

# Exploračná analýza

Portál: edu.ukf.sk - Vzdelávací portál - Univerzita  
Konštantína Filozofa, Nitra

Kurz: Počítačová analýza dát 21/22 (KI/PAD/21)

Kniha: Exploračná analýza

Vytlačil(a): Zuzana Pavlendová

Dátum: Streda, 1 december 2021, 17:49

# Opis

Metódy exploračnej analýzy slúžia na objavenie štruktúr, vytvorenie hypotéz, rozoznanie osobitostí a znázornenie fenoménov. Východiskovým bodom každej analýzy dát sú samotné dáta. Dáta nemusia spĺňať určité podmienky, ako sa žiada v inferenčnej analýze (napr. že dáta museli byť získané náhodným výberom). Ide hlavne o to, rozličnými spôsobmi znázorniť tieto dáta, rozoznať pravidelnosti a nepravidelnosti, štruktúry, vzory a osobitosti. V exploračnom procese hľadáme v dátach zaujímavé konfigurácie a vzťahy.

# Obsah

## **1. Popisná štatistika**

- 1.1. Početnosti
- 1.2. Charakteristiky polohy
- 1.3. Charakteristiky variability
- 1.4. Charakteristiky tvaru
- 1.5. Kvantily, percentily a kvartily

## **2. Vizualizácia dát**

- 2.1. Vizualizácia početností
- 2.2. Vizualizácia popisných charakteristík
- 2.3. Vizualizácia závislostí
- 2.4. Vizualizácia rozdelenia

## **3. Analýza reziduálnych hodnôt**

## **4. Transformácia dát**

## **5. Viacrozmerné prieskumné techniky**

# 1. Popisná štatistika

Cieľom popisnej štatistiky je organizácia a popis získaných dát. Patrí sem identifikácia extrémnych hodnôt, znázorňovanie dát a ich porovnávanie pomocou tabuliek a grafov. Umožňuje porozumieť veľkému množstvu dát pomocou vhodných popisných charakteristík polohy, variability a tvaru rozdelenia dát.

# 1.1. Početnosti

Vstupná dátová tabuľka je pri väčšom rozsahu súboru neprehľadná, preto ju upravujeme na tabuľku **rozdelenia početností**.

Tabuľku početností môžeme získať **jednoduchým alebo intervalovým triedením**. Ak sa hodnoty premennej opakujú, na zistenie početností používame **jednoduché triedenie**. V prípade, že hodnoty sú rôznorodé používame **intervalové triedenie**, t.j. zisťuje sa, koľko hodnôt sa nachádza vo vytvorených intervaloch.

V prípade kvalitatívnych/nominálnych premenných (napr. pohlavie) a vo väčšine prípadov aj ordinálnych premenných (napr. prospech) používame jednoduché triedenie. V prípade metrických premenných vo väčšine prípadov používame intervalové triedenie, pokiaľ hodnoty premennej sú dostatočne rôznorodé.

## Jednoduché triedenie

Premenná môže nadobúdať  $r \geq 2$  hodnôt  $(x_1, x_2, \dots, x_r)$ . Zo vstupnej dátovej tabuľky zistíme, koľkokrát sa jednotlivé hodnoty premennej opakujú v súbore - absolútne početnosti a vypočítame relatívne početnosti, percentá a kumulatívne početnosti. Ak je možné hodnoty usporiadať podľa veľkosti (intenzívne premenné), tak ich usporiadame od minimálnej hodnoty po maximálnu.

Hodnota premennej	Absolútna početnosť	Relatívna početnosť	Relatívna početnosť(%)	Kumulatívna početnosť
$x_1$	$n_1$	$n_1/n$	$(n_1/n)100$	$\sum_{j=1}^1 n_j = n_1$
$x_2$	$n_2$	$n_2/n$	$(n_2/n)100$	$\sum_{j=1}^2 n_j$
...	...	...	...	...
$x_{r-1}$	$n_{r-1}$	$n_{r-1}/n$	$(n_{r-1}/n)100$	$\sum_{j=1}^{r-1} n_j$
$x_r$	$n_r$	$n_r/n$	$(n_r/n)100$	$\sum_{j=1}^r n_j = n$
<b><math>\Sigma</math></b>	<b><math>n</math></b>	<b>1</b>	<b>100</b>	

Ak sme správne zostavili tabuľku početností, potom súčet absolútnych početností sa rovná celkovému počtu prípadov  $n$ , súčet relatívnych početností jednej celej, súčet percent 100% a posledná kumulatívna početnosť sa rovná celkovému počtu  $n$ .

Kumulatívne početnosti nemá význam počítat pre nominálne premenné. Ale v prípade metrických a ordinálnych premenných nám pomáhajú sprehľadniť daný súbor, tak napr. v prípade ordinálnej premennej *prospech*, vieme koľkí študenti mali trojku a lepšiu známku a pod. Kumulatívne početnosti sa dajú počítat aj z relatívnych početností, rovnako ako z absolútnych.

## Intervalové triedenie

V prípade, že dátový súbor je rôznorodý, zoskupujeme hodnoty do triednych intervalov. Usporiadame hodnoty podľa veľkosti od minimálnej po maximálnu hodnotu.

Počet intervalov  $r$  si určíme alebo odhadneme napr. podľa Sturgesovho pravidla

$$r \approx 1 + 3,3 \log n$$

Následne si vypočítame šírku intervalu  $h$ , ktorú zaokrúhlime nahor

$$h = \frac{R}{r} = \frac{x_{\max} - x_{\min}}{r}$$

Intervaly volíme tak, že z jednej strany sú otvorené a z druhej uzavreté. Ak by sme volili uzavreté intervaly z oboch strán, mohol by nastať prípad, že by jedna hodnota padla do oboch intervalov. Keď zvolíme interval zľava otvorený nezačínáme v minimálnej hodnote, ale v hodnote o niečo menšej. Celému triednemu intervalu priradíme hodnotu znaku rovnú stredu intervalu.

Naznačíme tvorbu triednych intervalov:  $a < x_{\min}, (a, a+h], x_1 = a + h/2, \dots$

Triedny interval	Stred intervalu	Absolútna početnosť	Relatívna početnosť	Relatívna početnosť(%)	Kumulatívna početnosť
	$x_1$	$n_1$	$n_1/n$	$(n_1/n)100$	$\sum_{j=1}^1 n_j = n_1$
	$x_2$	$n_2$	$n_2/n$	$(n_2/n)100$	$\sum_{j=1}^2 n_j$
...	...	...	...	...	...
	$x_{r-1}$	$n_{r-1}$	$n_{r-1}/n$	$(n_{r-1}/n)100$	$\sum_{j=1}^{r-1} n_j$
	$x_r$	$n_r$	$n_r/n$	$(n_r/n)100$	$\sum_{j=1}^r n_j = n$
<b><math>\Sigma</math></b>		<b><math>n</math></b>	<b>1</b>	<b>100</b>	

Absolútne početnosti predstavujú počet hodnôt, ktoré patria do príslušného triedneho intervalu.

### Kontingenčná tabuľka

Kontingenčné tabuľky sa používajú na zachytenie početností dvoch kvalitatívnych/nominálnych premenných. Premenná  $Y$  nadobúda  $S$  rôznych hodnôt a premenná  $X$  zas  $R$  rôznych hodnôt ( $S > 1, R > 1$ ).

V kontingenčnej tabuľke  $a_{rs}, r = 1, 2, \dots, R, s = 1, 2, \dots, S$  (Tabuľka 8) predstavuje počet tých štatistických jednotiek, na ktorých sa súčasne namerali (zistili) hodnoty  $x_r$  a  $y_s$  – pozorované absolútne početnosti. Početnosti hodnôt premennej  $X$  sa označujú ako riadkové početnosti  $r_.$  a početnosti hodnôt premennej  $Y$  sa označujú ako stĺpcové početnosti  $s_.$

$$r_ = \sum_{s=1}^S a_{rs}, s_ = \sum_{r=1}^R a_{rs}, n = \sum_{r=1}^R \sum_{s=1}^S a_{rs},$$

kde  $n$  je celkový počet štatistických jednotiek v danom súbore.

