Exploračná analýza

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 (Tukey, 1977; Hendl, 2004).

2.1 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.

2.1.1 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 $(x_1, x_2, ..., x_r)$. 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.

Tabuľka 6 Tabuľka početností – jednoduché triedenie

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	$(n_r/n)100$	$\sum_{j=1}^{r} n_j = n$
Σ	***n***	**1**	**100**	

Ak sme správne zostavili tabuľku početností (Tabuľka 6), 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 počítať 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 počítať aj z relatívnych početností, rovnako ako z absolútnych.

Intervalové triedenie

V prípade, že dátový súbor je rôznorodý, zoskupujeme hodnoty do triednych intervalov (Tabuľka 7). Usporiadame hodnoty podľa veľkosti od minimálnej po maximálnu hodnotu.

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

$$r \doteq 1 + 3{,}3 \log n .$$

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 obidvoch strán, mohol by nastať prípad, že by jedna hodnota padla do obidvoch intervalov. Keď zvolíme interval zľava otvorený nezačíname v minimálnej hodnote, ale v hodnote o niečo menšej. Celému triednemu intervalu priradíme hodnotu znaku rovnú stredu intervalu.

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

Tabuľka 7 Tabuľka početností – intervalové triedenie

Triedny interval	**Stred intervalu**	**Absolútna početnosť**	**Relatívna početnosť**	**Relatívna početnosť(%)**	**Kumulatívna početnosť**
$(a_0, a_1]$	x_1	n_1	n_1/n	$(n_1/n)100$	$\sum_{j=1}^{1} n_j = n_1$
$(a_1, a_2]$	x_2	n_2	n_2/n	$(n_2/n)100$	$\sum_{j=1}^{2} n_j$
⋮	⋮	⋮	⋮	⋮	⋮
$(a_{r-2}, a_{r-1}]$	x_{r-1}	n_{r-1}	n_{r-1}/n	$(n_{r-1}/n)100$	$\sum_{j=1}^{r-1} n_j$
$(a_{r-1}, a_r]$	x_r	n_r	n_r/n	$(n_r/n)100$	$\sum_{j=1}^{r} n_j = n$
Σ		***n***	**1**	**100**	

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}, $r = 1, 2, ..., R$, $s = 1, 2, ..., S$ (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_r a početnosti hodnôt premennej *Y* sa označujú ako stĺpcové početnosti s_s

$$r_r = \sum_{s=1}^{S} a_{rs} \,,\; s_s = \sum_{r=1}^{R} a_{rs} \,,\; n = \sum_{r=1}^{R}\sum_{s=1}^{S} a_{rs} \,,$$

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

Tabuľka 8 Kontingenčná tabuľka pozorovaných početností RxS

$X\backslash Y$	y_1	y_2	…	y_S	Σ
x_1	a_{11}	a_{12}	…	a_{1S}	r_1
x_2	a_{21}	a_{22}	…	a_{2S}	r_2
⋮	⋮	⋮	⋮	⋮	⋮
x_R	a_{R1}	a_{R2}	…	a_{RS}	r_R
Σ	s_1	s_2	…	s_S	n

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*x*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%).

2.1.2 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,

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

Geometrický priemer sa definuje vzťahom

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

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.

Harmonický priemer sa definuje vzťahom

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

keď sú všetky hodnoty premennej *X* kladné. Príkladom použitia harmonického priemeru je výpočet priemernej rýchlosti.

Kvadratický priemer sa definuje vzťahom

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

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

Pre priemery platí

$$x_{\min} \le \overline{x}_H \le \overline{x}_G \le \overline{x} \le \overline{x}_K \le x_{\max},$$

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

$$x_1 + \ldots + x_n = \overline{x} + \ldots + \overline{x}.$$

Geometrický priemer slúži k náhrade individuálnych x_i pri násobení, platí

$$x_1 \ldots x_n = \overline{x}_G \ldots \overline{x}_G.$$

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

$$\frac{1}{x_1} + \ldots + \frac{1}{x_n} = \overline{x}_H + \ldots + \overline{x}_H.$$

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

$$x_1^2 + \ldots + x_n^2 = \overline{x}_K^2 + \ldots + \overline{x}_K^2.$$

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

$$\overline{x} = \frac{1}{\sum n_i} \sum n_i x_i.$$

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

$$\tilde{x} = x_m \, .$$

V prípade párneho počtu hodnôt, $n = 2m$, medián je aritmetický priemer prostredných hodnôt, teda

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

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.

