Reprezentácia dát

Ing. Martin Mariš, Katedra regionalistiky a rozvoja vidieka, SPU, NITRA

- slovným opisom
- grafickým zobrazením

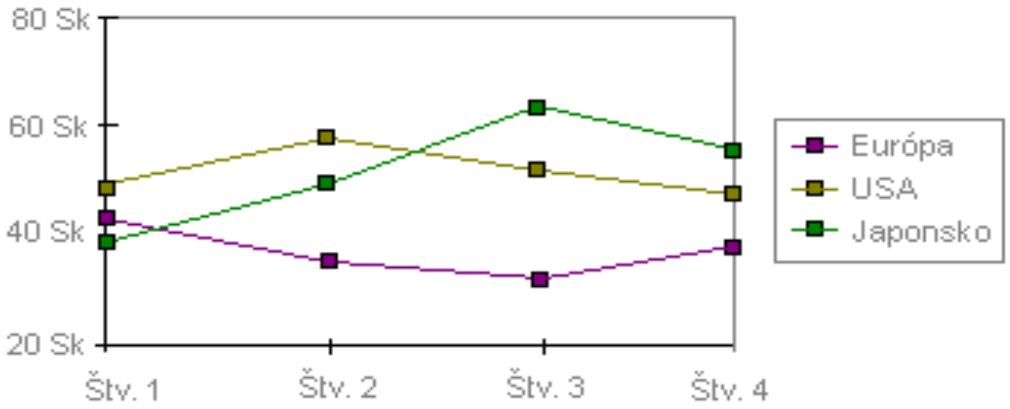
Typy grafov a ich použitie

Najčastejšie používané typy grafov:

- _čiarový graf
- _stĺpcový graf (horizontálny, vertikálny)
- _kruhový (koláčový)
- bodový
- plošný
- radarový
- povrchový, burzový atď...

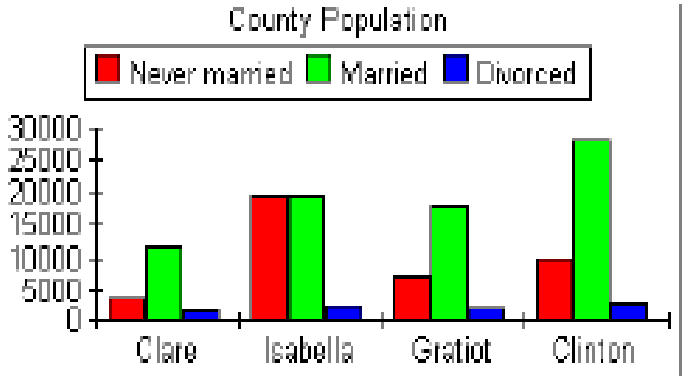
Čiarový graf

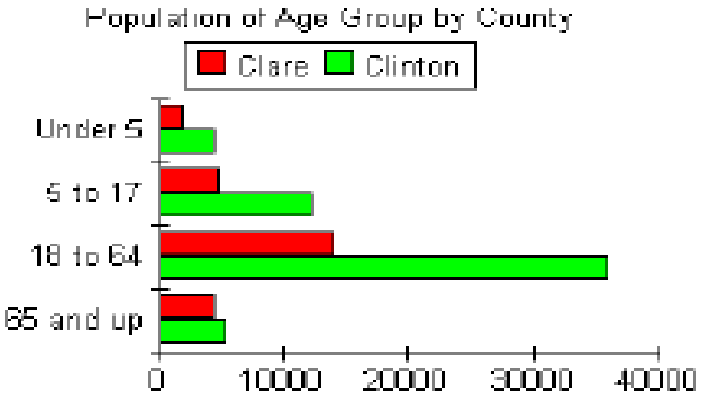
- Čiarový graf zobrazuje trendy vývoja údajov v rovnakých časových intervaloch