Pre dve dichotomické premenné by kontingenčná tabuľka mala 4 polia (2x2) na vyjadrenie všetkých interakčných početností.

Kontingenčná tabuľka môže obsahovať okrem absolútnych pozorovaných početností aj relatívne početnosti - percentá z celkového počtu (súčet percent vo všetkých poliach  $R \times S$  dáva celok - 100%), z počtu v riadkoch (súčet percent v riadkoch pre každý stĺpec dáva celok - 100%) a z počtu v stĺpcoch (súčet percent v stĺpcoch pre každý riadok dáva celok - 100%).

## 1.2. Charakteristiky polohy

Predstavujú charakteristiky, okolo ktorých sú hodnoty premennej  $X$  sústredené. Nazývajú sa taktiež miery polohy, resp. miery strednej hodnoty.

**Aritmetický priemer** predstavuje pomer medzi súčtom všetkých hodnôt súboru a počtom hodnôt v súbore (The arithmetic mean is the ratio between the sum of all values in a file and the number of values in a file),

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

**Geometrický priemer** sa definuje vzťahom (Geometric mean is defined by relation)

$$\bar{x}_G = \sqrt[n]{\prod_{i=1}^n x_i}$$

keď sú všetky hodnoty premennej  $X$  kladné. Je to  $n$ -tá odmocnina zo súčinu nameraných hodnôt. Príkladom použitia geometrického priemeru je výpočet priemeru za určité časové obdobie. (when all  $X$  variable values are positive. An example of using a geometric average is to calculate the average over a period of time.)

**Harmonický priemer** sa definuje vzťahom (The harmonic mean is defined by relation)

$$\bar{x}_H = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

keď sú všetky hodnoty premennej  $X$  kladné. Príkladom použitia harmonického priemeru je výpočet priemernej rýchlosti. (when all  $X$  variable values are positive. An example of the use of the harmonic average is the calculation of the average speed.)

**Kvadratický priemer** sa definuje vzťahom (The quadratic average is defined by the relation)

$$\bar{x}_K = \sqrt{\frac{\sum_{i=1}^n x_i^2}{n}}$$

Každý súbor má iba jeden priemer. Keďže sa priemer vypočíta zo všetkých hodnôt je ovplyvnený extrémnymi hodnotami.

Pre priemery platí

$$x_{\min} \leq \bar{x}_H \leq \bar{x}_G \leq \bar{x} \leq \bar{x}_K \leq x_{\max}$$

ak hodnoty premennej  $X$  sú kladné a rovnosť nastáva vtedy, ak sú všetky hodnoty rovnaké, inak platí ostrá nerovnosť.

Aritmetický priemer slúži k náhrade individuálnych  $x_i$  pri sčítaní, platí

$$x_1 + \dots + x_n = \bar{x} + \dots + \bar{x}$$

Geometrický priemer slúži k náhrade individuálnych  $x_i$  pri násobení, platí

$$x_1 \dots x_n = \bar{x}_G \dots \bar{x}_G$$



Harmonický priemer slúži k náhrade individuálnych  $x_i$  pri sčítaní prevrátených hodnôt, platí

$$\frac{1}{x_1} + \dots + \frac{1}{x_n} = \bar{x}_H + \dots + \bar{x}_H$$

Kvadratický priemer slúži k náhrade individuálnych  $x_i$  pri sčítaní štvorcov, platí

$$x_1^2 + \dots + x_n^2 = \bar{x}_K^2 + \dots + \bar{x}_K^2$$

Pre výpočet **aritmetického priemeru z tabuľky rozdelenia početností**, kde pre hodnoty  $x_i$  sú zadané ich početnosti  $n_i$ , použijeme vzťah

$$\bar{x} = \frac{1}{\sum n_i} \sum n_i x_i$$

Podobne môžeme vypočítať aj ďalšie priemery z tabuľky početností.

V prípade, že máme niekoľko priemerov vypočítaných z rôznych podmnožín dát a poznáme príslušné počty meraní  $n_i$ , môžeme vypočítať celkový priemer zo všetkých dát ako **vážený priemer**

$$\bar{x} = \frac{1}{\sum n_i} \sum n_i \bar{x}_i$$

**Medián** predstavuje prostrednú hodnotu súboru, ktorý je zoradený od najmenšej po najväčšiu hodnotu. V prípade nepárneho počtu hodnôt,  $n = 2m - 1$ , medián je (The median represents the middle value of a file that is sorted from the smallest to the largest value. In the case of an odd number of values,  $n = 2m - 1$ , the median is)

$$\tilde{x} = x_m$$

V prípade párneho počtu hodnôt,  $n = 2m$ , medián je aritmetický priemer prostredných hodnôt, teda (In the case of an even number of values,  $n = 2m$ , the median is the arithmetic mean of the middle values, ie)

$$\tilde{x} = \frac{x_m + x_{m+1}}{2}$$

Medián nie je ovplyvnený extrémnymi hodnotami.

**Modus** predstavuje najčastejšie sa vyskytujúcu hodnotu. Táto charakteristika sa uplatňuje hlavne u kategoriálnych dát. Súbor môže mať viacero modusov, v tom prípade hovoríme o viacvrcholovom rozdelení. V prípade dvoch modusov hovoríme, že súbor je dvojmodálny. (The mode is the most common value. This characteristic applies mainly to categorical data. A file can have multiple modes, in which case we are talking about a multilevel split. In the case of two modes, the file is binary.)

#### **Použitie jednotlivých charakteristík polohy:**

- Priemer používame hlavne pre metrické premenné, v prípade symetrického rozdelenia a použitia parametrických testov.
- Medián používame pre intenzívne premenné, v prípade, že chceme poznať stred rozdelenia dát, v prípade výskytu extrémnych hodnôt a zošikmeného rozdelenia.
- Modus používame pre premenné, v prípade, že rozdelenie má viac vrcholov.
- V prípade symetrického rozdelenia sú všetky tieto charakteristiky približne rovnaké.

See also: <https://towardsdatascience.com/on-average-youre-using-the-wrong-average-geometric-harmonic-means-in-data-analysis-2a703e21ea0>

See also: <http://core.ecu.edu/ofe/statisticsresearch/basic%20statistics%20l.pdf>

See also:

[arithmetic mean](#)

[geometric mean](#)

[harmonic mean](#)

[quadratic mean](#)

[median](#)

[mode](#)

## 1.3. Charakteristiky variability

### Measures of Variability

Dáta s rovnakou strednou hodnotou môžu mať rôznu rozptýlenosť. Veľkosť premenlivosti dát určujeme vhodne vybranou charakteristikou variability. Nazývajú sa taktiež miery rozptýlenosti.

**Variačné rozpätie** predstavuje rozdiel medzi maximálnou a minimálnou hodnotou premennej, (The variation range is the difference between the maximum and minimum values of the variable,)

$$R = x_{\max} - x_{\min}$$

Nevýhodou variačného rozpätia je veľká citlivosť voči extrémnym hodnotám.

**Rozptyl** je dôležitou mierou variability štatistického súboru a používa sa v inferenčnej analýze pri výpočte rôznych testovacích štatistík. Rozptyl je priemerná kvadratická odchýlka merania od aritmetického priemeru, (Variance is an important measure of variability in the statistical population and is used in inference analysis to calculate various test statistics. The variance is the average square deviation of the measurement from the arithmetic mean,)

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \cdot \left( \sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2 \right) = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$$

Čím je rozptyl väčší, tým sa údaje viac odchyľujú od priemeru. Pri väčších rozsahoch nie je veľký rozdiel medzi delením číslom  $n$  alebo  $n - 1$ . Delenie číslom  $n$  sa používa, ak počítame rozptyl pre všetky prvky populácie, pri výpočte rozptylu pre výber delíme číslom  $n - 1$ .