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.

2.1.3 Charakteristiky 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,

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

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,

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

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

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

Smerodajná (štandardná) chyba odhadu priemeru predstavuje podiel smerodajnej odchýlky a odmocniny z rozsahu súboru *n*. Čím je vzorka väčšia, tým je chyba menšia (pozri kapitolu 3). Počíta sa podľa vzorca

$$s_{\overline{x}} = \frac{s}{\sqrt{n}} .$$

Variačný koeficient sa používa pre porovnávanie variability viacerých štatistických znakov, predstavuje relatívnu mieru 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

$$v = \frac{s}{\overline{x}}\cdot 100 ,$$

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 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

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

Koeficient disperzie predstavuje relatívnu mieru variability, ktorá je iba málo ovplyvnená extrémnymi hodnotami. Vypočíta sa ako podiel priemernej odchýlky a mediánu,

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

Mediánová absolútna odchýlka 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),

$$MAD = \tilde{x}\{\left|x_i - \tilde{x}\right|\}.$$

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 (pozri kapitolu 2.1.5). 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.

2.1.4 Charakteristiky tvaru

Tvar rozdelenia dát hodnotíme charakteristikami tvaru – šikmosťou a špicatosťou.

Šikmosť 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

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

Špicatosť 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:

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

2.1.5 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 najpoužívanejší 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 najmenšej 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 $j = 1, 2, ..., n$.

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 $x_{0,75}$.

2.2 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.

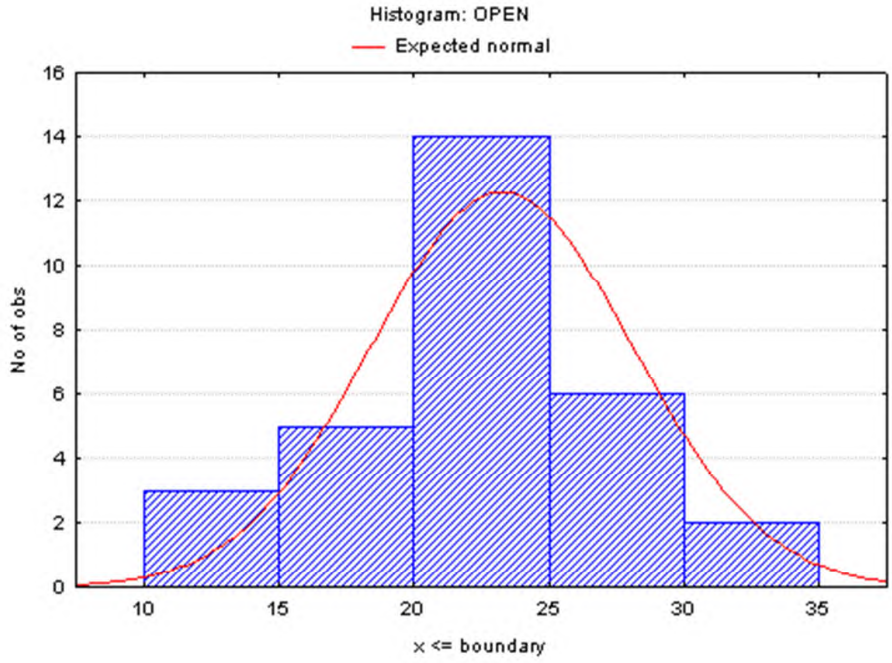
Pri analýze grafu hodnotíme: zhustenia, zhluky, medzery, extrémne hodnoty a tvar rozdelenia.

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.

2.2.1 Vizualizácia početností

Polygón (Obrázok 13) slúži k vizualizácii početností pri jednoduchom triedení a **histogram** (Obrázok 11) sa zvyčajne používa pri intervalovom triedení.



Obrázok 11 Histogram

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 (Obrázok 12) 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 interval <20, 25) sa nachádza iba jedna hodnota 2° 4 = 24.

