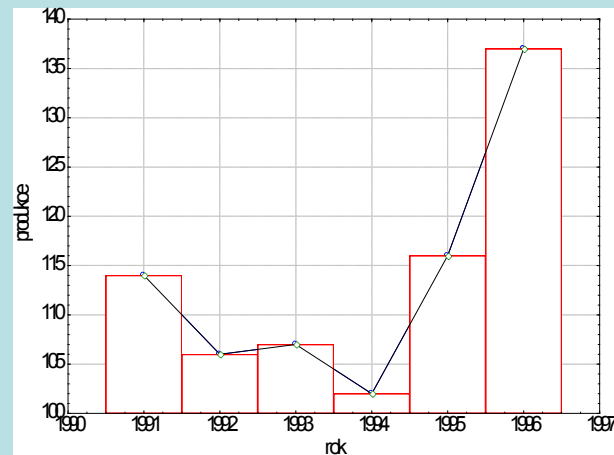
1991	1992	1993	1994	1995	1996
114	106	107	102	116	137

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

$$\bar{y} = \frac{1}{n-1} \sum_{i=2}^n \frac{y_{i-1} + y_i}{2} = \frac{1}{n-1} \left( \frac{y_1 + y_2}{2} + \frac{y_2 + y_3}{2} + \dots + \frac{y_{n-1} + y_n}{2} \right)$$

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:**  $\bar{y} = \frac{1}{10} \left( \frac{14194 + 14271}{2} + \frac{14271 + 14333}{2} + \dots + \frac{14631 + 14738}{2} \right) = 14345$



### Průměr intervalové časové řady

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

**Příklad:** Vypočítejte 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:**  $\bar{y} = \frac{1}{7} (1303,6 + \dots + 1433,8) = 1397$

## Dynamické charakteristiky časových řad

### Absolutní přírůstky

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

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

atd.

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

Průměrný absolutní přírůstek:  $\Delta \bar{y} = \frac{1}{n-1} \sum_{i=1}^n \Delta y_i$

### Relativní přírůstek

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

(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)

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

(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

$\bar{k} = \sqrt[n]{k_1 \cdot k_2 \cdot \dots \cdot k_n} = \sqrt[n]{\frac{y_n}{y_1}}$

Průměrný relativní přírůstek

$\bar{\delta} = \frac{\bar{k} - 1}{100}$

**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.

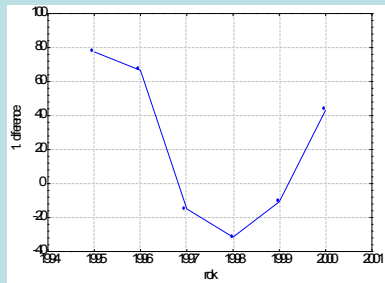
**Řešení:**

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

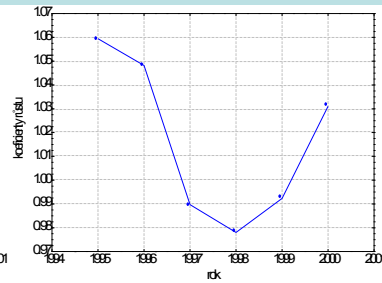
Průměrný absolutní přírůstek:  $\Delta = \frac{1433,8 - 1303,6}{7} = 17$ , tzn., že v období 1994 – 2000 rostl HDP průměrně o 21,7 miliard Kč ročně.

Průměrný koeficient růstu:  $k = \sqrt[7]{\frac{1433,8}{1303,6}} = 1,01$ , tzn., že v období 1994 – 2000 rostl HDP průměrně o 1,6% ročně.