Horizontálne a vertikálne stĺpcové grafy

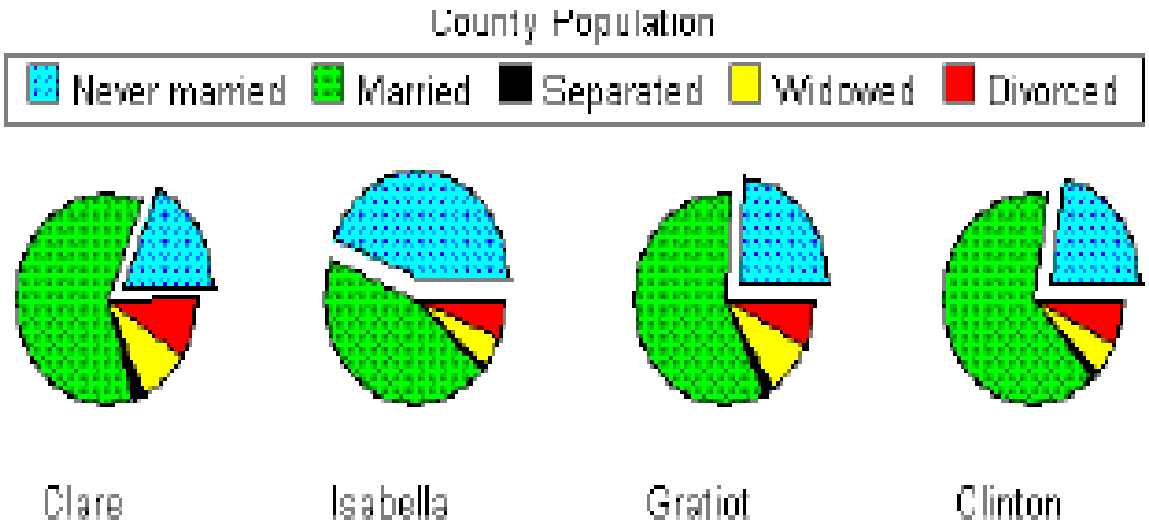
- sú vhodné na porovnávanie hodnôt a znázorňovanie trendov





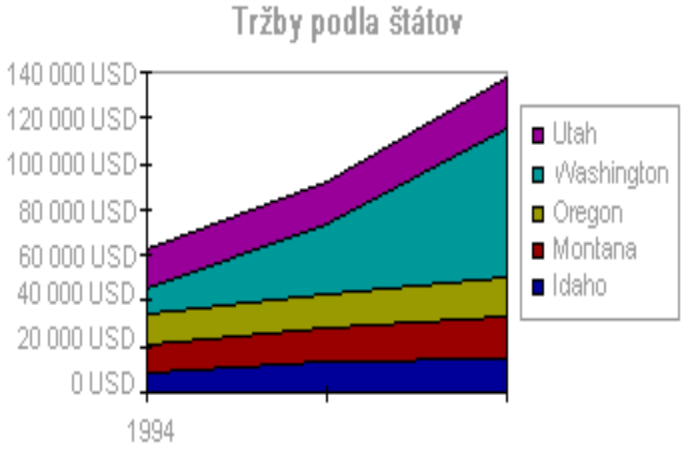
Kruhový (koláčový) graf

- ilustrujú vzťah medzi časťou a celkom a sú vhodné na znázorňovanie pomerov a proporcií