**Smerodajná (štandardná) odchýlka** predstavuje druhú odmocninu rozptylu. V prípade, že sú všetky dáta rovnaké, smerodajná odchýlka je rovná nule. Počíta sa podľa vzorca (The standard deviation is the square root of the variance. If all data are the same, the standard deviation is zero. It is calculated according to the formula)

$$s = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2}$$

**Smerodajná (štandardná) chyba odhadu priemeru** predstavuje podiel smerodajnej odchýlky a odmocniny z rozsahu súboru  $n$ . Čím je vzorka väčšia, tým je chyba menšia. Počíta sa podľa vzorca (The standard error of the mean is the ratio of the standard deviation to the root of the range  $n$ . The larger the sample, the smaller the error. It is calculated according to the formula)

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

**Variačný koeficient** sa používa pre porovnávanie variability viacerých štatistických znakov, predstavuje relatívnu mieru variability. (The coefficient of variation is used to compare the variability of several statistical features, representing the relative measure of variability.) Vypočíta sa ako podiel smerodajnej odchýlky a priemeru. Nezávisí na jednotkách, v ktorých sú hodnoty premennej vyjadrené, na rozdiel od rozptylu a smerodajnej odchýlky. Variačný koeficient sa definuje vzťahom

$$v = \frac{s}{\bar{x}} \cdot 100$$

keď sú všetky hodnoty premennej  $X$  kladné.

Ak je hodnota variačného koeficientu väčšia ako 50%, aritmetický priemer stráca význam, pretože štatistický súbor je heterogénny, nesúrodý a aritmetický priemer ho nemôže reprezentovať. V takom prípade namiesto aritmetického priemeru ako strednú hodnotu používame medián.

**Priemerná absolútna odchýlka** (Average absolute deviation) sa najčastejšie používa ako miera rozptýlenosti okolo aritmetického priemeru alebo mediánu. Dá sa interpretovať dvoma spôsobmi. Po prvé, je to priemerný rozdiel medzi hodnotami a priemerom pri ignorovaní znamienok. Po druhé, je to priemerný rozdiel medzi každými dvoma hodnotami pri ignorovaní znamienok. Z matematického hľadiska má však prioritu medián. Počíta sa podľa vzorca

$$d_{Me} = \frac{1}{n} \sum_{i=1}^n |x_i - \tilde{x}|$$

**Koeficient disperzie** (coefficient of dispersion) predstavuje relatívnu mieru variability, ktorá je iba málo ovplyvnená extrémnymi hodnotami. Vypočíta sa ako podiel priemernej odchýlky a mediánu,

$$d = \frac{\frac{1}{n} \sum_{i=1}^n |x_i - \tilde{x}|}{\tilde{x}} 100$$

**Mediánová absolútna odchýlka** (Median absolute deviation) je charakteristika variability, ktorá nie je ovplyvnená extrémnymi hodnotami. Vypočíta sa ako medián z absolútnych hodnôt odchýlok jednotlivých meraní od mediánu. Označuje sa *MAD* (Median Absolute Deviation),

$$MAD = \tilde{x}\{|x_i - \tilde{x}|\}$$

**Kvartilové rozpätie** predstavuje rozdiel medzi tretím  $Q_{III}$  a prvým  $Q_I$  kvartilom (75. a 25. percentilom)

$$Q = Q_{III} - Q_I$$

čo reprezentuje oblasť stredných 50% hodnôt premennej. Táto miera variability nie je ovplyvnená extrémnymi hodnotami premennej.

#### **Použitie jednotlivých charakteristík variability:**

- Smerodajná odchýlka a rozptyl merajú rozptýlenosť okolo priemeru a používajú sa, keď priemer je vhodný ako miera strednej hodnoty.
- Smerodajná odchýlka a rozptyl sú silne ovplyvnené extrémnymi hodnotami, preto v tomto prípade uprednostňujeme kvartilové rozpätie, mediánovú absolútnu odchýlku, respektíve priemernú absolútnu odchýlku od mediánu.
- V prípade silne zošikmeného rozdelenia, smerodajná odchýlka a rozptyl neposkytujú dobrú informáciu o rozptýlenosti dát.
- V prípade, že chceme posúdiť relatívnu veľkosť rozptýlenosti dát od priemeru použijeme variačný koeficient.
- V prípade, že chceme posúdiť relatívnu veľkosť rozptýlenosti dát od mediánu použijeme koeficient disperzie.

#### **See also:**

The variation range

Variance

The standard deviation

The standard error of the mean

The coefficient of variation

Average absolute deviation

coefficient of dispersion

Median absolute deviation

## 1.4. Charakteristiky tvaru

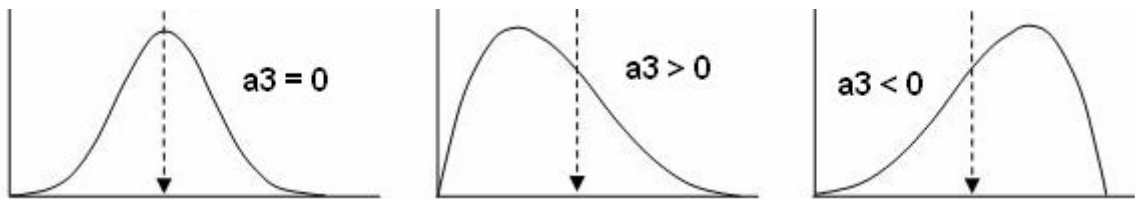
Tvar rozdelenia dát hodnotíme charakteristikami tvaru – šikmostou a špicatostou.

**Šikmost'  $a_3$**  meria stupeň asymetrie rozdelenia premennej. Kladná hodnota znamená, že priemer je väčší ako medián, teda väčšina hodnôt je menšia ako priemer. V tomto prípade je rozdelenie zošikmené doľava. Záporná hodnota znamená, že medián je väčší ako priemer a teda väčšina hodnôt je väčšia ako priemer. V takomto prípade je rozdelenie zošikmené doprava. Hodnoty blízke 0 znamenajú symetrické rozdelenie, čo znamená, že priemer a medián sa rovnajú. Počíta sa nasledovne:

$$a_3 = \frac{m_3}{s^3}$$

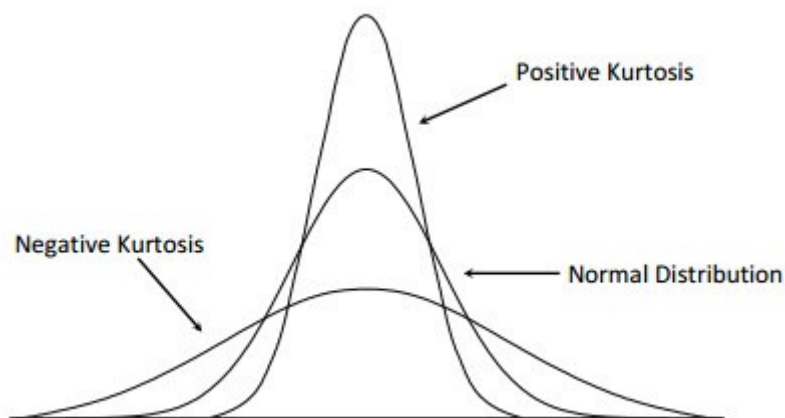
kde

$$m_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k$$



**Špicatost'  $a_4$**  meria stupeň strmosti rozdelenia premennej. Kladná hodnota znamená, že rozdelenie je špicatejšie. Záporná hodnota znamená, že rozdelenie je plochejšie. Daná je vzťahom

$$a_4 = \frac{m_4}{s^4} - 3$$



### Použitie jednotlivých charakteristík tvaru:

- Šikmost' použijeme, ak chceme zistiť, či sa častejšie vyskytujú nižšie hodnoty ako vyššie alebo naopak.
- Špicatost' použijeme, ak chceme zistiť, akým spôsobom sa vlastne hodnoty premennej koncentrujú okolo priemeru.