Graf 1. diferencí:

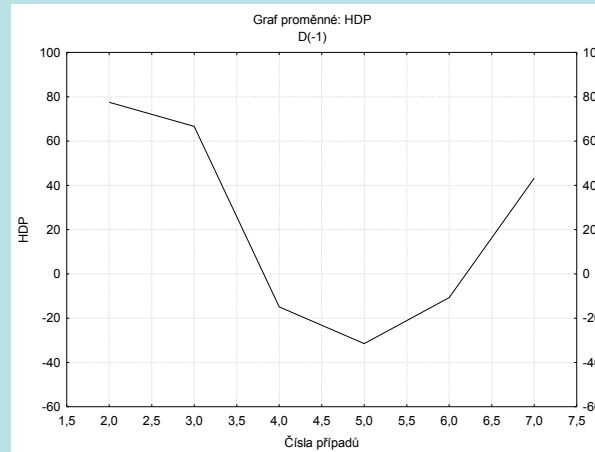


Graf koeficientů růstu:



## 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:  $\delta = \frac{X_i - X_{i-1}}{X_{i-1}}$  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:  $k_i = \frac{X_i}{X_{i-1}}$  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	3 HDP	4 RP	5 KR
1	1303,				
2	1381,	77,5	1303,	0,059	1,059
3	1447,	66,6	1381,	0,048	1,048
4	1432,	-14,9	1447,	-0,010	0,989
5	1401,	-31,5	1432,	-0,02	0,978
6	1390,	-10,7	1401,	-0,00	0,992
7	1433,	43,2	1390,	0,031	1,031
8			1433,		

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

$$\Delta = \frac{1438 - 1306}{1306} = 0,101 \text{ a } k = \frac{1438}{1306} = 1,101.$$

## 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  $t \in \{1, \dots, d\}$  a  $t \in \{n-d+1, \dots, n\}$  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 průměru (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.

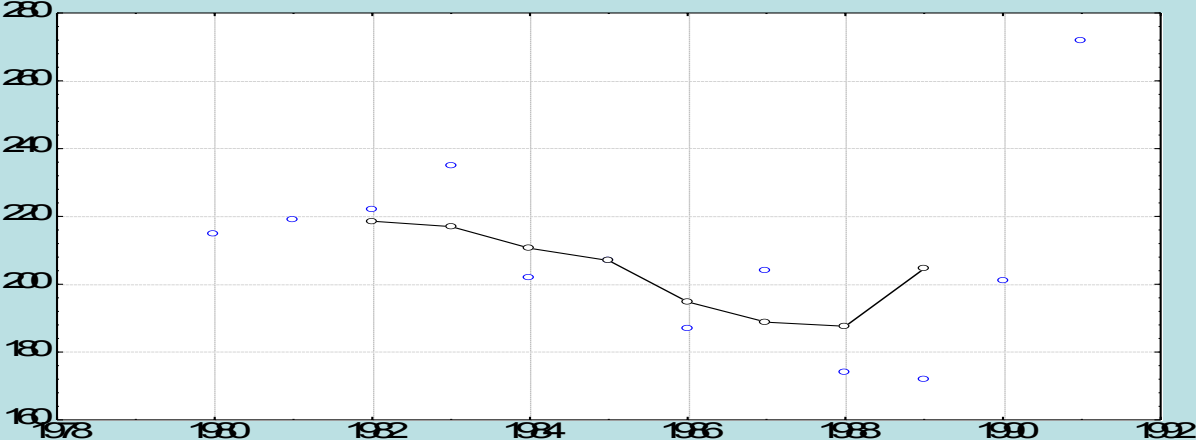
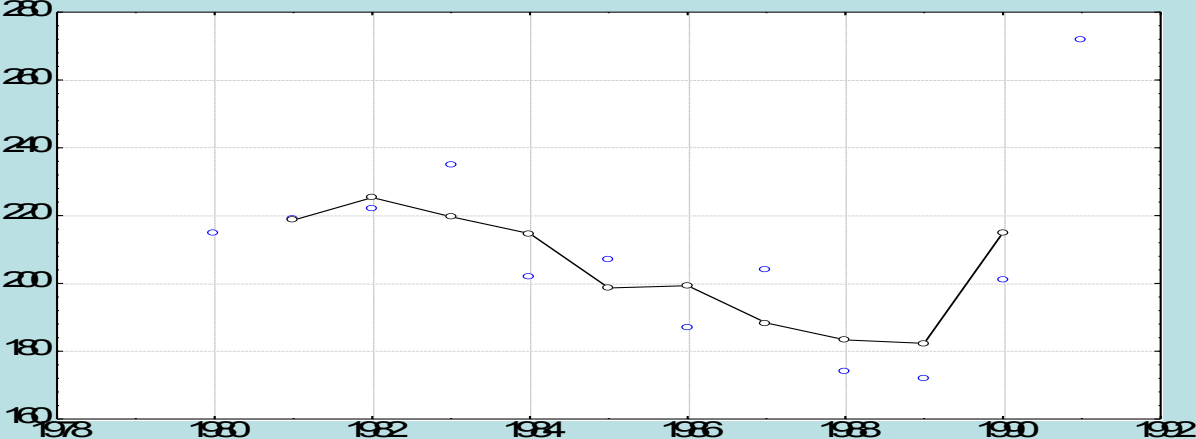
### Řešení pomocí systému STATISTICA:

