

# Osnova přednášky Opakování statistických pojmů

## I. Jednorozměrný a vícerozměrný datový soubor

1. Pořízení jednorozměrného datového souboru
2. Bodové rozložení četností
3. Intervalové rozložení četností
4. Číselné charakteristiky datového souboru
5. Diagnostické grafy

## II. Úvod do testování hypotéz

1. Nulová a alternativní hypotéza
2. Chyba 1. a 2. druhu
3. Tři způsoby testování hypotéz
4. Testy normality dat

## 1. Pořízení jednorozměrného a vícerozměrného datového souboru

**Jednorozměrný soubor:** Na množině objektů  $\{\varepsilon_1, \dots, \varepsilon_n\}$  zjišťujeme hodnoty znaku  $X$  (např. u 6 domácností zjišťujeme počet členů).

Hodnotu znaku  $X$  na objektu  $\varepsilon_i$  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  $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ .

**Vícerozměrný datový soubor:** Vzniká v situacích, kdy na  $n$  objektech sledujeme hodnoty  $p$  znaků  $(X_1, \dots, X_p)^T$ .

Má tvar matice  $n \times p$ :

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

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).

## 2. 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  $x_{[j]}$

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

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

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

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

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

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

**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

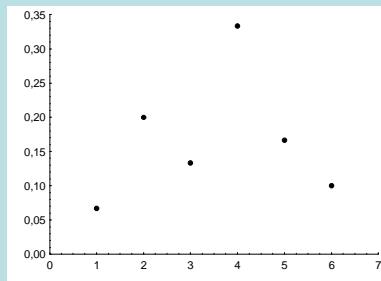
Vytvořte tabulku rozložení četností. Nakreslete grafy četnostní funkce a empirické distribuční funkce. Dále nakreslete sloupkový diagram a polygon četností počtu členů domácnosti.

**Řešení:**

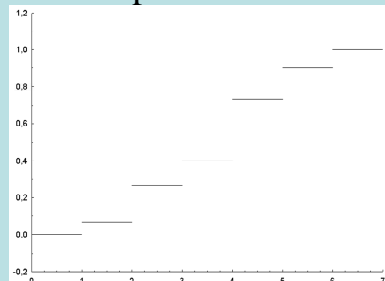
Tabulka rozložení četností

$x_{[j]}$	$n_i$	$p_i$	$N_i$	$F_i$
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

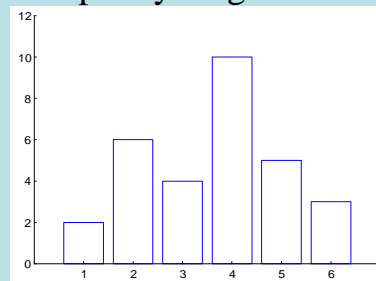
Graf četnostní funkce



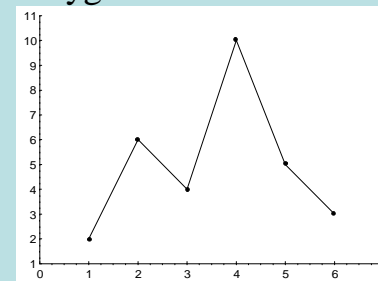
Graf empirické distribuční funkce



Sloupkový diagram



Polygon četností



### 3. Intervalové rozložení četností

Je-li počet variant znaku  $X$  velký, přiřazujeme četnosti nikoli jednotlivým variantám, ale třídícím intervalům  $(u_1, u_2)$ , ...,  $(u_r, u_{r+1})$  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:**  $f(x) = \begin{cases} f_j & \text{pro } u_j < x \leq u_{j+1}, j=1, \dots, r \\ 0 & \text{jinak} \end{cases}$  (grafem hustoty četnosti je histogram)

**Intervalová empirická distribuční funkce:**  $F(x) = \int_{-\infty}^x f(t) dt$ .

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

Výdaje	(35,65)	(65,95)	(95,125)	(125,155)	(155,185)	(185,215)
Počet dom.	7	16	27	14	4	2

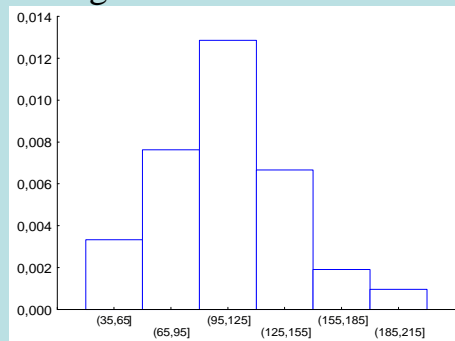
Sestavte tabulku rozložení četností, nakreslete histogram a graf intervalové empirické distribuční funkce.