## 1.5. Kvantily, percentily a kvartily

**Kvantil**  $x_q$  je hodnota, pod ktorou leží definovaná časť údajov. Hladina  $q$  určuje relatívny podiel dát, ktoré sa nachádzajú pod kvantilom  $x_q$ , kde  $0 < q < 1$ .

Medián predstavuje najpoužívanější kvantil  $\tilde{x} = x_{0,5}$ . Kvantil súboru je hodnota  $k$ -tej časti, ak je súbor rozdelený na  $n$  rovnakých častí (hodnoty sú zoradené od najmenej po najväčšiu).

Výpočet kvantilu:

$j = [qn]$ , kde  $q$  je hladina kvantilu  $x_q$ ,  $n$  je celkový počet meraní/prípadov a operácia  $[]$  znamená zaokrúhlenie na menšie celé číslo.

Ak  $qn = [qn]$ , potom  $x_q = (x_j + x_{j+1})/2$ , inak  $x_q = x_{j+1}$ , kde  $x_j$  sú hodnoty premennej zoradené podľa veľkosti a  $j = 1, 2, \dots, n$ .

**Percentil** je kvantil, ktorého hladina je udávaná v percentách. Percentily delia súbor na 100 častí.

**Kvartily** predstavujú percentily s hladinou 25%, 50% a 75%. Kvartily delia súbor na 4 časti.

$Q_I$  je prvý/dolný kvartil, respektíve 25. percentil alebo  $x_{0,25}$ .

$Q_{II}$  je druhý kvartil, respektíve 50. percentil alebo medián  $x_{0,5}$ .

$Q_{III}$  je tretí/horný kvartil, respektíve 75. percentil alebo  $x_{0,75}$ .

## 2. Vizualizácia dát

Vizualizácia dát sa riadi zásadou „obrázok je viac ako tisíc slov“. Pomocou grafu je jednoduchšie detekovať konfigurácie a štruktúry. Grafickými metódami hľadáme extrémne hodnoty, rozoznávame zhluky v dátach, kontrolujeme rozdelenie dát a predpoklady, skúmame vzťahy medzi premennými, porovnávame miery strednej hodnoty a rozptýlenia alebo skúmame dáta závislé na čase.

Grafické metódy sú vhodné pre ukázanie širších vlastností dát. V prípade, že chceme uviesť vybrané údaje v presnom tvare, je lepšie ich zobrazovať v tabuľkách.

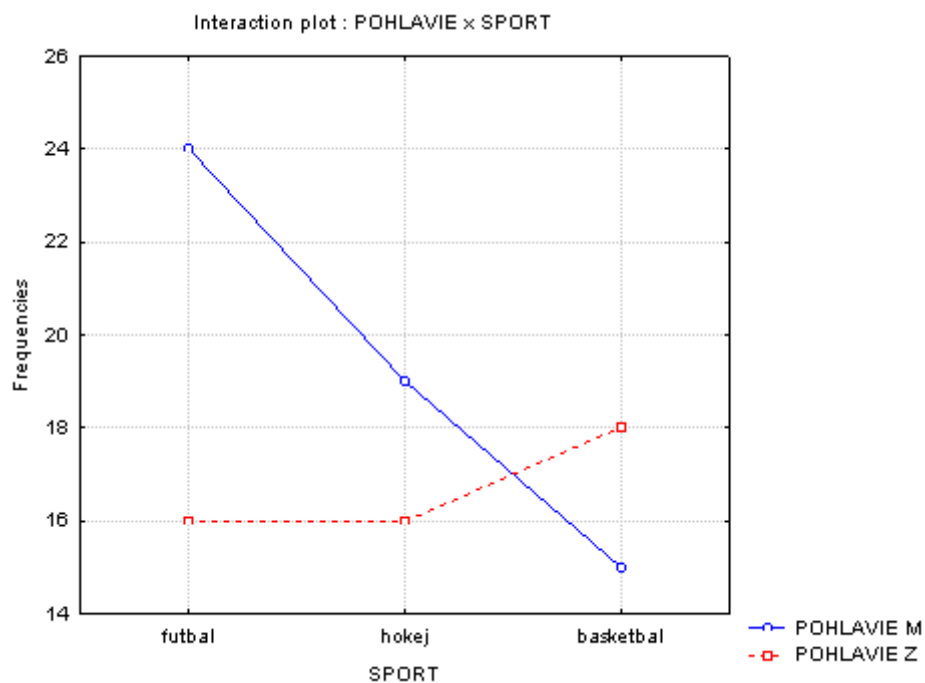
Pri analýze grafu hodnotíme: zhustenia, zhluky, medzery, extrémne hodnoty a tvar rozdelenia.

Grafy môžeme zoskupovať podľa rôznych kritérií, napríklad 2D grafy, 3D grafy, kategorizované grafy a pod. V našom prípade si ich rozdelíme podľa použitia. V žiadnom prípade však neobsiahneme všetky možnosti, ale pokúsime sa prezentovať tie najdôležitejšie. Niektoré grafy sú natoľko špecifické, že sú iba súčasťou konkrétnych analýz. Príkladom takého grafu je dendrogram, ktorý je súčasťou zhlukovej analýzy a slúži k vizualizácii zhlukov v priestore dát.

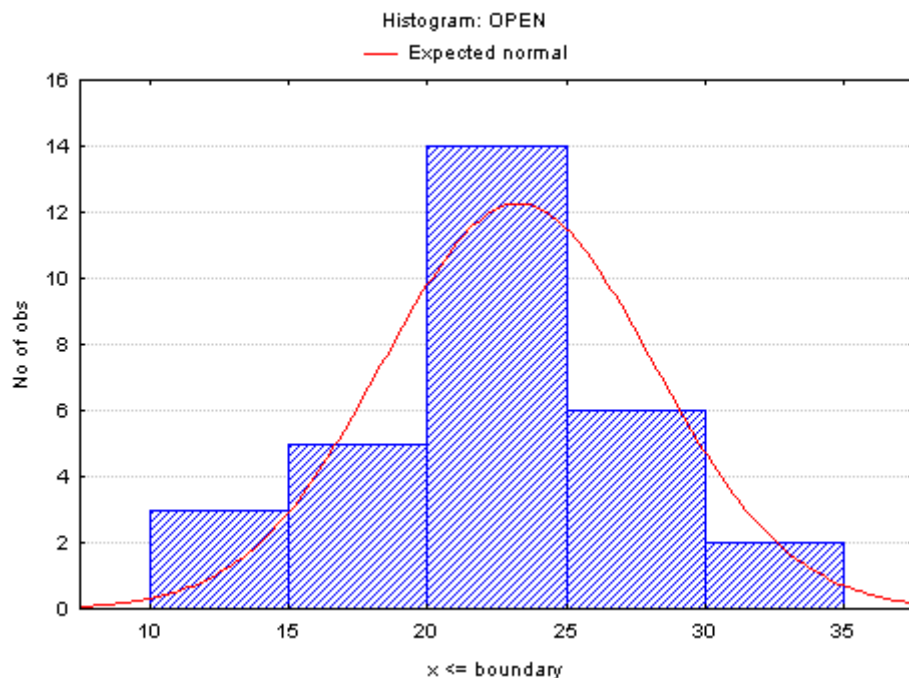


## 2.1. Vizualizácia počtostí

**Polygón** slúži k vizualizácii počtostí pri jednoduchom triedení



**Histogram** sa zvyčajne používa pri intervalovom triedení.



Intervaly, resp. kategórie sú reprezentované šírkou stĺpca (os  $X$ ) a počet prípadov, ktoré padnú do intervalu, resp. do danej kategórie je reprezentovaný výškou stĺpca (os  $Y$ ).