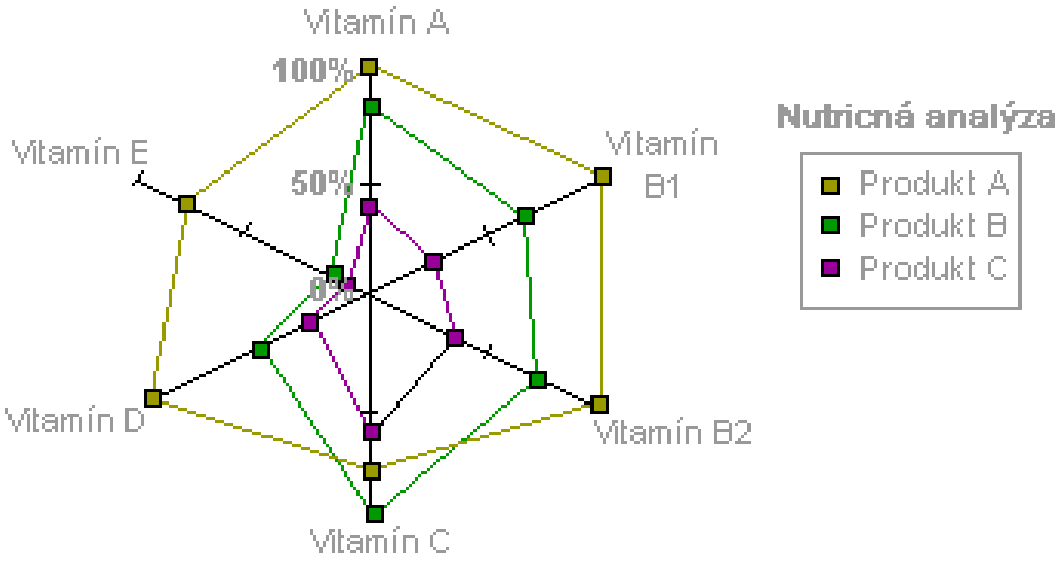
Plošný graf

- plošný graf znázorňuje rozsah zmien za určitú dobu, súčet hodnôt môže vyjadrovať i vzťah častí k celku



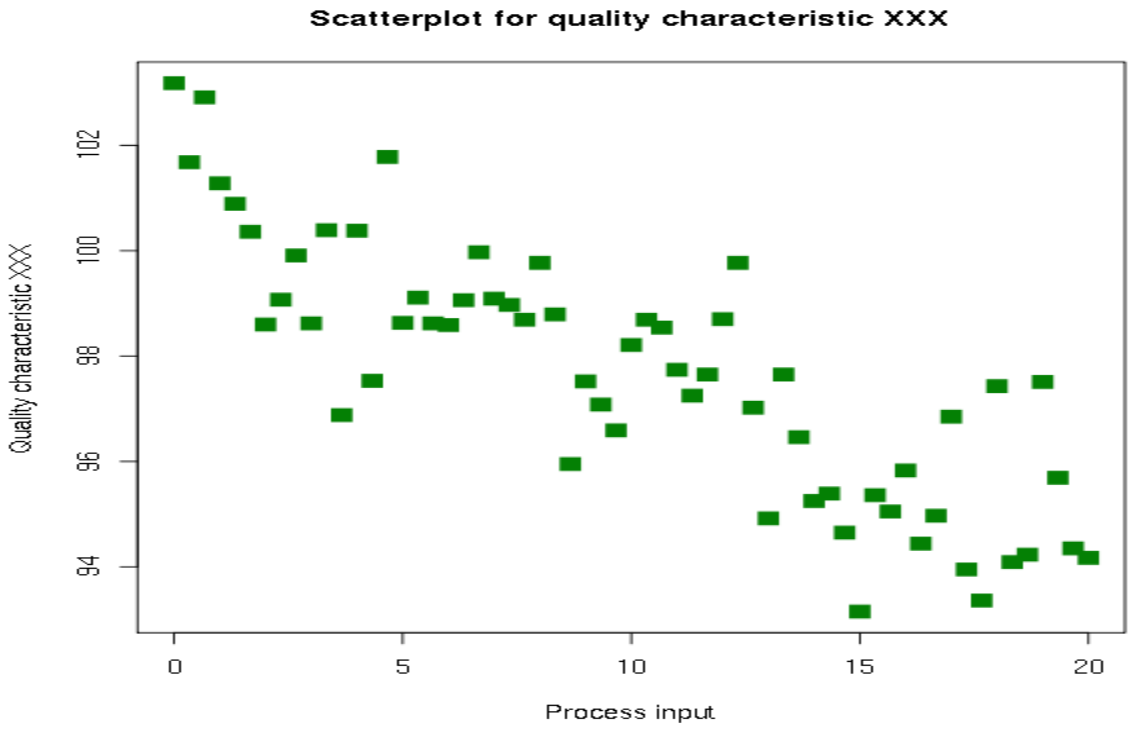
Radarový graf

- radarový graf sa používa na porovnanie súhrnných hodnôt niekoľkých radov údajov