**Řešení:**

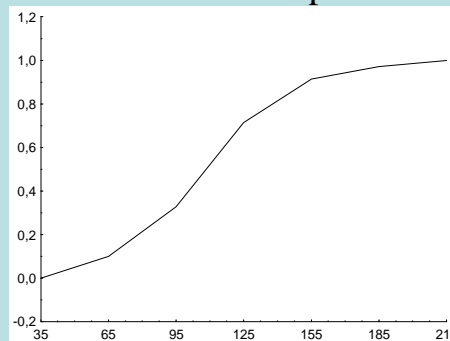
Tabulka rozložení četností

$(u_j, u_{j+1}]$	$n_j$	$p_j$	$f_j$	$N_j$	$F_j$
(35,65)	7	7/70	7/2100	7	7/70
(65,95)	16	16/70	16/2100	23	23/70
(95,125)	27	27/70	27/2100	50	50/70
(125,155)	14	14/70	14/2100	64	64/70
(155,185)	4	4/70	4/2100	68	68/70
(185,215)	2	2/70	2/2100	70	1

Histogram



Graf intervalové empirické distribuční funkce



## 4. 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. nejčtenější varianta či střed nejčtenějšího 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:

$$n\alpha = \begin{cases} \text{celé číslo } c \Rightarrow x_\alpha = \frac{x_{(c)} + x_{(c+1)}}{2} \\ \text{necelé číslo} \Rightarrow \text{zaokrouhlíme nahoru na nejbližší celé číslo } c \Rightarrow x_\alpha = x_{(c)} \end{cases}$$

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

$$q = 7 - 4 = 3$$

## 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, ...

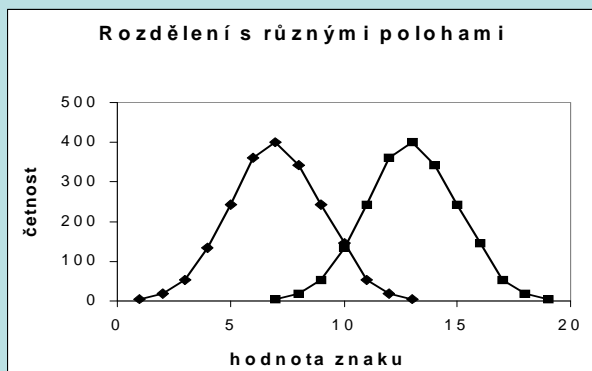
Společný znak poměrových znaků: poměrový znak má přirozený počátek, ke kterému jsou vztahovány všechny další hodnoty znaku.

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  $\frac{1}{n} \sum_{i=1}^n (x_i - m) = \frac{1}{n} \sum_{i=1}^n x_i - \frac{1}{n} \sum_{i=1}^n m = m - \frac{1}{n} \cdot n \cdot m = 0 = 0$ .

- 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.

## Charakteristiky variability intervalových a poměrových znaků

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

**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)

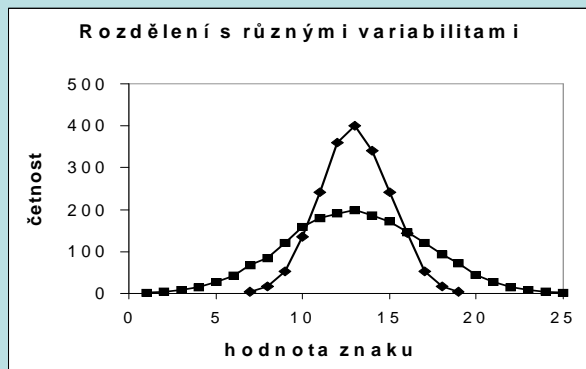
**směrodatná odchylka**  $s = \sqrt{s^2}$ .