**Listový graf** je ekvivalentný graf k histogramu. Od histogramu sa odlišuje tým, že máme stále prehľad o jednotlivých hodnotách, t.j. vieme presne, ktoré hodnoty sa nachádzajú v danom intervale. Hodnoty naľavo od ° predstavujú stonky a hodnoty napravo listy. Napríklad v prvom intervale  $<20, 25)$  sa nachádza iba jedna hodnota  $2^4 = 24$ .

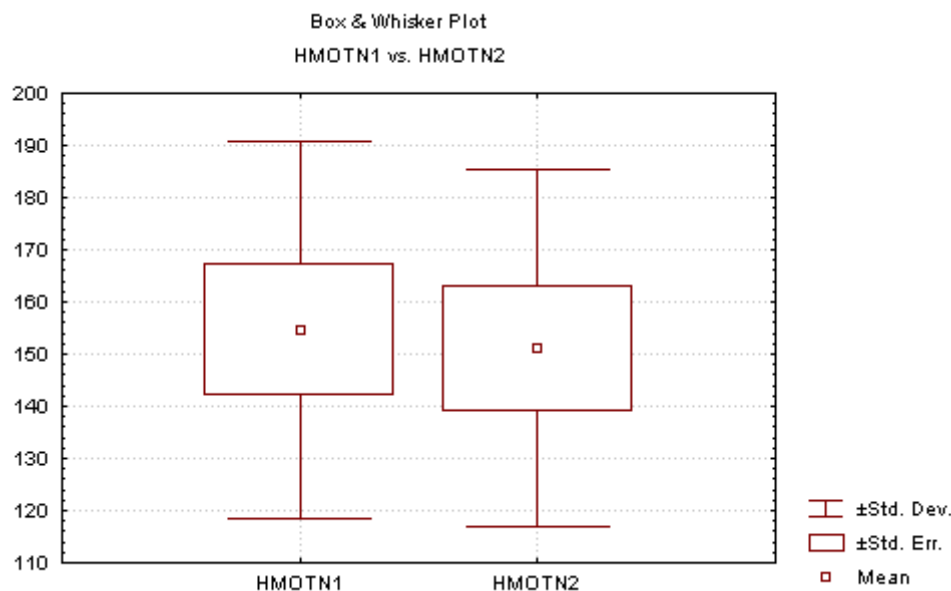
stonka° list	N	
2° 4 - - - -	1	
2° 55577999 - - - -	8	
3° 00124 - - - -	5	25%
3° 5556667777899999 - - - -	16	medián
4° 000111222344444 - - - -	15	75%
4° 556 - - - -	3	
5° - - - -	0	
min = 24,00000 max = 46,00000 Total N:	48	

**Kruhový diagram** predstavuje najvhodnejšie znázornenie relatívnych početností alebo počtu percent.

**Interakčné grafy** (Obrázok Polygón) predstavujú najlepšiu možnosť vizualizácie kontingenčnej tabuľky. Na osi *X* sú hodnoty jednej premennej a na osi *Y* sú početnosti. V grafe sú zobrazené dva polygóny pre každú úroveň druhej kategorickej premennej, čím sú vyjadrené interakčné početnosti medzi dvoma premennými. Interakčný graf môžeme považovať za **kategorizovaný polygón**. Okrem interakčného grafu k vizualizácii môžeme použiť **kategorizované histogramy**, kde na osi *X* sú hodnoty jednej premennej, na osi *Y* sú početnosti a pre každú úroveň druhej premennej sa zobrazí jeden histogram. Taktiež môžeme použiť **3D histogramy**, kde na osi *X* a *Y* sú hodnoty premenných a na osi *Z* sú početnosti kombinácií hodnôt premenných.

## 2.2. Vizualizácia popisných charakteristík

K zobrazovaniu popisných štatistík je vhodný **krabicový graf**. Pomocou tohto grafu môžeme posudzovať a porovnávať miery polohy a rozptýlenosť hodnôt okolo nich, taktiež môžeme posudzovať zošikmenie a prítomnosť extrémnych hodnôt.



Podľa toho, aké popisné charakteristiky krabicový graf zobrazuje, rozlišujeme dve základné zobrazenia:

- **medián/kvartilové rozpätie/variačné rozpätie,**
- **aritmetický priemer/smerodajná chyba/smerodajná odchýlka.**

Ak premenná má symetrické rozdelenie – priemer ako miera polohy má zmysel, tak súbor charakterizujeme najčastejšie priemerom, smerodajnou chybou a smerodajnou odchýlkou. V opačnom prípade, keď sa v súbore vyskytujú extrémne hodnoty, súbor charakterizujeme najčastejšie mediánom, kvartilovým rozpätím a variačným rozpätím.

Okrem základného delenia podľa zobrazených popisných charakteristík rozlišujeme nasledovné typy:

- **medián/kvartilové rozpätie/variačné rozpätie,**

1.	2.	3.	4.	5.
minimum	dolný kvartil	medián	horný kvartil	maximum
$x_{min}$	$Q_1$	$\tilde{x}$	$Q_3$	$x_{max}$

Interval predstavuje variačné rozpätie  $R$ , väčší obdĺžnik kvartilové rozpätie  $Q$  a menší medián (Tabuľka 9). Ak je medián blízko dolného alebo horného kvartilu, rozdelenie je zošikmené.

- **medián/kvartilové rozpätie/1,5\*kvartilové rozpätie,**

1.	2.	3.	4.	5.
dolný kvartil -1,5*kvartilové rozpätie	dolný kvartil	medián	horný kvartil	horný kvartil +1,5*kvartilové rozpätie
$Q_1 - 1,5 \cdot Q$	$Q_1$	$\tilde{x}$	$Q_3$	$Q_3 + 1,5 \cdot Q$

Tento typ krabicového grafu sa odlišuje od predchádzajúceho tým, že interval je tvorený poslednou hodnotou pod  $Q_3 + 1,5 \cdot Q$  a poslednou hodnotou premennej nad  $Q_1 - 1,5 \cdot Q$ . Hodnoty mimo intervalu pokladáme za podozrivé – možné extrémne prípady, ktoré nie sú určené premennou, ale vznikli chybným meraním alebo chybným zápisom.

- aritmetický priemer/smerodajná chyba/smerodajná odchýlka,

1.	2.	3.	4.	5.
priemer -1*smerodajná odchýlka	priemer -1*smerodajná chyba	priemer	priemer +1*smerodajná chyba	priemer +1*smerodajná odchýlka
$\bar{x} - s$	$\bar{x} - s_z$	$\bar{x}$	$\bar{x} + s_z$	$\bar{x} + s$

Interval  $\pm$  predstavuje  $\bar{x} \pm s$ , väčší obdĺžnik  $\square \bar{x} \pm s_z$  a menší  $\square$  priemer  $\bar{x}$ .

- aritmetický priemer/smerodajná chyba/1,96\*smerodajná chyba,

1.	2.	3.	4.	5.
priemer -1,96*smerodajná chyba	priemer -1*smerodajná chyba	priemer	priemer +1*smerodajná chyba	priemer +1,96*smerodajná chyba
$\bar{x} - 1,96 * s_z$	$\bar{x} - s_z$	$\bar{x}$	$\bar{x} + s_z$	$\bar{x} + 1,96 * s_z$

Interval  $\pm$  predstavuje  $\bar{x} \pm 1,96 * s_z$ , väčší obdĺžnik  $\square \bar{x} \pm s_z$  a menší  $\square$  priemer  $\bar{x}$ . Interval zobrazuje 95% interval spoľahlivosti priemeru.