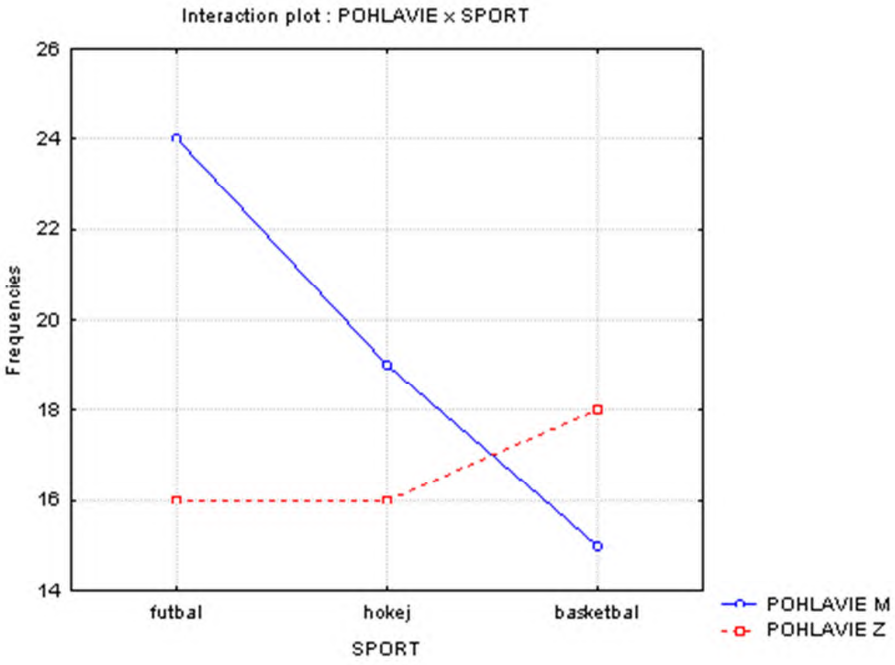
stonka° list	N	
2° 4	1	
2° 55577999	8	
3° 00124	5	25%
3° 5556667777899999	16	medián
4° 000111222344444	15	75%

4°	556	.	.	.	.	3	
5°		.	.	.	.	0	
min = 24,00000		max = 46,00000			Total N:	48	

Obrázok 12 Listový graf

Kruhový diagram predstavuje najvhodnejšie znázornenie relatívnych početností alebo počtu percent.

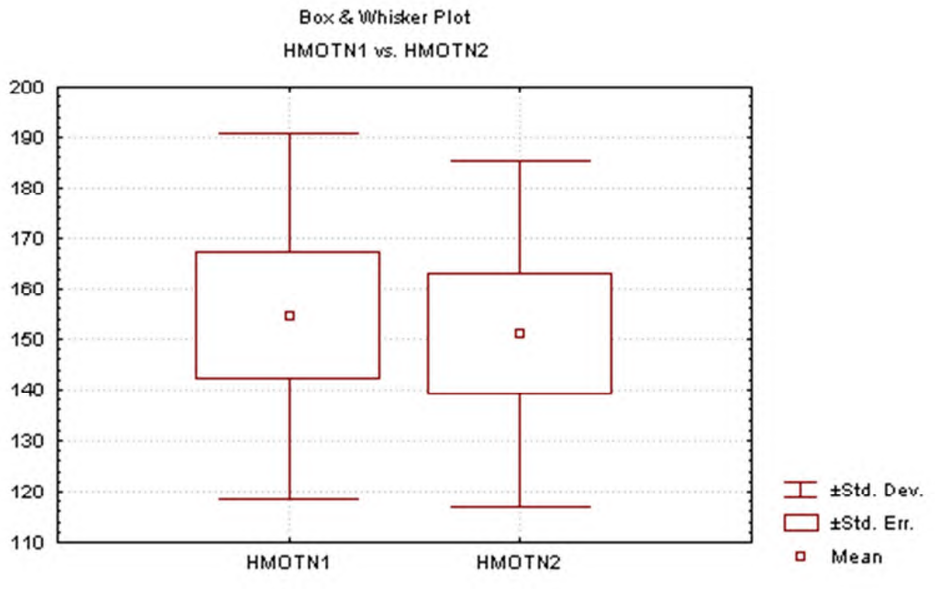
Interakčné grafy (Obrázok 13) predstavujú najlepšiu možnosť vizualizácie kontingenčnej tabuľky. Na osi *X* sú hodnoty jednej premennej a na osi *Y* sú početnosti. V grafe sú zobrazené dva polygóny pre každú úroveň druhej kategorickej premennej, čím sú vyjadrené interakčné početnosti medzi dvoma premennými. 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.