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

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



Upozornění: Pohlížíme-li na datový soubor jako na výběrový soubor, bude ve jmenovateli vzorce pro rozptyl  $n-1$ , nikoliv  $n$  a výběrový rozptyl budeme považovat za nestranný odhad populačního rozptylu  $\sigma^2$ .

### Vlastnosti rozptylu:

- Rozptyl je nulový pouze tehdy, když jsou všechny hodnoty stejné, jinak je kladný.
- Rozptyl centrovaných hodnot je roven původnímu rozptylu, neboť  $\frac{1}{n} \sum_{i=1}^n [(x_i - m) - 0]^2 = \frac{1}{n} \sum_{i=1}^n (x_i - m)^2 = s^2$ .
- 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  $s^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - m^2$ .
- 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.

## 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**  $s^2 = \frac{1}{n} \sum_{j=1}^r n_j (x_{[j]} - m)^2 = \sum_{j=1}^r p_j (x_{[j]} - m)^2$  (výpočetní vzorec:  $s^2 = \frac{1}{n} \sum_{j=1}^r n_j x_{[j]}^2 - m^2 = \sum_{j=1}^r p_j x_{[j]}^2 - m^2$ ).

**Příklad 4.:** 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

Vypočtěte průměr, směrodatnou odchylku a koeficient variace počtu odpracovaných hodin.

**Řešení:**

$$\text{Vážený průměr: } m = \frac{1}{n} \sum_{j=1}^r n_j x_{[j]} = \frac{1}{35} (4 \cdot 184 + 6 \cdot 185 + 7 \cdot 186 + 6 \cdot 187 + 7 \cdot 188 + 5 \cdot 189) = 186,6$$

$$\text{Vážený rozptyl: } s^2 = \frac{1}{n} \sum_{j=1}^r n_j x_{[j]}^2 - m^2 = \frac{1}{35} (4 \cdot 184^2 + 6 \cdot 185^2 + 7 \cdot 186^2 + 6 \cdot 187^2 + 7 \cdot 188^2 + 5 \cdot 189^2) - 186,6^2 = 2,5257$$

$$\text{Vážená směrodatná odchylka: } s = \sqrt{s^2} = \sqrt{2,5257} = 1,59 \text{ h} = 1 \text{ h } 35 \text{ min}$$

$$\text{Koeficient variace: } \frac{s}{m} 100\% = \frac{1,59}{186,6} 100\% = 0,85\%$$

Vidíme, že zaměstnanci odpracovali za měsíc v průměru 186,6 h, přičemž směrodatná odchylka dosahuje 0,85 % průměrné odpracované doby.

## Šikmost

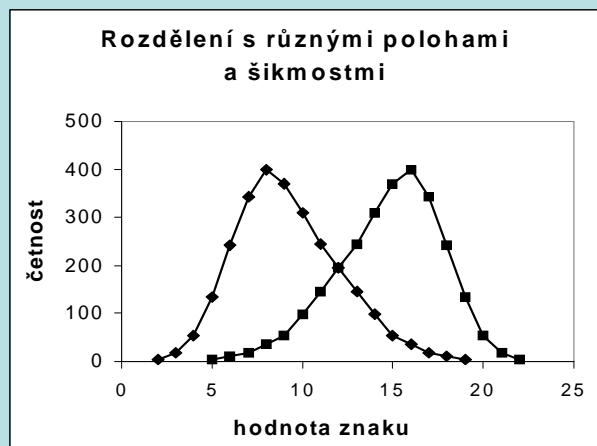
$$\alpha_3 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - m)^3}{s^3}$$
 - měří nesouměrnost rozložení četností kolem průměru.

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

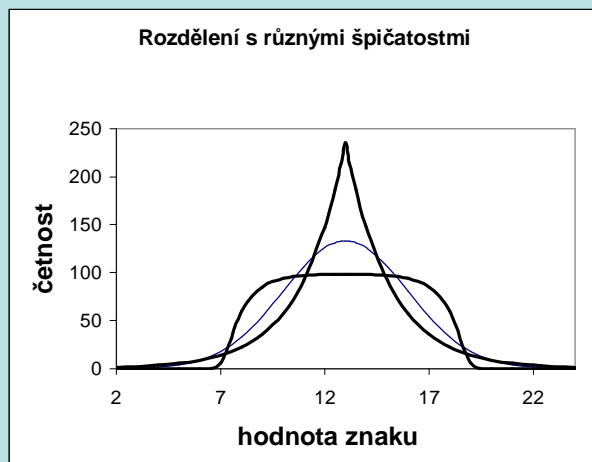
$$\alpha_4 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - m)^4}{s^4} - 3$$
 - měří koncentraci rozložení četností kolem průměru.

