Metódy spracovania dát v kvantitatívnom výskume

Seminár k diplomovej práci

Peter Vankúš

Základný súbor vs. výberový súbor

- **Základný súbor** (inakšie tiež zvaný populácia) tvoria všetky subjekty, na ktoré sa majú výsledky výskumu vzťahovať. Ak sa majú výsledky vzťahovať na všetkých siedmakov v SR, potom základný súbor tvoria všetci žiaci siedmych ročníkov v SR.

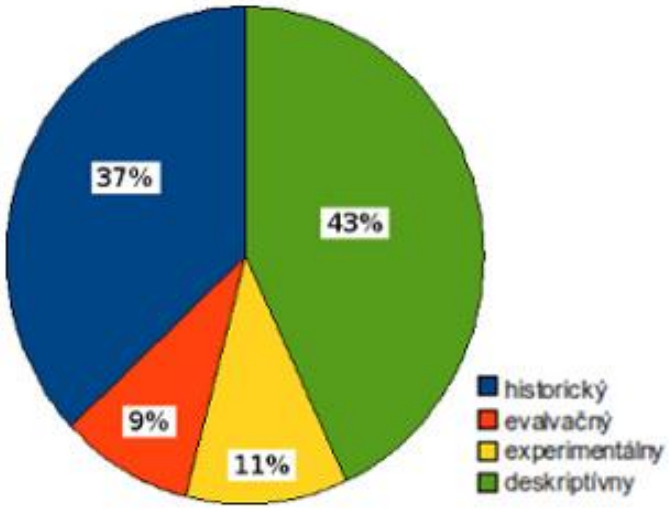
Základný súbor vs. výberový súbor

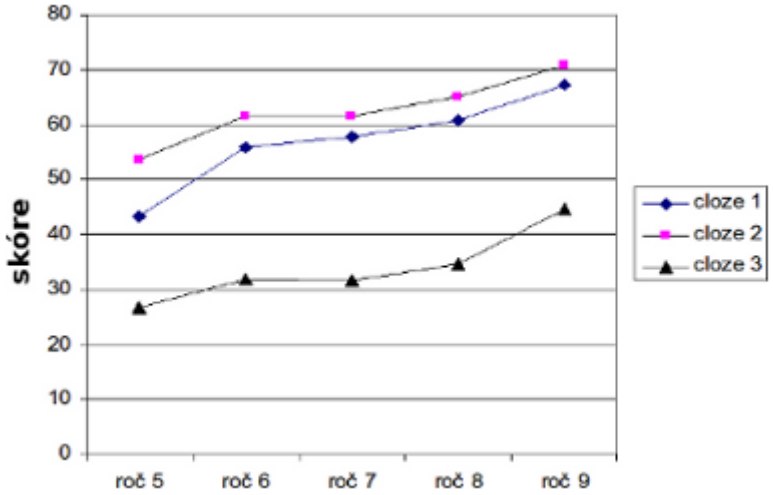
- Je zrejmé, že výskumník nemá možnosť zorganizovať výskum s celým základným súborom. Preto z nich urobí určitý výber. Osoby, ktoré vybral, nazývame, **výberový súbor** (alebo inakšie **výskumná vzorka**).

Gavora, Peter a kol. 2010. Elektronická učebnica pedagogického výskumu. [online]. Bratislava : Univerzita Komenského, 2010. Dostupné na: http://www.e-metodologia.fedu.uniba.sk/ ISBN 978–80–223–2951–4.

Opis súboru – deskriptívna štatistika

- Prehľadové tabuľky
- Diagramy





Opisné charakteristiky

- Absolútne a relatívne početnosti
- Stredné hodnoty: aritmetický priemer, modus, medián
- Minimum, maximum

	počet	%
do 18	11	10,0
19-24	46	41,8
25-29	13	11,8
30-39	16	14,5
40-49	10	9,2
50-	14	12,7
Spolu	110	100

Opisné charakteristiky

- **Aritmetický priemer (arithmetic mean)** je najčastejšie používaná stredná hodnota.