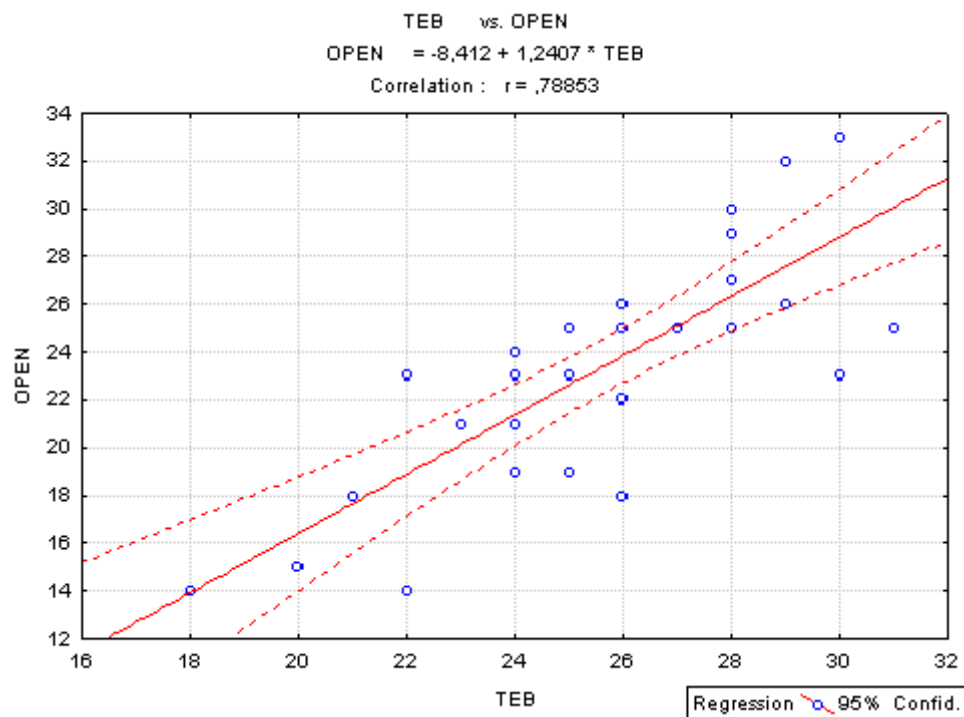
- aritmetický priemer/smerodajná odchýlka/1,96\*smerodajná odchýlka.

1.	2.	3.	4.	5.
priemer -1,96*smerodajná odchýlka	priemer -1*smerodajná odchýlka	priemer	priemer +1*smerodajná odchýlka	priemer +1,96*smerodajná odchýlka
$\bar{x} - 1,96 * s$	$\bar{x} - s$	$\bar{x}$	$\bar{x} + s$	$\bar{x} + 1,96 * s$

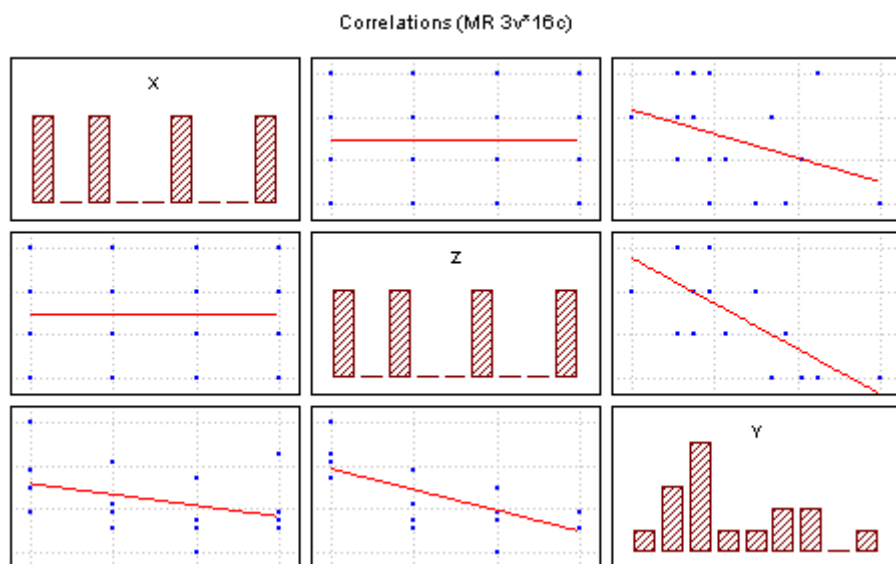
Interval  $\pm$  predstavuje  $\bar{x} \pm 1,96 * s$ , väčší obdĺžnik  $\square \bar{x} \pm s$  a menší  $\square$  priemer. Interval zobrazuje 95% interval spoľahlivosti pre individuálne pozorovania okolo priemeru.

## 2.3. Vizualizácia závislostí

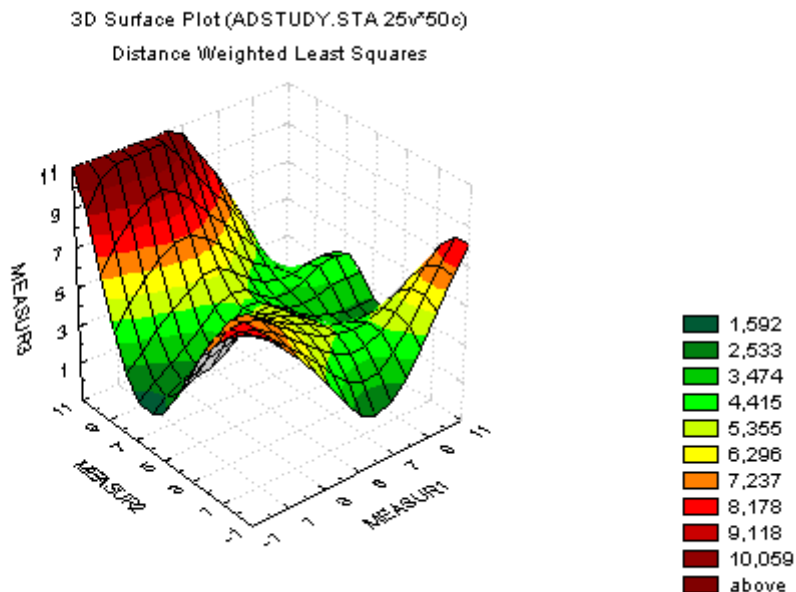
**Korelačný/bodový graf** je vizualizáciou lineárnej závislosti medzi dvoma intenzívnymi/kvantitatívnymi premennými. V prípade priamoúmernej závislosti (hodnoty sa menia spoločne jedným smerom) sú body v korelačnom grafe preložené rastúcou priamkou – body, v prípade nepriamoúmernej závislosti (hodnoty sa menia spoločne opačným smerom) sú body preložené klesajúcou priamkou a v prípade nezávislosti konštantnou priamkou – hodnoty sa spolu nemenia ani jedným smerom. Jediný extrémista vo veľkom súbore môže významne znížiť silnú závislosť, ale aj vyrobiť silnú závislosť tam, kde žiadna nie je. Preto je nutné preskúmať korelačný graf.