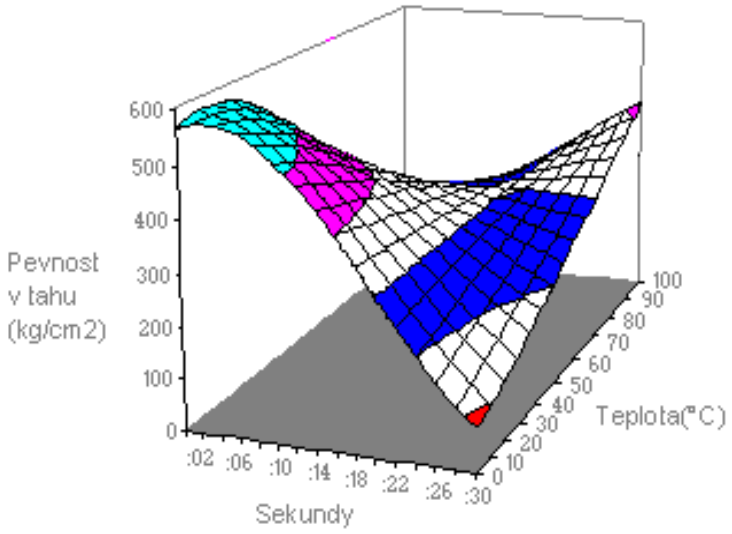
Bodový graf

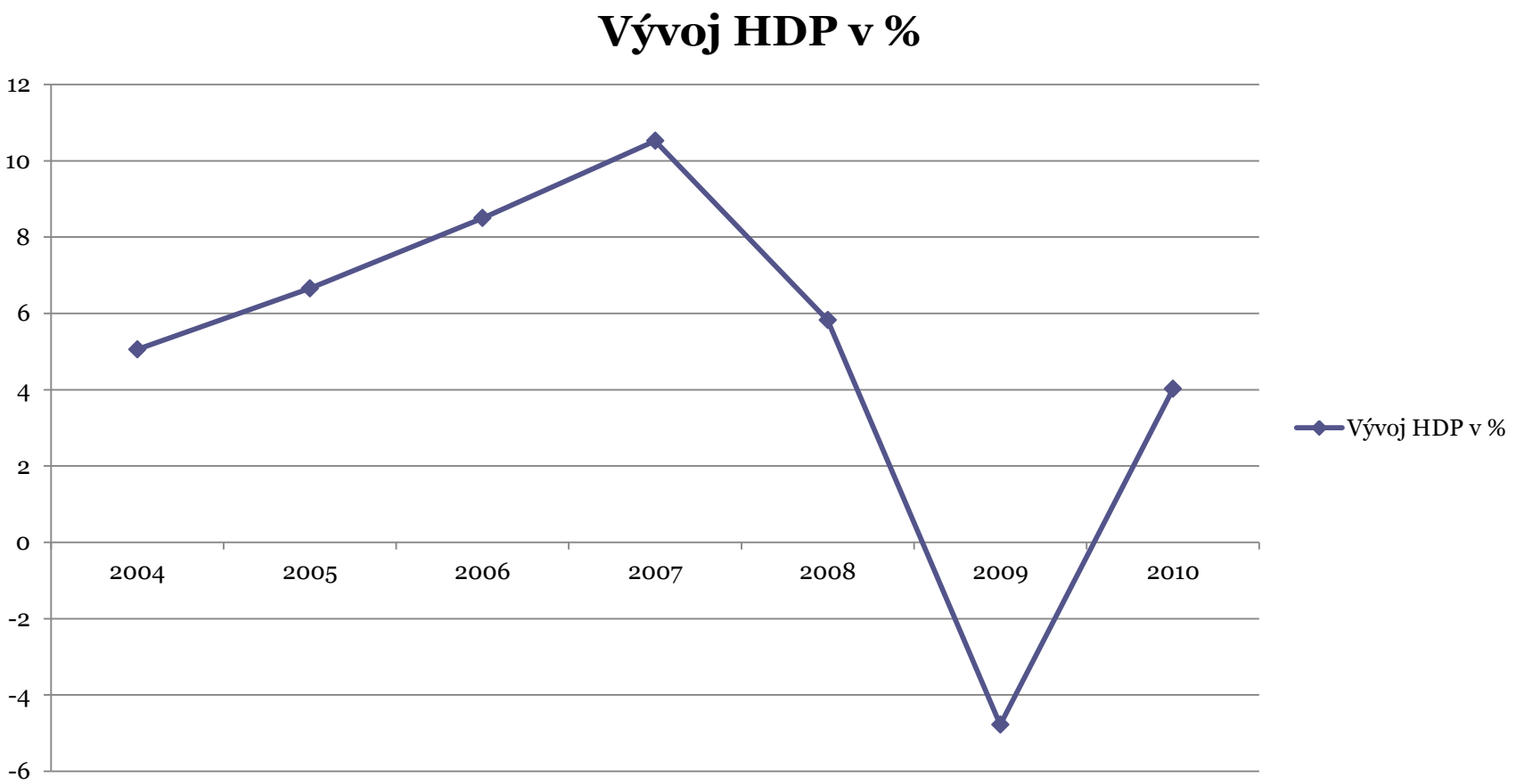
- odhaľuje trendy alebo štruktúru údajov a pomáhajú zisťovať príčinné súvislosti