Obrázok 13 Interakčný graf

2.2.2 Vizualizácia popisných charakteristík

K zobrazovaniu popisných štatistík je vhodný **krabicový graf** (Obrázok 14). Pomocou tohto grafu môžeme posudzovať a porovnávať miery polohy a rozptýlenosť hodnôt okolo nich, taktiež môžeme posudzovať zošikmenie a prítomnosť extrémnych hodnôt.



Obrázok 14 Krabicový graf

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,**

Tabuľka 9 Päť hodnôt krabicového grafu medián/kvartilové rozpätie/variačné rozpätie

1.	2.	3.	4.	5.
minimum	dolný kvartil	medián	horný kvartil	maximum

x_{min}	Q_I	$\tilde{x}$	Q_{III}	x_{max}

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

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

*Tabuľka 10 Päť hodnôt krabicového grafu 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_I - 1,5*Q	Q_I	$\tilde{x}$	Q_{III}	Q_{III} + 1,5*Q

Tento typ krabicového grafu sa odlišuje od predchádzajúceho tým, že interval je tvorený poslednou hodnotou pod Q_{III} + 1,5*Q a poslednou hodnotou premennej nad Q_I - 1,5*Q (Tabuľka 10). 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,**

Tabuľka 11 Päť hodnôt krabicového grafu 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
$\bar{x} - s$	$\bar{x} - s_{\bar{x}}$	$\bar{x}$	$\bar{x} + s_{\bar{x}}$	$\bar{x} + s$

Interval predstavuje $\bar{x} \pm s$, väčší obdĺžnik $\bar{x} \pm s_{\bar{x}}$ a menší priemer $\bar{x}$ (Tabuľka 11).

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

*Tabuľka 12 Päť hodnôt krabicového grafu 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
$\bar{x} - 1{,}96 * s_{\bar{x}}$	$\bar{x} - s_{\bar{x}}$	$\bar{x}$	$\bar{x} + s_{\bar{x}}$	$\bar{x} + 1{,}96 * s_{\bar{x}}$

Interval predstavuje $\bar{x} \pm 1{,}96 * s_{\bar{x}}$, väčší obdĺžnik $\bar{x} \pm s_{\bar{x}}$ a menší priemer $\bar{x}$ (Tabuľka 12).

Interval zobrazuje 95% interval spoľahlivosti priemeru (pozri kapitolu 3).

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

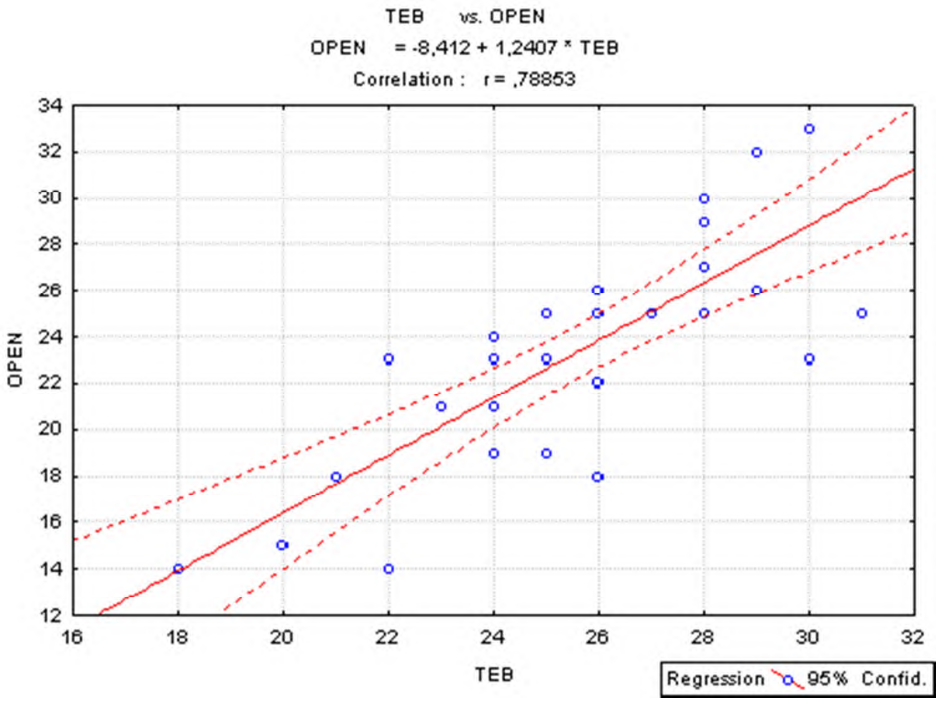
*Tabuľka 13 Päť hodnôt krabicového grafu aritmetický priemer/smerodajná odchýlka/1,96*smerodajná odchýlka*

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 predstavuje $\bar{x} \pm 1{,}96 * s$, väčší obdĺžnik $\bar{x} \pm s$ a menší priemer $\bar{x}$ (Tabuľka 13). Interval zobrazuje 95% interval spoľahlivosti pre individuálne pozorovania okolo priemeru (pozri kapitolu 3).