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

$$\bar{x}_w = \frac{x_1 . w_1 + x_2 . w_2 + \ldots + x_n . w_n}{w_1 + w_2 + \ldots + w_n}$$

Príklad: Známky z matematiky v určitej triede sú uvedené v tabuľke

známka	1	2	3	4	5
počet žiakov s danou známkou	14	6	5	4	1

Opisné charakteristiky

- Aritmetický priemer je citlivý na neobyčajne malé alebo veľké hodnoty
- Aritmetický priemer je citlivý na hrubé chyby
- Aritmetický priemer by sa nemal brať do úvahy, ak
  - 1. je štatistické rozdelenie dát viacvrcholové
  - 2. je štatistické rozdelenie dát asymetrické

KULČÁR, LADISLAV. Harmonický priemer a jeho praktická aplikácia [online]. http://www.sachovespravy.eu, [cit. 2022-11-21]. Dostupné online.

Opisné charakteristiky

- **Geometrický priemer** $\overline{x}_G$

$$\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdots x_n} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$$

Opisné charakteristiky

- Je zrejmé, že geometrický priemer má zmysel iba pre dáta, v ktorých sú všetky hodnoty kladné čísla.
- Geometrický priemer sa na rozdiel od aritmetického priemeru používa na koeficienty (napr. percentá).
- Napr. na výpočet priemerného rastu:

Ak rast cien bol postupne 20 %, 10 %, potom 15 % pokles a 10 % rast, tak priemerný rast sa rovná $(1{,}20 \cdot 1{,}10 \cdot 0{,}85 \cdot 1{,}10)^{1/4} \cong 1{,}054$, čiže priemerný rast je približne 5,4 %.

Toto číslo vyjadruje, že výsledná cena by bola taká istá aj v prípade, ak by rast bol konštantný, každý rok 5,4 % (lebo $1{,}054^4 \cong 1{,}2 \cdot 1{,}1 \cdot 0{,}85 \cdot 1{,}1$).

Opisné charakteristiky

- **Modus** je najčastejšie sa vyskytujúca hodnota v súbore.
- Modus identifikuje „najpopulárnejšie“ skóre .
- Nepoužíva sa však často, a to najmä pre svoju nízku stabilitu, prejavujúcu sa značnou fluktuáciou hodnôt medzi rôznymi výbermi z tej istej populácie.

Opisné charakteristiky

- **Medián** je prostredná hodnota súboru zoradeného podľa veľkosti. Ak usporiadame hodnoty súboru podľa veľkosti od najmenšej po najväčšiu, medián rozdelí súbor na dve rovnako početné časti.
- Ak je počet hodnôt súboru *n* nepárny, je tá hodnota, ktorá má v usporiadanom súbore poradové číslo *(n + 1) / 2.*
- Ak je *n* párne, potom je medián rovný aritmetickému priemeru hodnôt s poradovými číslami *(n/2)*, *(n/2)+1.*

Opisné charakteristiky

- Základná výhoda mediánu, ako štatistického ukazovateľa je tá skutočnosť, že nie je ovplyvnený extrémnymi hodnotami. Preto sa často používa aj v prípade šikmých rozdelení, pri ktorých aritmetický priemer poskytuje zvyčajne nevhodné výsledky. Napr. v súbore *{ 1, 2, 2, 3, 9 }* sa medián (tak isto ako modus) rovná dvom, čo je viditeľne vhodnejší ukazovateľ prevažujúcej tendencie, ako aritmetický priemer, ktorý tu je *3,4*.
- Ďalšia výhoda je tá, že medián sa dá definovať na každom súbore lineárne usporiadanom podmienkou „menší alebo sa rovná“, aj keď nejde o súbor čísel. Napríklad medián súboru *{absolvent ZŠ, vyučený, vyučený s maturitou, vysokoškolák}* sa rovná hodnote *„vyučený“,* ak kategórie vzdelania považujeme za usporiadané podľa náročnosti školy.
- Nevýhodné je použiť medián pri súboroch, v ktorých sledovaný znak nadobúda len jednu z dvoch možností. Tam sa medián správa rovnako ako modus: je hrubým meradlom vlastností rozdelenia a v prípade, že obidve kategórie sú zastúpené zhruba rovnako, je veľmi nestabilný.