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	198	215,0		
2	198	219,0	218,0	
3	198	222,0	225,0	218,0
4	198	235,0	219,0	217,0
5	198	202,0	214,0	210,0
6	198	207,0	198,0	207,0
7	198	187,0	199,0	194,0
8	198	204,0	188,0	188,0
9	198	174,0	183,0	187,0
10	198	172,0	182,0	204,0
11	199	201,0	215,0	
12	199	272,0		



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  $\beta_0, \beta_1, \dots, \beta_k$  a známých funkcích  $\varphi_0(t), \varphi_1(t), \dots, \varphi_k(t)$ , které již neobsahují žádné neznámé parametry, tj.

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

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

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

## Nejdůležitější typy trendových funkcí

Volba typu trendové funkce se provádí

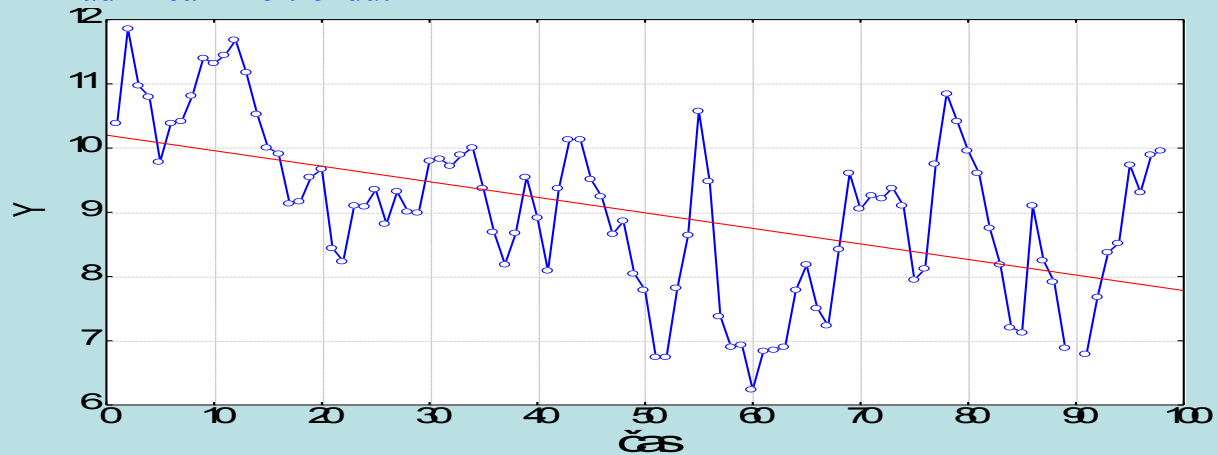
- na základě teoretických znalostí a zkušeností se zkoumanou veličinou  $Y_t$
- pomocí grafu časové řady
- pomocí informativních testů založených na jednoduchých charakteristikách časové řady

a) **Lineární trend**

Analytické vyjádření:  $f(t) = \dots$

Informativní test: 1. difference ( $\Delta = \dots$ ,  $t = \dots$ ) jsou přibližně konstantní.