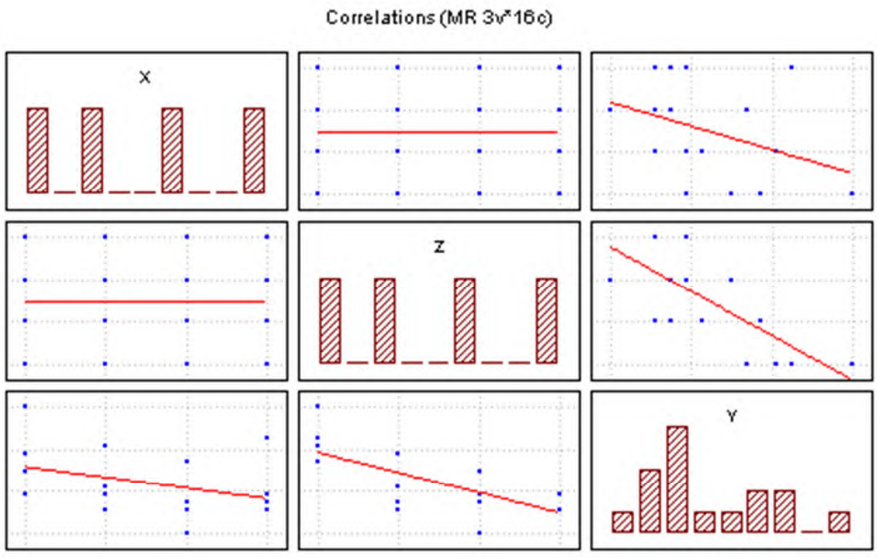
2.2.3 Vizualizácia závislostí

Korelačný/bodový graf (Obrázok 15) 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.



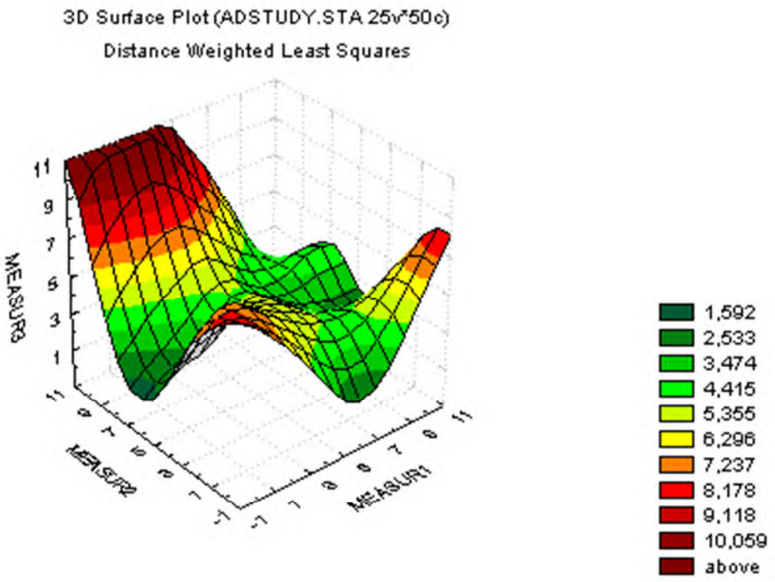
Obrázok 15 Korelačný/bodový graf

Maticový graf (Obrázok 16) je vizualizáciou lineárnej závislosti medzi viacerými intenzívnymi/kvantitatívnymi premennými.