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



V případě vícerozměrného datového souboru pro znak  $X_j$  zavedeme průměr  $m_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$ , rozptyl

$s_j^2 = \frac{1}{n} \sum_{i=1}^n (x_{ij} - m_j)^2$  a pro dvojici znaků  $(X_j, X_k)$  zavedeme kovarianci  $s_{jk} = \frac{1}{n} \sum_{i=1}^n (x_{ij} - m_j)(x_{ik} - m_k)$

a koeficient korelace  $r_{jk} = \frac{1}{n} \sum_{i=1}^n \frac{x_{ij} - m_j}{s_j} \frac{x_{ik} - m_k}{s_k} = \frac{s_{jk}}{s_j s_k}$ .

Průměry uspořádáme do **vektoru průměrů**  $(m_1, \dots, m_p)^T$ , rozptyly a kovariance do **varianční**

**matice**  $\begin{pmatrix} s_1^2 & \cdots & s_{1p} \\ \cdots & \cdots & \cdots \\ s_{p1} & \cdots & s_p^2 \end{pmatrix}$  a koeficienty korelace do **korelační matice**  $\begin{pmatrix} 1 & \cdots & r_{1p} \\ \cdots & \cdots & \cdots \\ r_{p1} & \cdots & 1 \end{pmatrix}$ . Tyto matice jsou

symetrické, protože kovariance a koeficient korelace jsou symetrické.

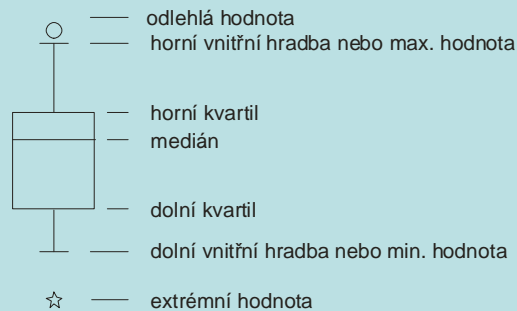
Koeficient korelace  $r_{jk}$  nás informuje o síle lineární závislosti mezi znaky  $X_j, X_k$ .

## 5. Diagnostické grafy

### a) 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  $(-\infty, x_{0,25} - 3q)$ .

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

U 30 domácností byl zjišťován počet členů.

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

Pro tyto údaje sestrojte krabicový diagram.

## Řešení:

Připomeneme nejprve definici  $\alpha$ -kvantilu. 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:

$$n\alpha = \begin{cases} \text{celé číslo } c \Rightarrow x_\alpha = \frac{x_{(c)} + x_{(c+1)}}{2} \\ \text{necelé číslo} \Rightarrow \text{zaokrouhlíme nahoru na nejbližší celé číslo } c \Rightarrow x_\alpha = x_{(c)} \end{cases}$$

Algoritmus:

$$n\alpha = \begin{cases} \text{celé číslo } c \Rightarrow x_\alpha = \frac{x_{(c)} + x_{(c+1)}}{2} \\ \text{necelé číslo} \Rightarrow \text{zaokrouhlíme nahoru na nejbližší celé číslo } c \Rightarrow x_\alpha = x_{(c)} \end{cases}$$

Data:

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

V našem případě 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

Dolní kvartil je 2, tedy aspoň čtvrtina domácností má aspoň dva členy.

Medián je 4, tedy aspoň polovina domácností má aspoň 4 členy.