Plošný graf

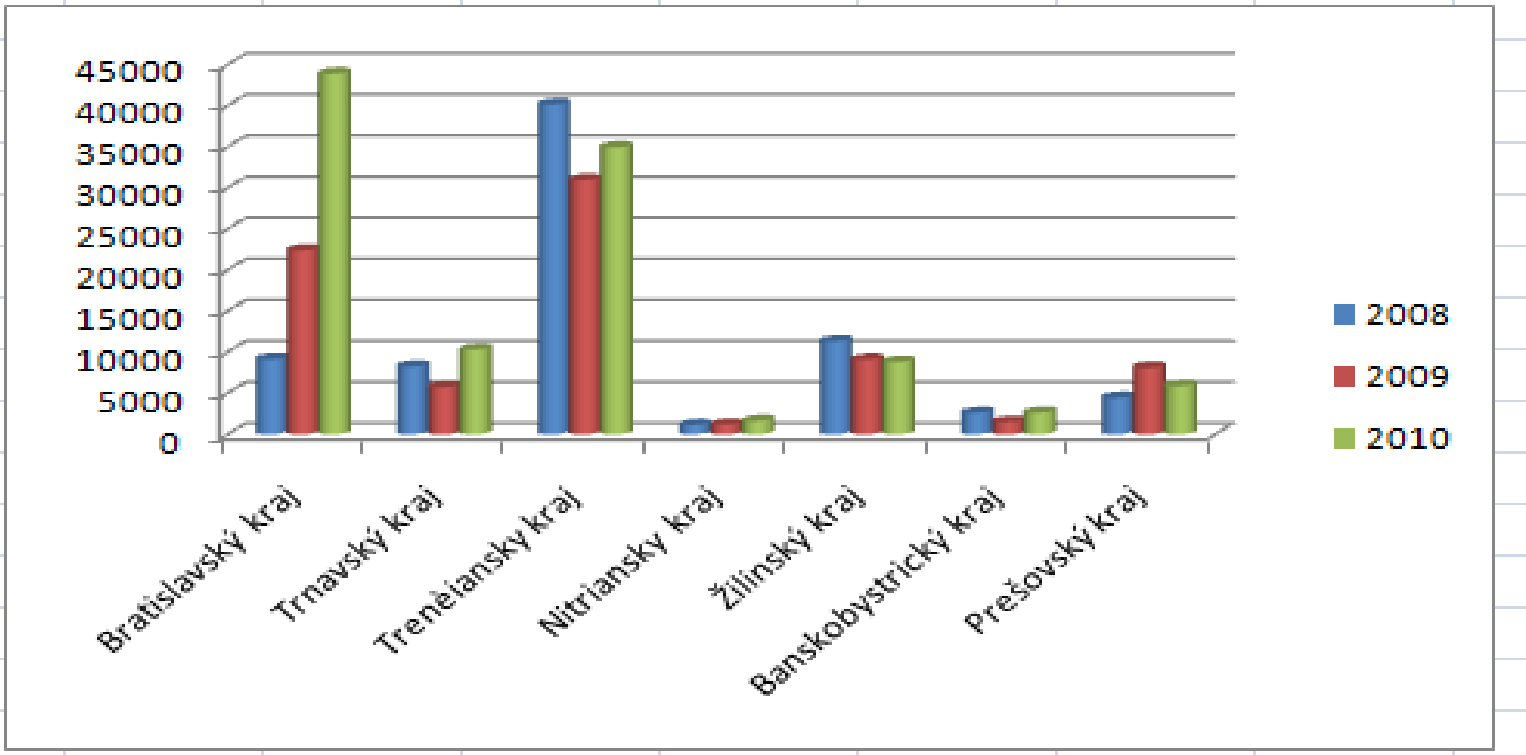
- Povrchový graf sa používa pri hľadaní optimálnej kombinácie medzi dvomi množinami údajov





Interpretácia: Na danej krivke grafu možno pozorovať rastúci trend vývoja HDP od r.2004,kt. konštantne rástol priemerným tempom 7,4% až do r.2009, kedy došlo k jeho prudkému prepadu o 4,7%. V r.2010 sa trendová línia opätovne vrátila k rastu.