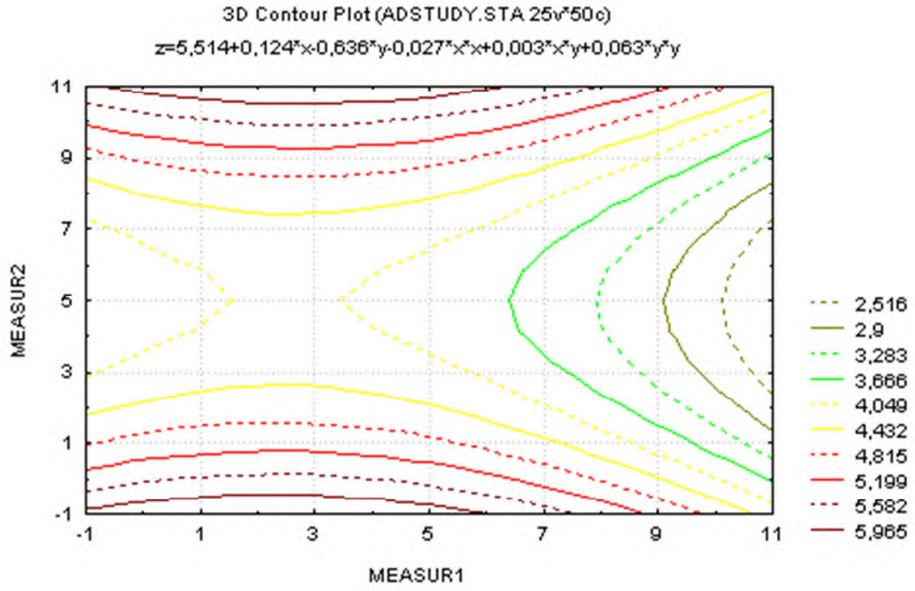
Obrázok 16 Maticový graf

Povrchový graf (Obrázok 17) vizualizuje funkčnú závislosť intenzívnych/kvantitatívnych premenných - závislej premennej *Y* na nezávislých premenných X_1 a X_2.



Obrázok 17 Povrchový graf

Vrstevnicový graf (Obrázok 18) 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).



Obrázok 18 Vrstevnicový graf

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.

2.2.4 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** (pozri kapitolu 3.1.4).

2.3 Analýza reziduálnych hodnôt

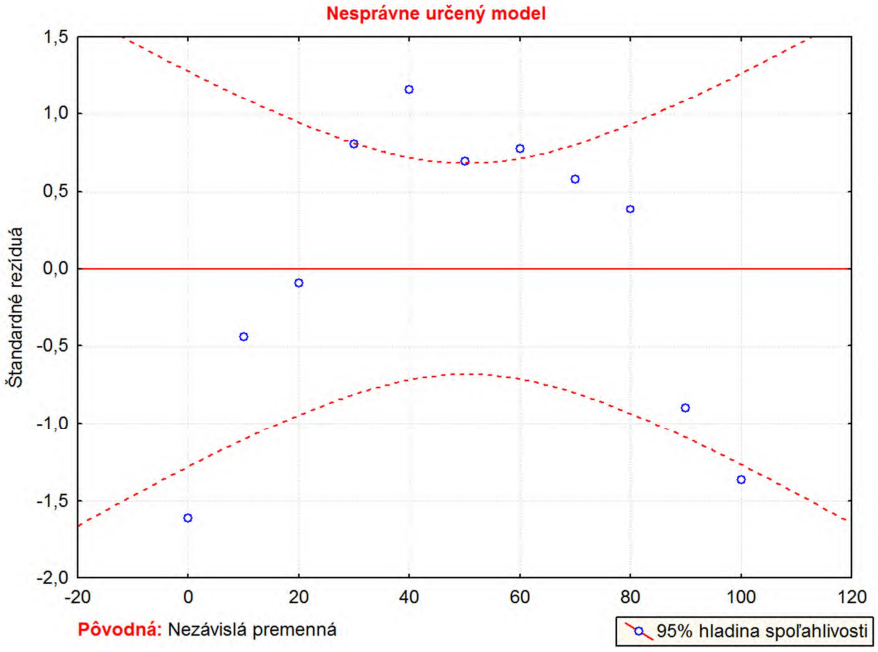
Analýza reziduálnych hodnôt vychádza zo základnej predstavy:

Dáta = predikcia modelom (funkcia) + reziduálna hodnota.