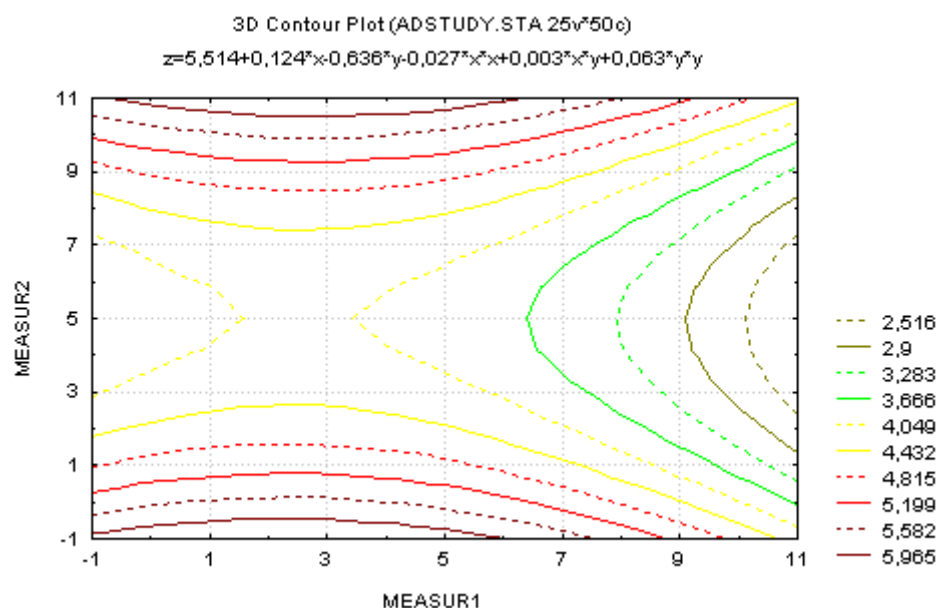
**Maticový graf** je vizualizáciou lineárnej závislosti medzi viacerými intenzívnymi/kvantitatívnymi premennými.



**Povrchový graf** vizualizuje funkčnú závislosť intenzívnych/kvantitatívnych premenných - závislej premennej  $Y$  na nezávislých premenných  $X_1$  a  $X_2$ .



**Vrstevnicový graf** je príbuzným povrchového grafu. Predstavuje topografickú mapu zostrojenú z troch intenzívnych/kvantitatívnych premenných. Na osi *X* je jedna premenná, druhá je na osi *Y* a tretia premenná je reprezentovaná izočiarami (čiarami s rovnakou hodnotou).



Spomínaný **interakčný graf** by sme mohli pokladať za vizualizáciu závislostí dvoch nominálnych premenných, respektíve **3D histogram** alebo **kategorizovaný histogram**. Rovnako aj **grafom priemerov a intervalov spoľahlivosti** môžeme vizualizovať závislosť medzi kvantitatívnou a kvalitatívnou premennou.

## 2.4. Vizualizácia rozdelenia

Histogramom môžeme určiť základný tvar rozdelenia a identifikovať prítomnosť extrémnych hodnôt. Histogram môže mať symetrický tvar alebo zošikmený doľava respektíve doprava. Taktiež môže mať jeden alebo viac vrcholov. Histogramy niekedy prekladáme ideálnym očakávaným rozdelením. Za účelom porovnania rozdelenia premennej s normálnym - symetrickým rozdelením používame **histogram preložený Gaussovou krivkou** alebo **normálny graf pravdepodobnosti**.

### 3. Analýza reziduálnych hodnôt

Analýza reziduálnych hodnôt vychádza zo základnej predstavy:

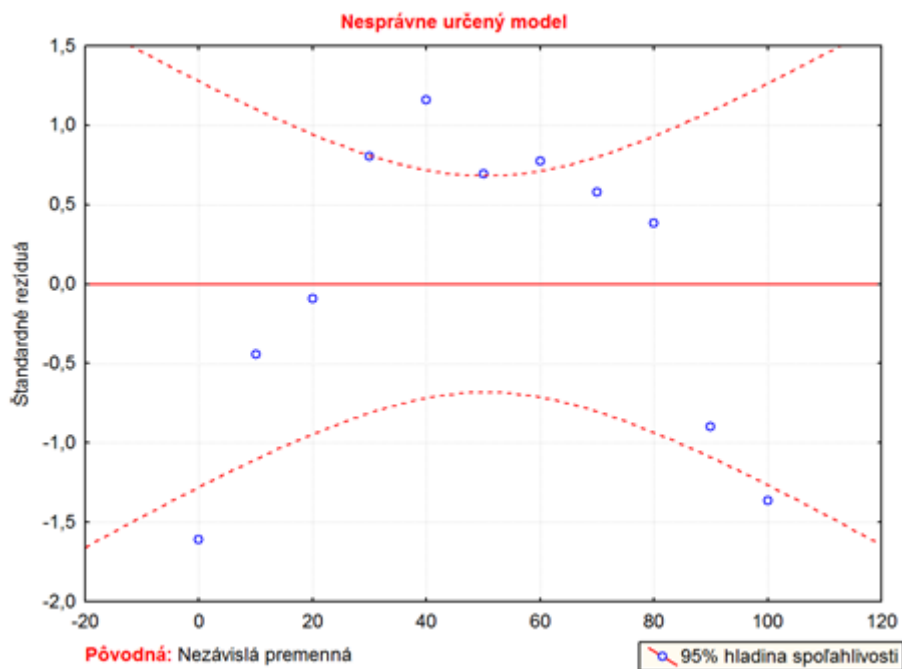
$$\text{Dáta} = \text{predikcia modelom (funkcia)} + \text{reziduálna hodnota}.$$

Ak od dát (pozorované hodnoty) odpočítame hodnoty získané z modelu (očakávané hodnoty), získame chyby (reziduálne hodnoty) a ich analýzou môžeme posúdiť zostrojený model.

Tento typ analýzy slúži k overeniu validity modelu a k jeho vylepšovaniu, pretože pomáha odhaliť aspekty vzťahov, ktoré model nezohľadňuje.

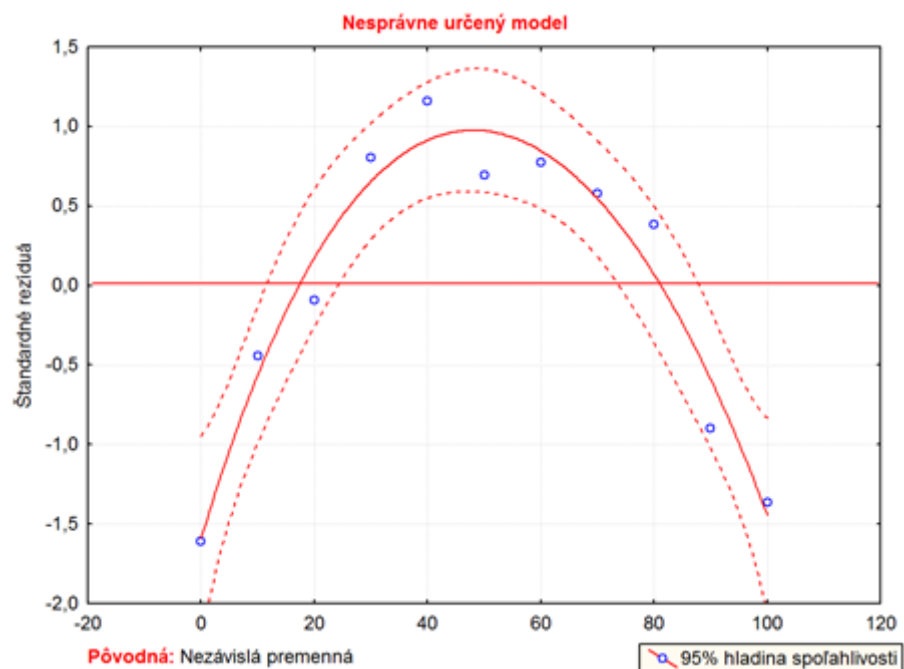
Napríklad pomocou analýzy reziduálnych hodnôt môžeme overiť stabilitu regresného modelu, t.j. identifikovať nesprávnosť zvoleného modelu – prostredníctvom zobrazenia korelačného/bodového grafu štandardných rezíduí a nezávislej premennej. Keď je model správne určený, tak sú body náhodne rozmiestnené okolo vodorovnej osi. Ak sa body zhľukujú okolo nejakej priamky rôznej od vodorovnej osi alebo okolo nejakej krivky, je to známka nesprávne zvoleného modelu. V našom prípade sa body zhľukujú okolo paraboly, je nutné transformovať nezávislú premennú  $X = X^2$