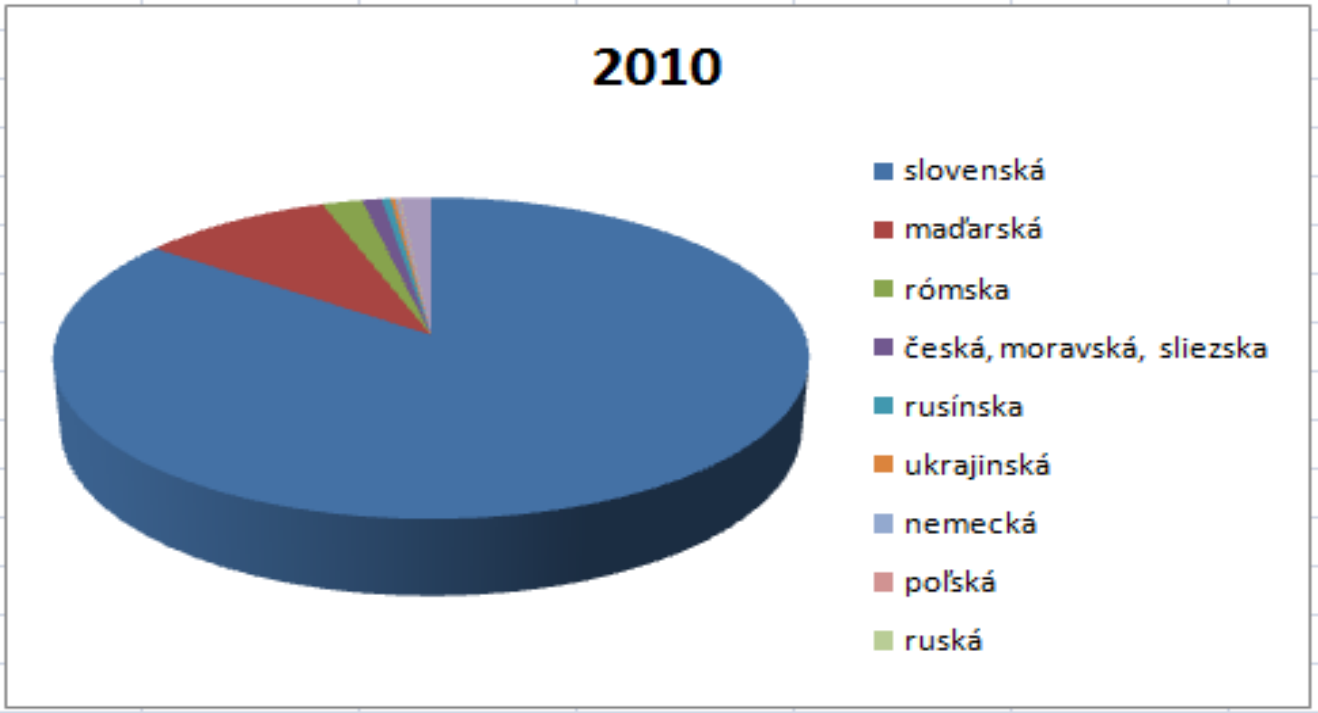
Obr.1: Výdavky na vývoj podľa krajov v tis.eur



Zdroj : Regdat

Na danom obrázku možno vidieť prehľad vývoja výdavkov na výskum v r. 2008 – 2010 podľa krajov. Najväčší rast výdavkov zaznamenal BSK, na úrovni 380% oproti r.2008, čo v absolútnom vyjadrení predstavuje 34 769 tis.eur. Najmenší rast výdavkov zaznamenal BBSK , len o 1,49% v absolútnom vyjadrení 38 tis. eur....atď....