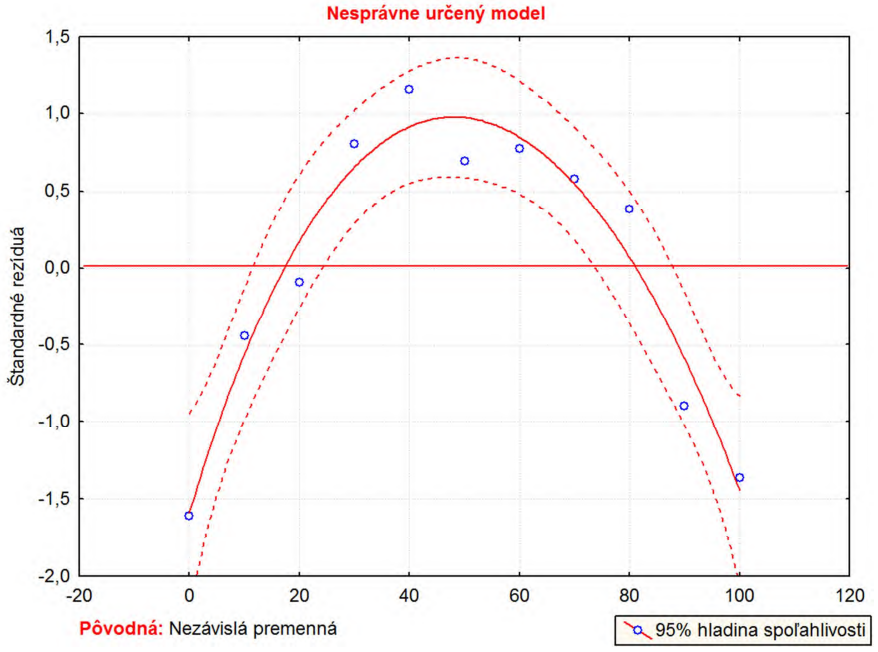
Ak od dát (pozorované hodnoty) odpočítame hodnoty získané z modelu (očakávané hodnoty), získame chyby (reziduálne hodnoty) a ich analýzou môžeme posúdiť zostrojený model.

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

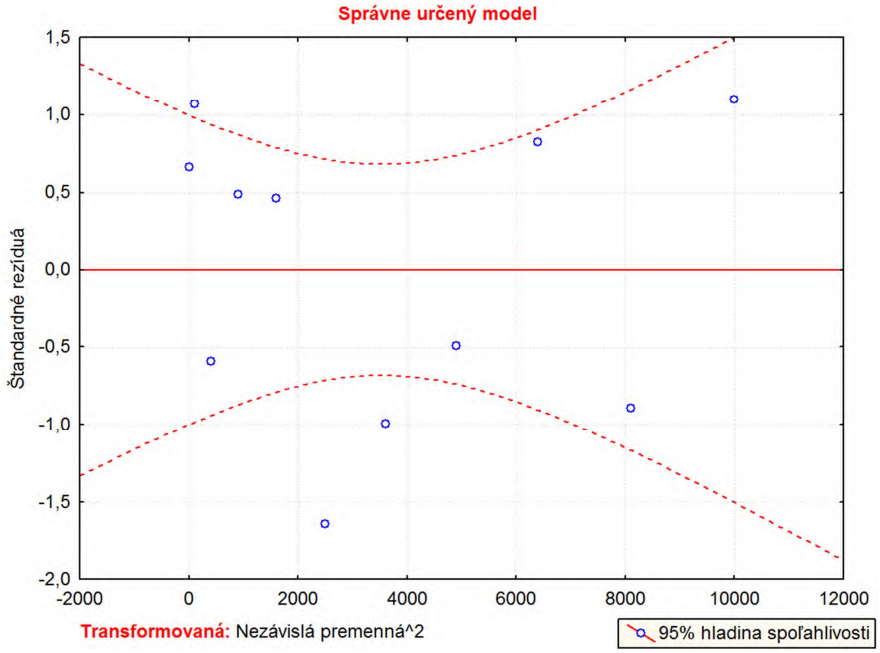
Napríklad pomocou analýzy reziduálnych hodnôt môžeme overiť stabilitu regresného modelu, t.j. identifikovať nesprávnosť zvoleného modelu – prostredníctvom zobrazenia korelačného/bodového grafu štandardných rezíduí a nezávislej premennej. Keď je model správne určený, tak sú body náhodne rozmiestnené okolo vodorovnej osi. Ak sa body zhlukujú okolo nejakej priamky rôznej od vodorovnej osi alebo okolo nejakej krivky, je to známka nesprávne zvoleného modelu. V našom prípade sa body zhlukujú okolo paraboly, je nutné transformovať nezávislú premennú $X = X^2$ (Obrázok 19, Obrázok 20, Obrázok 21).