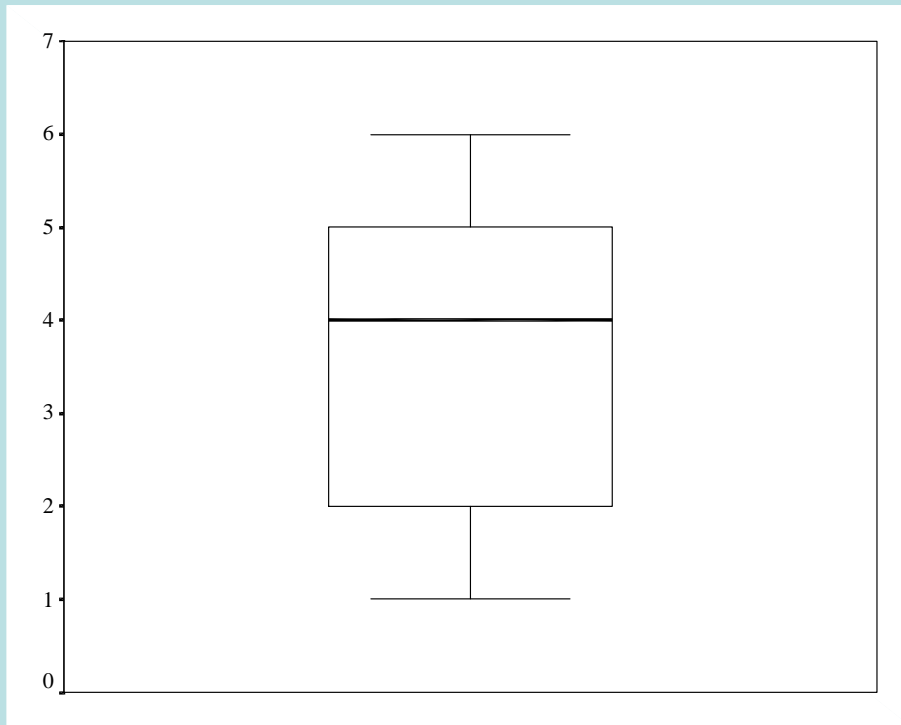
Horní kvartil je 5, tedy aspoň tři čtvrtiny domácností mají aspoň 5 členů.

Vypočteme kvartilovou odchylku:  $q = x_{0,75} - x_{0,25} = 5 - 2 = 3$ .

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

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

Nakonec sestrojíme krabicový diagram:



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

## b) Normální pravděpodobnostní graf (NP-plot)

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

**Způsob konstrukce:** na vodorovnou osu vynášíme uspořádané hodnoty  $x_{(1)} \leq \dots \leq x_{(n)}$  a na

svislou osu kvantily  $u_{\alpha_j}$ , kde  $\alpha_j = \frac{3j-1}{3n+1}$ , 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  $(x_{(j)}, u_{\alpha_j})$  budou ležet na přímce.

Pro data z rozložení s kladnou šikmostí se dvojice  $(x_{(j)}, u_{\alpha_j})$  budou řadit do konkávní křivky, zatímco pro data z rozložení se zápornou šikmostí se dvojice  $(x_{(j)}, u_{\alpha_j})$  budou řadit do konvexní křivky.

## Příklad

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 NP plotu posuďte, zda se tato data řídí normálním rozložením.

## Řešení:

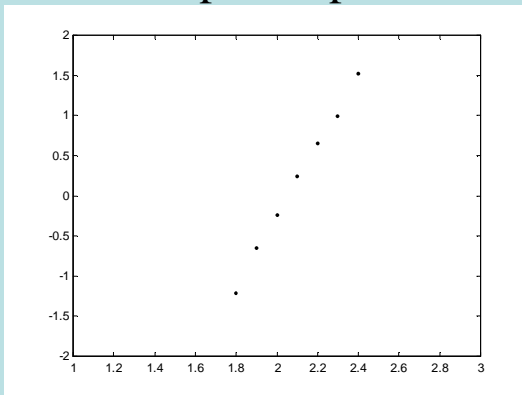
uspořádané hodnoty	1,8	1,8	1,9	2	2	2,1	2,1	2,2	2,3	2,4
pořadí	1	2	3	4	5	6	7	8	9	10
průměrné pořadí	1,5	1,5	3	4,5	4,5	6,5	6,5	8	9	10

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

vektor hodnot  $\alpha_j = \frac{3j-1}{3n+1} = (0,1129; 0,2581; 0,4032; 0,5968; 0,7419; 0,8387; 0,9355)$ ,

vektor kvantilů  $u_{\alpha_j} = (-1,2112; -0,6493; -0,245; 0,245; 0,6493; 0,9892; 1,5179)$ .