Příklad lineárního trendu:

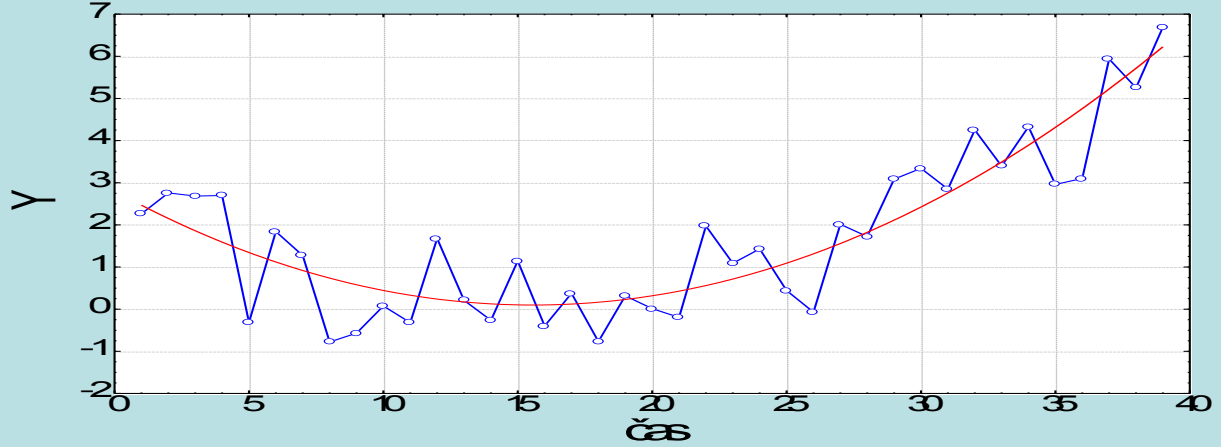


b) **Kvadratický trend**

Analytické vyjádření:  $f(t) = a \cdot t^2 + b \cdot t + c$

Informativní test: 1. diference mají přibližně lineární trend, 2. diference ( $\Delta^2 f(t) = f(t+2) - 2f(t+1) + f(t)$ ) jsou přibližně konstantní.

Příklad kvadratického trendu:

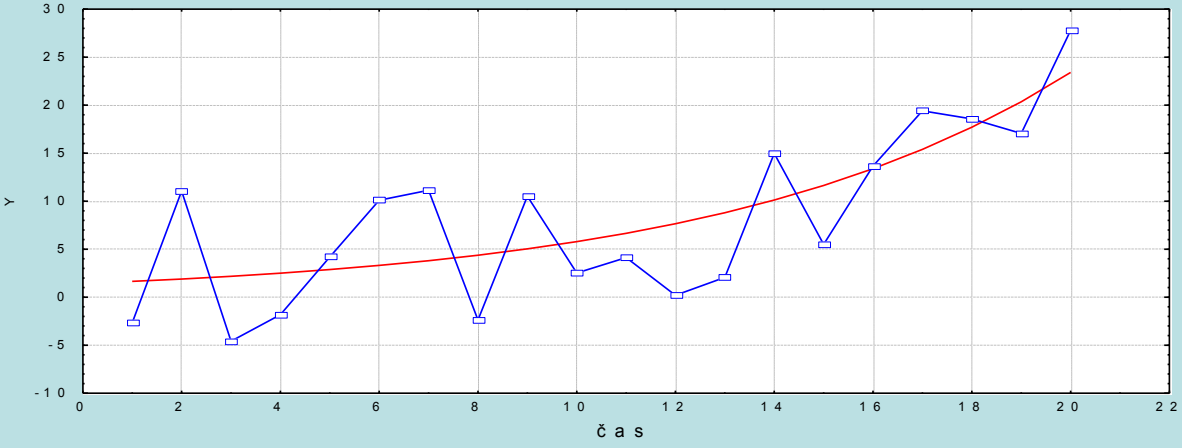


c) Exponenciální trend

Analytické vyjádření:  $f(t) = \dots$

Informativní test: koeficienty růstu ( $k_t = \dots, t = \dots, n$ ) jsou přibližně konstantní.