Obr.2.: Národnostná štruktúra obyvateľstva v r.2010



Zdorj: Regdat

Z hľadiska národnostnej štruktúry obyv. v SR dominujú občania Slovenskej národnosti na úrovni 85%, nasledujúc Maďarský občania na úrovni 9% atď...

Opisné štatistiky

- slúžia na opis výberového súboru prostredníctvom polohy a variability štatistického (skúmaného) znaku vo výberovom súbore
- stredné hodnoty
- miery variability

Kvantily

- Kvantily Q sú hodnoty náhodnej premennej, ktoré delia jej rozdelenie pravdepodobnosti na α častí, pričom každá časť má pravdepodobnosť $\frac{1}{\alpha}$
- najčastejšie kvantily: kvartil, decil, percentil ...

❑ Excel:

- funkcia QUARTILE
- výber poľa
- výber kvartilu: Q_1, Q_2, Q_3, Q_4

Stredné hodnoty

- čísla charakterizujúce úroveň hodnôt znaku v štatistickom súbore, nachádzajú sa medzi minimálnou a maximálnou hodnotou znaku v súbore, sú jednoznančne určené a majú vecne informatívny zmysel
- priemery
- ďalšie stredné hodnoty

- aritmetický priemer -$\bar{x}$ - najčastejšie používaná stredná hodnota (prvý začiatočný moment premenne X)
- jednoduchý aritmetický priemer

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

- vážený aritmetický priemer

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

❑ Excel:

- Funkcia AVERAGE (jednoduchý priemer)
- predstavuje hodnotu, ktorá je pre daný súbor reprezentatívna
- necitlivý k extrémnym veličinám

Harmonický priemer

- jednoduchý harmonický priemer

$$\overline{x_h} = \frac{n}{\sum_{j=1}^{m} \frac{1}{x_j}}$$

- vážený harmonický priemer

$$\overline{x_h} = \frac{n}{\sum_{i=1}^{m} \frac{n_i}{x_i}}$$

- používa sa vtedy ak je medzi hodnotami znaku a výsledným javom nepriamy vzťah
- Excel: funkcia HARMEAN

- príklad: určte priem. čas na výrobu 1 súčiastky, ak vieme, že sa používajú 2 stroje: na staršom trvá výroba 6 min. na novšom 4 min.
- medzi počtom vyrobených súčiastok a časom potrebným n výrobu je nepriamy vzťah, použijeme harmonický priemer

$$\overline{x_h} = \frac{n}{\sum_{j=1}^{m} \frac{1}{x_j}} = \frac{2}{\frac{1}{6}+\frac{1}{4}} = 4{,}8 \text{ min.}$$

- predpokladajme, že v podniku sú k dispozícii 3 stroje staršieho typu a 2 modernejšieho typu, použijeme vážený harmonický priemer

$$\overline{x_h} = \frac{n}{\sum_{i=1}^{m} \frac{n_i}{x_i}} = \frac{5}{\frac{3}{6}+\frac{2}{4}} = 5 \text{ min.}$$

Geometrický priemer

- používa sa na spriemerovanie veličín, medzi ktorými je multiplikatívny vzťah. V praxi sa najčastejšie používa pri výpočte priemerného koeficienta rastu (hdp, inflácie atď.)
- jednoduchý geometrický priemer

$$\overline{x_g} = \sqrt[n]{x_1 * x_2 \dots \dots x_n} = \sqrt[n]{\prod_{i=1}^{m} x_n}$$

- jednoduchý vážený geometrický priemer

$$\overline{x_g} = \sqrt[n]{x^{n_1} * x^{n_2} \dots \dots x^{n_m}} = \sqrt[n]{\prod_{i=1}^{m} xi^{n_i}}$$

- Excel : funkcia GEOMEAN

Medián, modus

- jeden z kvantilov, rozdeľujúci štatistický súbor na 2 rovnako početné časti
- nepárny počet štatistický jednotiek:

$$\tilde{x}=x_r,\ \mathrm{r}=\frac{n+1}{2}$$

- párny počet štatistický jednotiek:

$$\tilde{x}=x_r,\ \mathrm{r}=\frac{n}{2}$$

- intervalové rozdelenie početnosti

$$\tilde{x}=a+h*\frac{\frac{n+1}{2}-\sum_{i=1}^{r-1}n_i}{n_{\tilde{x}}}$$

a –dolná hranica intervalu
h – rozpätie mediánového intervalu
$\sum_{i=1}^{r-1}n_i$ - absolútna početnosť po mediánový interval
$n_{\tilde{x}}$- absolútna početnosť mediánového intervalu