Normální pravděpodobnostní graf



Závěr:

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



### c) **Dvourozměrný tečkový diagram**

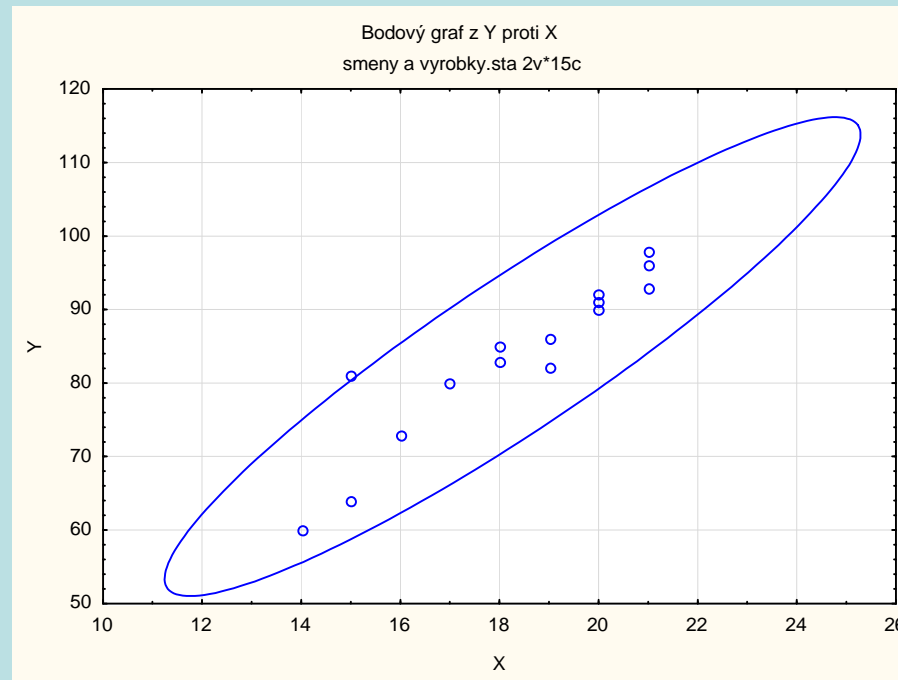
Slouží ke grafickému znázornění vztahu mezi dvěma znaky  $X_j$ ,  $X_k$ . Na vodorovnou osu vyneseme hodnoty  $x_j$ , na svislou hodnoty  $x_k$  a do příslušných průsečíků nakreslíme tolik teček, jaká je absolutní četnost dvojice  $(x_j, x_k)$ . Jedná-li se o náhodný výběr z dvourozměrného normálního rozložení, měly by tečky zhruba rovnoměrně vyplnit vnitřek elipsovitého obrazce. Vrstevnice hustoty dvourozměrného normálního rozložení jsou totiž elipsy.

## Příklad

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.

Pomocí dvourozměrného tečkového diagramu se zakreslenou 95% elipsou konstantní hustoty pravděpodobnosti posud'te, zda tato data lze považovat za realizace náhodného výběru z dvourozměrného normálního rozložení.