*Body nie sú náhodne rozmiestnené okolo vodorovnej osi*

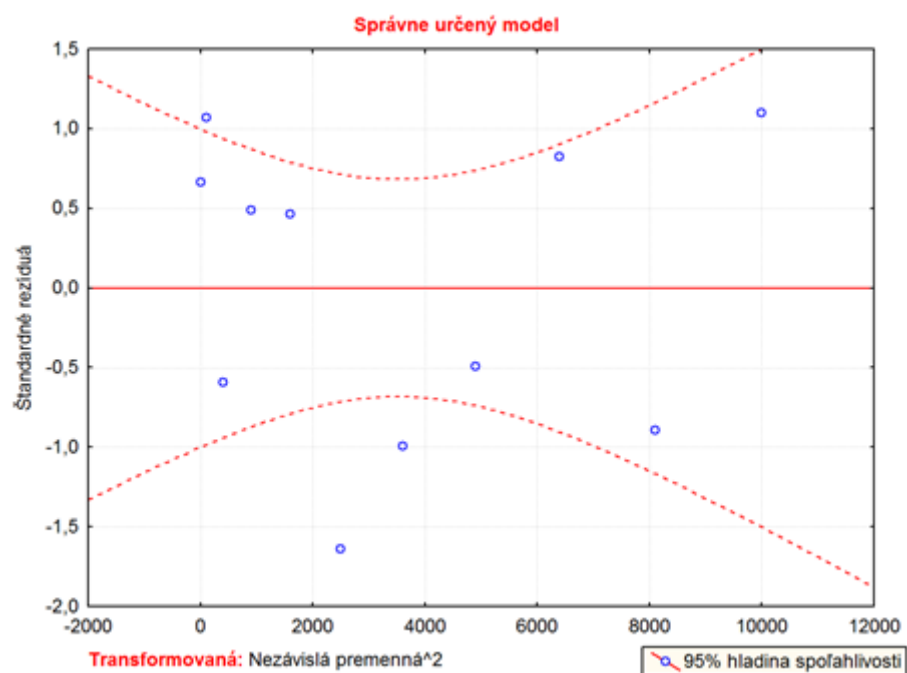


*Body sa zhľukujú okolo paraboly*





*Body sú náhodne rozmiestnené okolo vodorovnej osi*



## 4. Transformácia dát

### Základné transformácie

Transformovať údaje môžeme z viacerých dôvodov, napríklad transformujeme údaje pri prechode na nové jednotky merania, alebo odpočítame od údajov mieru polohy, čím získame centrované dáta.

### Štandardizácia

$$\text{Štandardizovaná hodnota} = (\text{pozorovaná hodnota} - \text{priemer}) / \text{smerodajná odchýlka}$$

Namiesto priemeru môžeme použiť medián a smerodajnej odchýlky kvartilové rozpätie. Dôsledkom štandardizácie je, že priemer (medián) štandardizovaných dát je 0 a ich smerodajná odchýlka (kvartilové rozpätie) je 1.

Dáta so symetrickým rozdelením štandardizované priemerom a smerodajnou odchýlkou sú symetricky rozdelené okolo nuly a ich hodnoty sú približne v rozmedzí -3 až 3. Hodnoty mimo tohto rozmedzia sa pokladajú za podozrivé – možné extrémne hodnoty.

Štandardizácia pomocou priemeru a smerodajnej odchýlky sa dá vyjadriť pomocou lineárnej funkcie:

$$y_i = a + bx_i, \text{ kde } a = -\frac{\bar{x}}{s_x}, b = \frac{1}{s_x} \text{ a } i = 1, 2, \dots, n.$$

### Lineárna transformácia

Lineárna transformácia nemení typ tvaru rozdelenia dát:

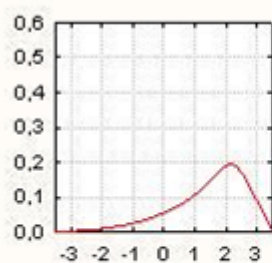
$$y_i = a + bx_i, b > 0,$$

$$\bar{y} = a + b\bar{x}, s_y = bs_x.$$

### Nelineárne transformácie

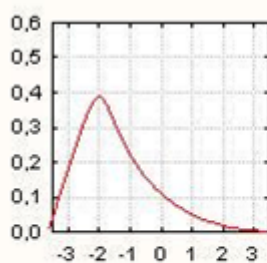
Cieľom nelineárnej transformácie dát je linearizácia nelineárnych vzťahov, dosiahnutie konštantného rozptylu a zmena tvaru rozdelenia.

*Problémy s tvarom rozdelenia, transformácie dát k normalite*



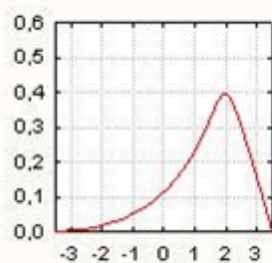
Transformácia:  
 $X = X^3$

Efekt:  
Znižuje extrémne  
zošikmenie doprava



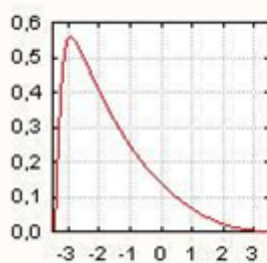
Transformácia:  
 $X = X^{(1/2)}$

Efekt:  
Znižuje zošikmenie  
doľava



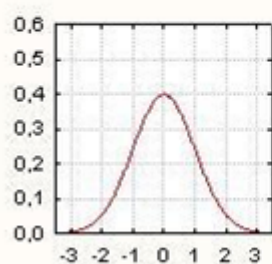
Transformácia:  
 $X = X^2$

Efekt:  
Znižuje zošikmenie  
doprava



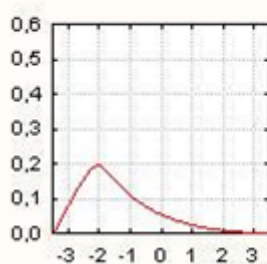
Transformácia:  
 $X = \ln X$

Efekt:  
Znižuje zošikmenie  
doľava



Transformácia:  
 $X = X^1$

Efekt:  
Žiadny účinok



Transformácia:  
 $X = -X^{(-1/2)}$

Efekt:  
Znižuje extrémne  
zošikmenie doľava

## 5. Viacrozmerné prieskumné techniky

Viacrozmerné prieskumné techniky sa zaoberajú analýzou vzťahov medzi skupinami premenných, vo vnútri skupín premenných a rozdielmi správania sa premenných v rôznych subpopuláciách.

Z väčšej časti sa viacrozmerné prieskumné techniky používajú za účelom klasifikácie, resp. segmentácie a znižovania počtu dimenzií. Konkrétne, za účelom segmentácie používame **zhlukovú analýzu**, klasifikácie **diskriminačnú analýzu a klasifikačné stromy** a k znižovaniu počtu dimenzií **faktorovú analýzu, analýzu hlavných komponentov, viacrozmerné škálovanie a korešpondenčnú analýzu**.

Medzi viacrozmerné prieskumné techniky patrí aj **kanonická analýza**, ktorá slúži k posudzovaniu vzťahov medzi dvoma skupinami kvantitatívnych premenných a **analýza spoľahlivosti/položiek** k posúdeniu kvality meracej procedúry. Práve poslednej menovanej analýze sa budeme hlbšie venovať (pozri kapitolu 5) pre jej dôležitý, aj keď často opomínaný význam – ak bude meracia procedúra nekvalitná, tak výsledky nebudú mať žiadnu výpovednú hodnotu, bez ohľadu na to, akú pokročilú metódu na spracovanie použijeme.

Viacrozmerné prieskumné techniky síce zaraďujeme do exploračnej analýzy, avšak treba zdôrazniť, že tieto metódy obsahujú aj prvky inferenčnej analýzy. Napríklad súčasťou analýzy spoľahlivosti je aj analýza rozptylu pre opakované merania, ktorá sa používa napr. za účelom posúdenia obtiažnosti úloh.