Příklad exponenciálního trendu:

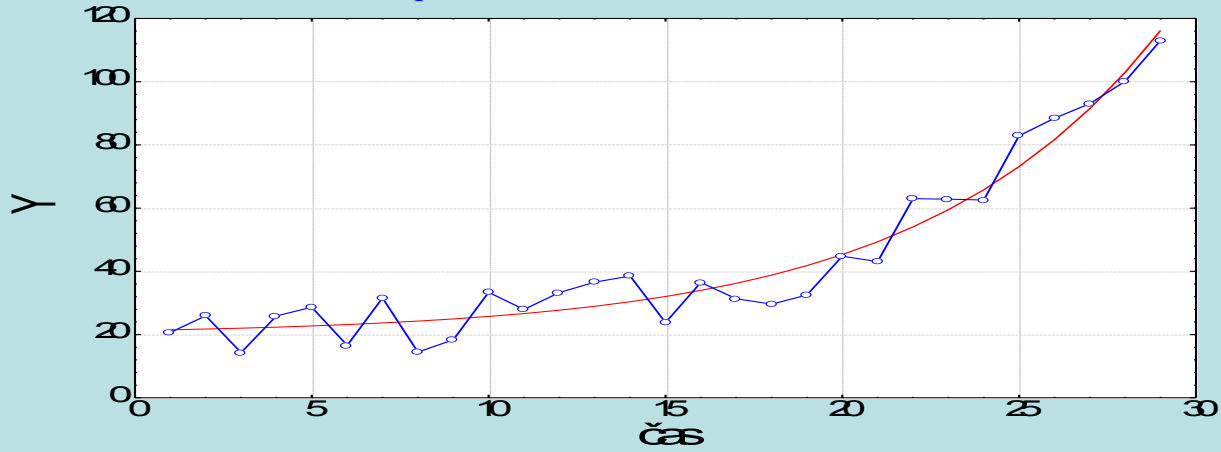


d) **Modifikovaný exponenciální trend**

Analytické vyjádření:  $f(t) =$

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

**Příklad modifikovaného exponenciálního trendu**

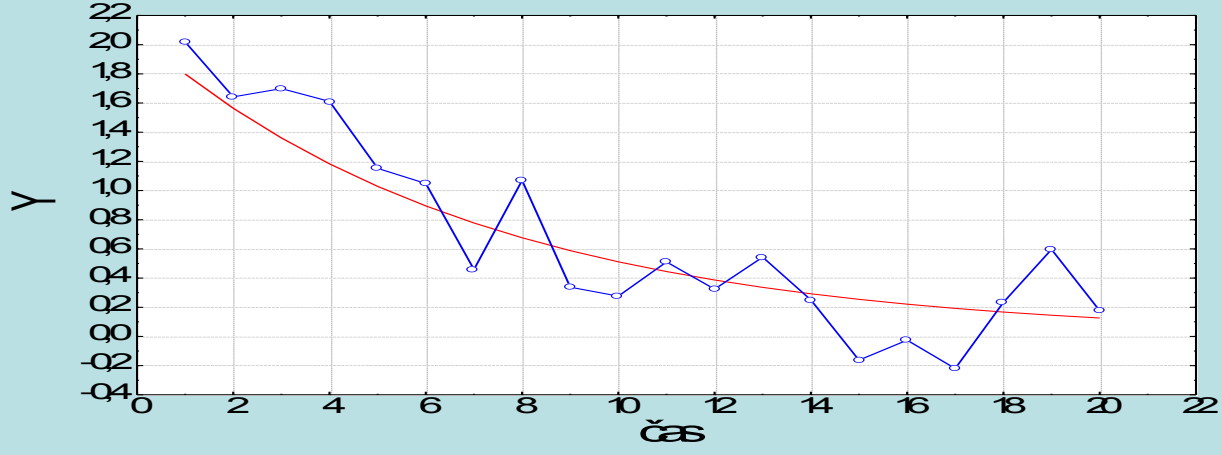


e) **Logistický trend**

Analytické vyjádření:  $f(t) = \frac{K}{1 + e^{-\alpha t}}$

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

Příklad logistického trendu:

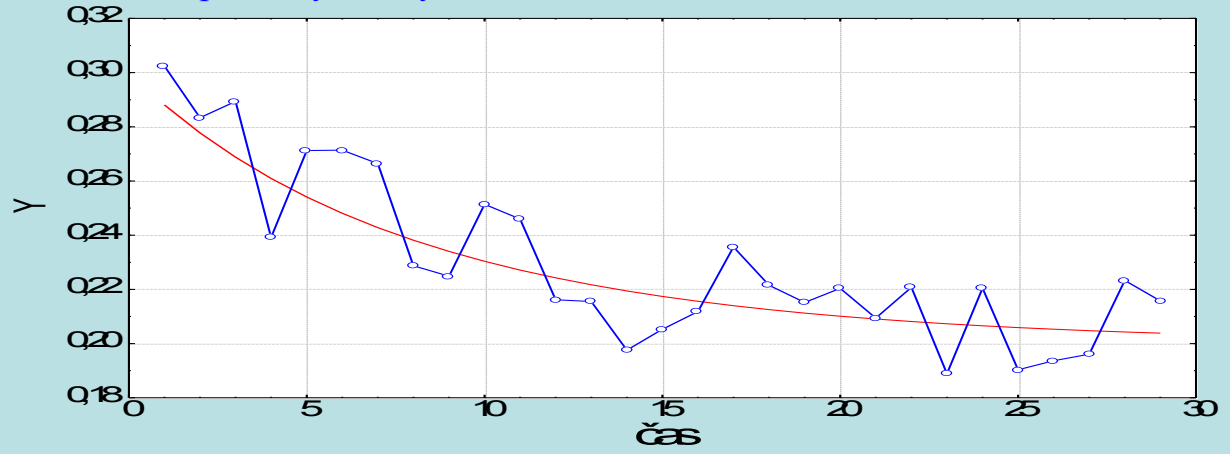


f) Gompertzova křivka

Analytické vyjádření:  $f(t) = \beta$

Informativní test: podíly  $\frac{\ln y_{t+1} - y_{t+1}}{\ln y_t - y_t}$  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  $f(t), \hat{f}(t)$ ,  $t = 1, 2, \dots, n$  by se měly řadit do přímky se směrnici 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  $f(t) =$  .

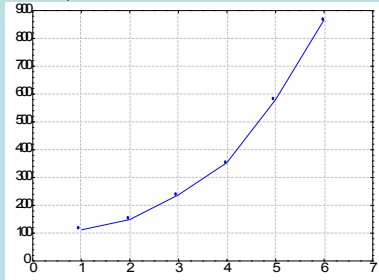
Odhadněte jeho parametry.

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

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

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

ad a)

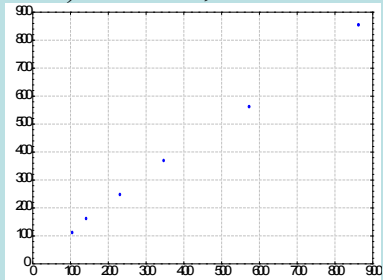


ad b) Koeficienty růstu:  $149/112 = 1,33$ ,  $238/149 = 1,597$ ,  $354/238 = 1,487$ ,  $580/354 = 1,628$ ,  $867/580 = 1,495$ . Vidíme, že koeficienty růstu jsou přibližně konstantní.

ad c) Model  $f(t) = b_0 \cdot b_1^t$  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)  $y_7 = 68,57875 \cdot 1,522265^7 = 1299,38$

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



Jak index determinace, tak graf  $f(t)$  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 : ln Y							
R= ,99801042 R2= ,99602479 Upravené R2=							
F(1,4)=1002,2 p<,00001 Směrod. chyba odhad							
N=6	b*	Sm.chy z b*	b	Sm.chy z b	t(4)	p-hoc	NPro =exp(
Abs.ci			4,227	0,051	81,79	0,000	68,57
cas	0,998	0,031	0,420	0,013	31,65	0,000	1,522

Vidíme, že  $Y = 68,57875 \cdot 1,52265^{\text{cas}}$

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 Details vybereme Proložení Exponenciální.