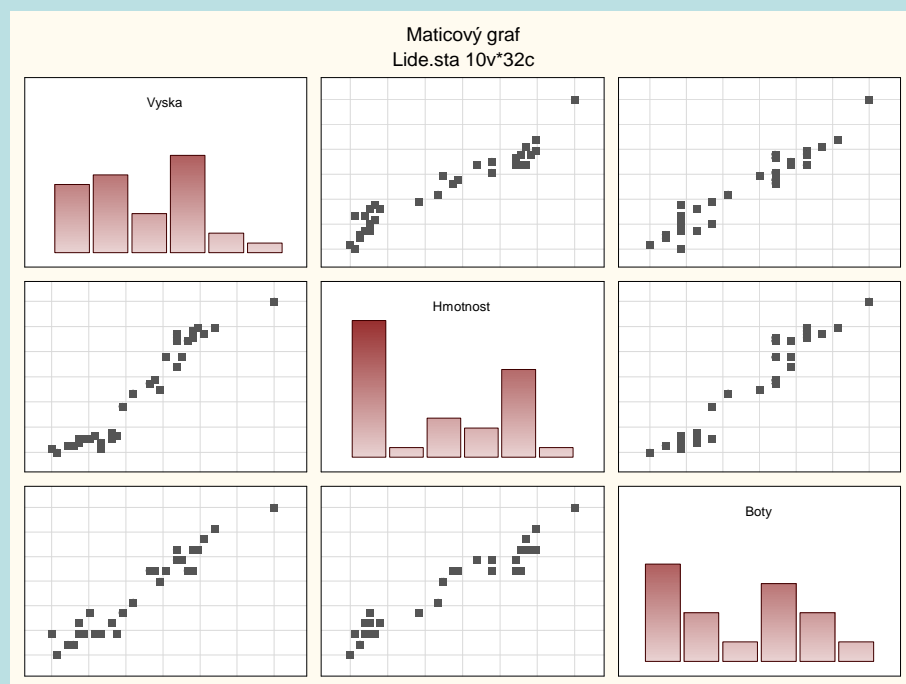


Obrázek svědčí o tom, že předpoklad dvourozměrné normality je oprávněný a že mezi počtem směn a počtem výrobků bude existovat určitý stupeň přímé lineární závislosti, tzn., že u dělníků, kteří měli vysoký resp. nízký počet směn, lze očekávat vysoký resp. nízký počet výrobků.

## d) Maticový graf

Používá se ke grafickému znázornění  $p$ -rozměrného datového souboru, obsahuje  $p \times p$  grafů uspořádaných do čtvercového schématu. Na hlavní diagonále jsou histogramy znaků  $X_1, \dots, X_p$ , mimo hlavní diagonálu pak dvourozměrné tečkové diagramy dvojic znaků.

Ukázka maticového grafu:



## II. 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.

## 1. Nulová a alternativní hypotéza

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

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

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

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

**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.

## 2. 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

### 3. Tři způsoby testování hypotéz

#### a) 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:

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

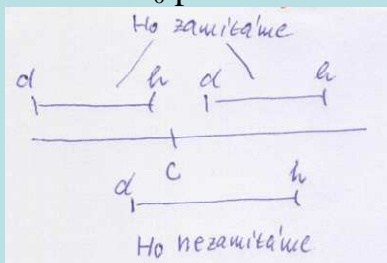
Kritický obor v případě pravostranné alternativy má tvar:

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

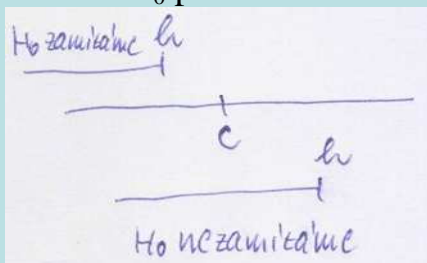
## b) Testování pomocí intervalu spolehlivosti

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

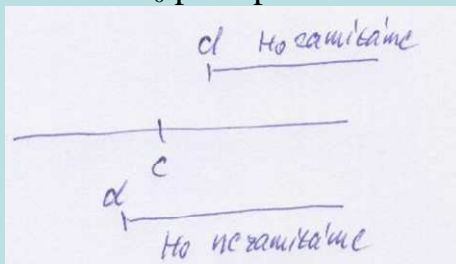
Pro test  $H_0$  proti oboustranné alternativě sestrojíme oboustranný interval spolehlivosti.



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



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



### c) 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  $\alpha$ .

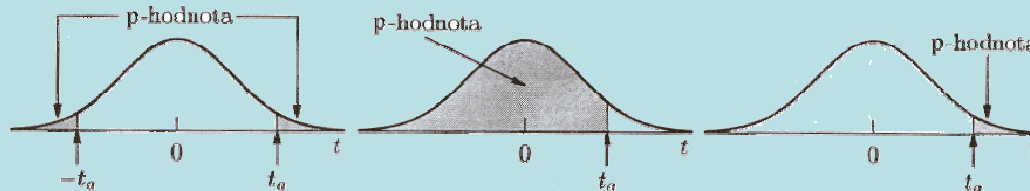
Způsob výpočtu p-hodnoty:

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  $h(\vartheta)$ . 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.)