Obrázok 19 Body nie sú náhodne rozmiestnené okolo vodorovnej osi



Obrázok 20 Body sa zhlukujú okolo paraboly



Obrázok 21 Body sú náhodne rozmiestnené okolo vodorovnej osi

2.4 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

Štandardizovaná hodnota = (pozorovaná hodnota – priemer)/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:

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

Lineárna transformácia

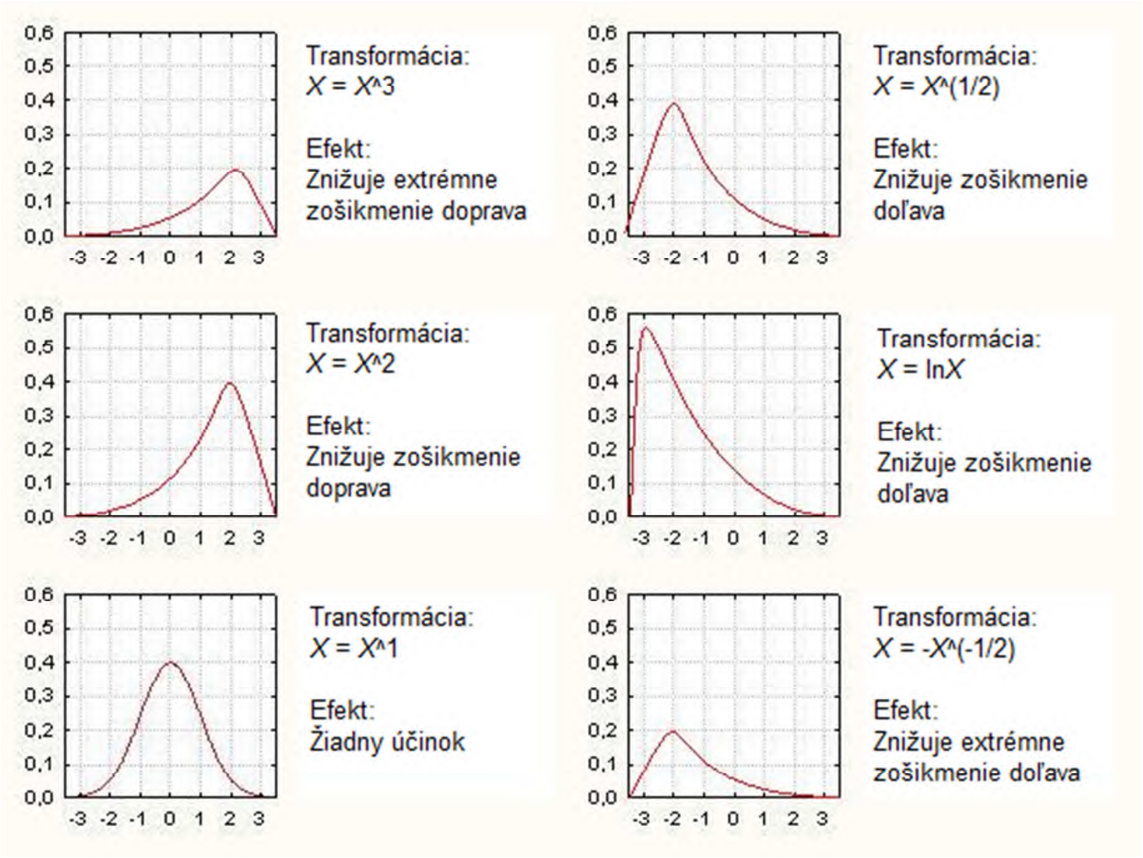
Lineárna transformácia nemení typ tvaru rozdelenia dát:

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

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

Nelineárne transformácie

Cieľom nelineárnej transformácie dát je linearizácia nelineárnych vzťahov, dosiahnutie konštantného rozptylu a zmena tvaru rozdelenia (Obrázok 22).



Obrázok 22 Problémy s tvarom rozdelenia, transformácie dát k normalite

2.5 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 analýza spoľahlivosti je aj analýza rozptylu pre opakované merania, ktorá sa používa napr. za účelom posúdenia obtiažnosti úloh.