- medián je necitlivý voči extrémnym veličinám
- Excel: funkcia MEDIAN
- Modus je najčastejšie vyskytujúca sa hodnota v empirickom štatistickom súbore
- Excel: funkcia MODE

Miery variability

- variačné rozpätie – rozdiel medzi max. a min. hodnotou znaku

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

- kvantilové rozpätie – rozdiel medzi horným a dolným kvantilom

$$R_q = Q_{\alpha-1} - Q_1$$

Rozptyl a štandardná (smerodajná) odchýlka

- Rozptyl - s^2 - druhý centrálny moment, definovaný ako aritmetický priemer zo štvorcov odchýlok hodnôt znaku od aritmetického priemeru
- jednoduchý

$$s^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2$$

- vážený

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_j - \bar{x})^2 \, n_i$$

- Excel: funkcia VAR (jednoduchý)
- Štandardná (smerodajná odchýlka) – vyjadruje variabilitu súboru v pôvodných merných jednotkách, v štatistickom skúmaní má veľké uplatnenie

$$s = \sqrt{s^2}$$

- Excel: funkcia STDEV

Miery asymetrie

- šikmosť rozdelenia špecifikuje, že sa častejšie vyskytujú nižšie hodnoty ako vyššie (zošikmenie vľavo) alebo, že sa častejšie vyskytujú vyššie hodnoty ako nižšie (zošikmenie vpravo)
- ak šikmosť nadobúda hodnotu „0“ ide o symetrické rozdelenie ak $x < 0$ – zošikmenie vpravo a ak $x > 0$ –zošikmenie vľavo
- Excel: funkcia SKEW
- špicatosť rozdelenia porovnáva tvar s normálnym rozdelením, ak $x > 0$- špicatejšie rozdelenie ako normálne, ak $x < 0$ – plochšie rozdelenie ako normálne
- čím vyššia špicatosť tým viac sú znaky sústredené v okolí nejakej hodnoty znaku
- Excel: funkcia KURT

Štandardizácia hodnôt

- stanovenie polohy štatistického znaku vo výberovom súbore

$$z_i = \frac{x_j - \bar{x}}{s_x}$$

- $\bar{x}$ - priemer
- s_x- smerodajná odchýlka
- ± 1- 68% hodnôt znaku
- ±2 – 95% hodnôt znaku
- ±3 – 99,7% hodnôt znaku
- ±4> extrémne veličiny
- Excel: funkcia STANDARDIZE
- priemer
- smerodajná odchýlka

Súhrnný štatistický opis výberového súboru

- Excel:
  - Údaje
  - Analýza dát
  - Opisné štatistiky (popisná štatistika)

Analýza štatistickej závislosti

- Korelačná analýza
- koeficient korelácie
- Excel: funkcia CORREL
- výber poľa 1.premennej
- výber poľa 2.premennej
- 0 – žiadna závislosť medzi premennými
- +1 – kladná korelačná závislosť (priama závislosť), s rastom 1. premennej rastie i 2.premenná
- -1 – záporná korelačná závislosť (nepriama závislosť) s rastom 1.premennej 2.premenná klesá

Korelačná matica

- Excel: funkcia CORRELATION
- analýza údajov
- CORRELATION

Regresné modelovanie závislosti

- príčinna závislosť
- príčinou je nezávislá premenná X
- následkom je nezávislá premenná Y

$$Y = f(x)$$

$$\hat{y} = b_0 + b_1 x1 + b_2 x2 + \cdots b_m xm$$

- b_0, $b_1 \ldots b_m$ sú regresné koeficienty
- x – nezávislá premenná
- $x1$, $x2 \ldots xm$
- $\hat{y}$ - odhad hodnôt závislej premennej podľa regresnej funkcie
- Excel: funkcia REGRESSION

❑ dôležité:

- Multiple R –miera intenzity korelácie
- R Square – variabilita závislej premennej vysvetlená nezávislými premennými
- Significance F – hladina významnosti, ak $<\alpha=0{,}05$ model je štatisticky významný