## 4. Testy normality dat

K ověřování normality dat slouží celá řada testů, které jsou podrobně popsány ve statistické literatuře. Zde se omezíme na tři testy, které jsou implementovány v systému STATISTICA, a to **Kolmogorovův – Smirnovův test** a jeho **Lilieforsovu variantu**, **Shapiroův – Wilkův test** a **Andersonův – Darlingův 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.

## a) 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  $\Phi_T(x)$ .

Nechť  $F_n(x)$  je výběrová distribuční funkce.

Testovou statistikou je statistika  $D_n = \sup_{-\infty < x < \infty} |F_n(x) - \Phi_T(x)|$ .

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  $\sqrt{\frac{1}{2n} \ln \frac{2}{\alpha}}$ .

**Upozornění:** Nulová hypotéza musí specifikovat distribuční funkci zcela přesně, včetně všech jejích případných parametrů. Např. K-S test lze použít pro testování hypotézy, že náhodný výběr  $X_1, \dots, X_n$  pochází z rozložení  $Rs(0,1)$ , což se využívá při testování generátorů náhodných čísel. Pokud však parametry distribuční funkce odhadujeme z výběru, změní se rozložení testové statistiky  $D_n$  a jde o Lilieforsův test. Příslušné modifikované kvantily byly určeny pomocí simulačních studií.

## b) Shapirův – Wilkův test

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

Testová statistika má tvar:

$$W = \frac{\sum_{i=1}^m a_i^{(n)} [X_{(n-i+1)} - X_{(i)}]^2}{\sum_{i=1}^m (X_i - M)^2},$$

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.

### c) Andersonův – Darlingův test

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

Testová statistika má tvar:

$$AD = -\frac{1}{n} \left[ \sum_{i=1}^n (2i-1) \left\{ \ln \Phi \left( \frac{x_{(i)} - m}{s} \right) + \ln \left( 1 - \Phi \left( \frac{x_{n+1-(i)} - m}{s} \right) \right) \right\} \right] - n ,$$

kde  $x_{(i)}$  jsou vzestupně uspořádané realizace náhodného výběru,  $\Phi$  je distribuční funkce rozložení  $N(0,1)$ .

Hypotéza  $H_0$  se zamítá na hladině významnosti  $\alpha$ , je-li vypočítaná hodnota testové statistiky  $AD$  větší než kritická hodnota  $D_{1-\alpha}$ . Pro velký rozsah výběru se přibližná 95% kritická hodnota počítá podle vzorce

$$D_{0,95} = 1,0348 \left( 1 - \frac{1,013}{n} - \frac{0,93}{n^2} \right)$$



### Příklad:

Jsou dány hodnoty 10, 12, 8, 9, 16. Pomocí Lilieforsova testu, S – W testu a A – D testu testujte na hladině významnosti 0,05 hypotézu, že 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.

#### Provedení Lilieforsova a S-W testu:

V menu vybereme Statistiky – Základní statistiky/tabulky – Tabulky četností – OK, Proměnné X – OK. Na záložce zvolíme Normalita a zaškrtneme Lilieforsův test a Shapiro – Wilkův W test – Testy normality.

Proměnná	Testy normality (Tabulka1)				
	N	max D	Lilliefors p	W	p
X	5	0,224085	p > .20	0,912401	0,482151

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.

#### Provedení A - D testu:

Statistiky – Rozdělení & simulace – proložení dat rozděleními – OK – Proměnné Spojité: X – na záložce Spojité proměnné ponecháme zaškrtnuté pouze Normální, na záložce Možnosti vybereme Anderson – Darling – OK – Souhrnné statistiky rozdělení.

	Souhrn rozdělení for Proměnná: x (Tabulka4)							
	K-S d	K-S p-hodn.	AD stat.	AD p-hodn.	Chí-kvadrát	Chí-kvadr. p-hodn.	Chí-kvadr. SV	Posun (práh/poloha)
Normální (poloha,měřítko)	0,224085	0,915101	0,295219	0,940172				

Testová statistika A – D testu je 0,2952, odpovídající p-hodnota je 0,9402, tedy hypotézu o normalitě nezamítáme na hladině významnosti 0,05.