Opisné charakteristiky

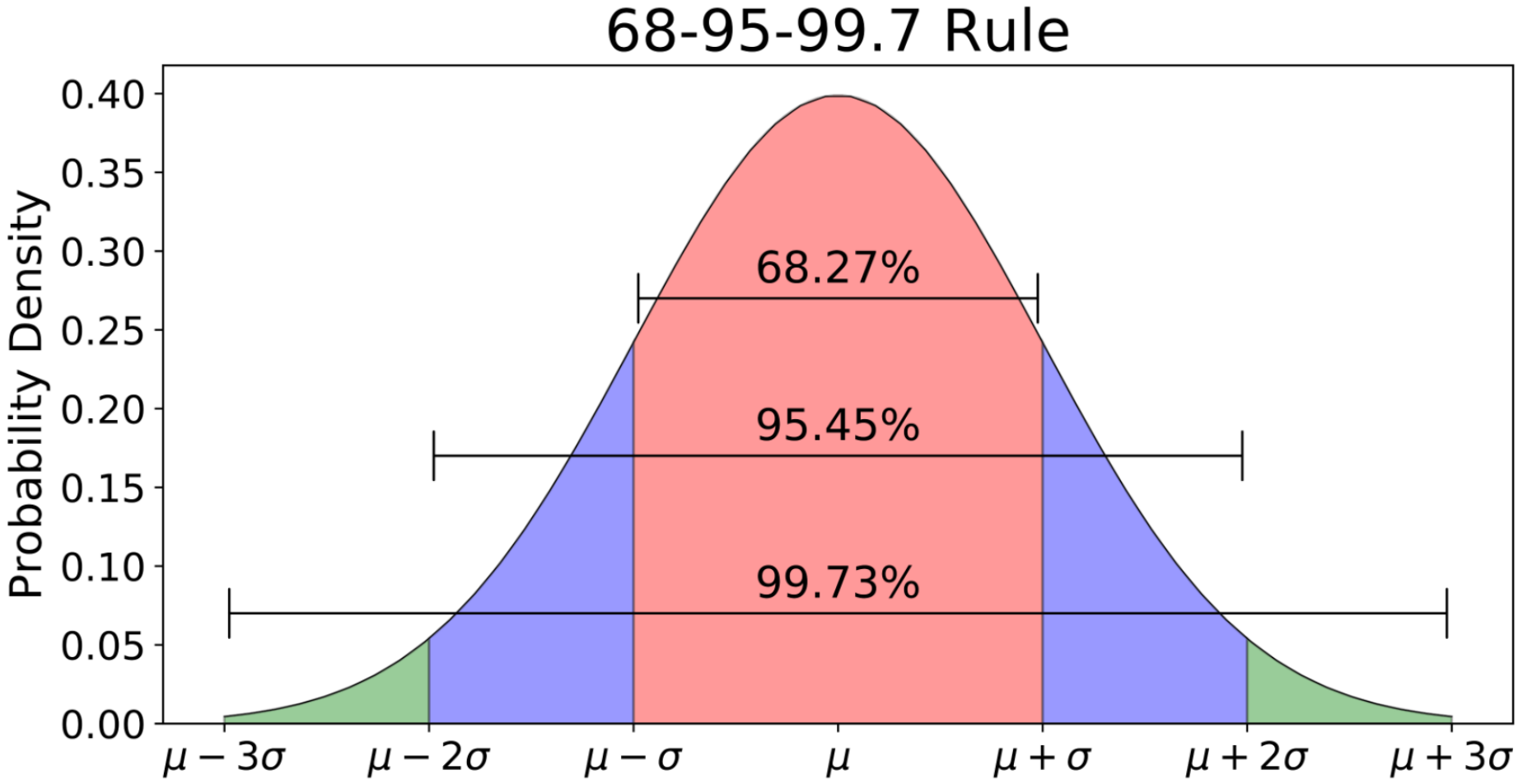
- **Smerodajná odchýlka (standard deviation)** je najdôležitejšia a najpoužívanejšia miera variability.
- Hovorí o tom, ako široko sú rozložené hodnoty v množine.
- Podobne ako aritmetický priemer sa určuje zo všetkých hodnôt súboru, čo zaručuje jej vysokú stabilitu. Označuje sa písmenom ***s*** *(SD)* a vypočíta sa podľa vzorca

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

Opisné charakteristiky

- s^2 sa označuje ako rozptyl (variance) (iné názvy: variancia, disperzia, stredná kvadratická odchýlka)

Opisné charakteristiky



Opisné charakteristiky

- Variačné rozpätie (range) $R = x_{max} - x_{min}$

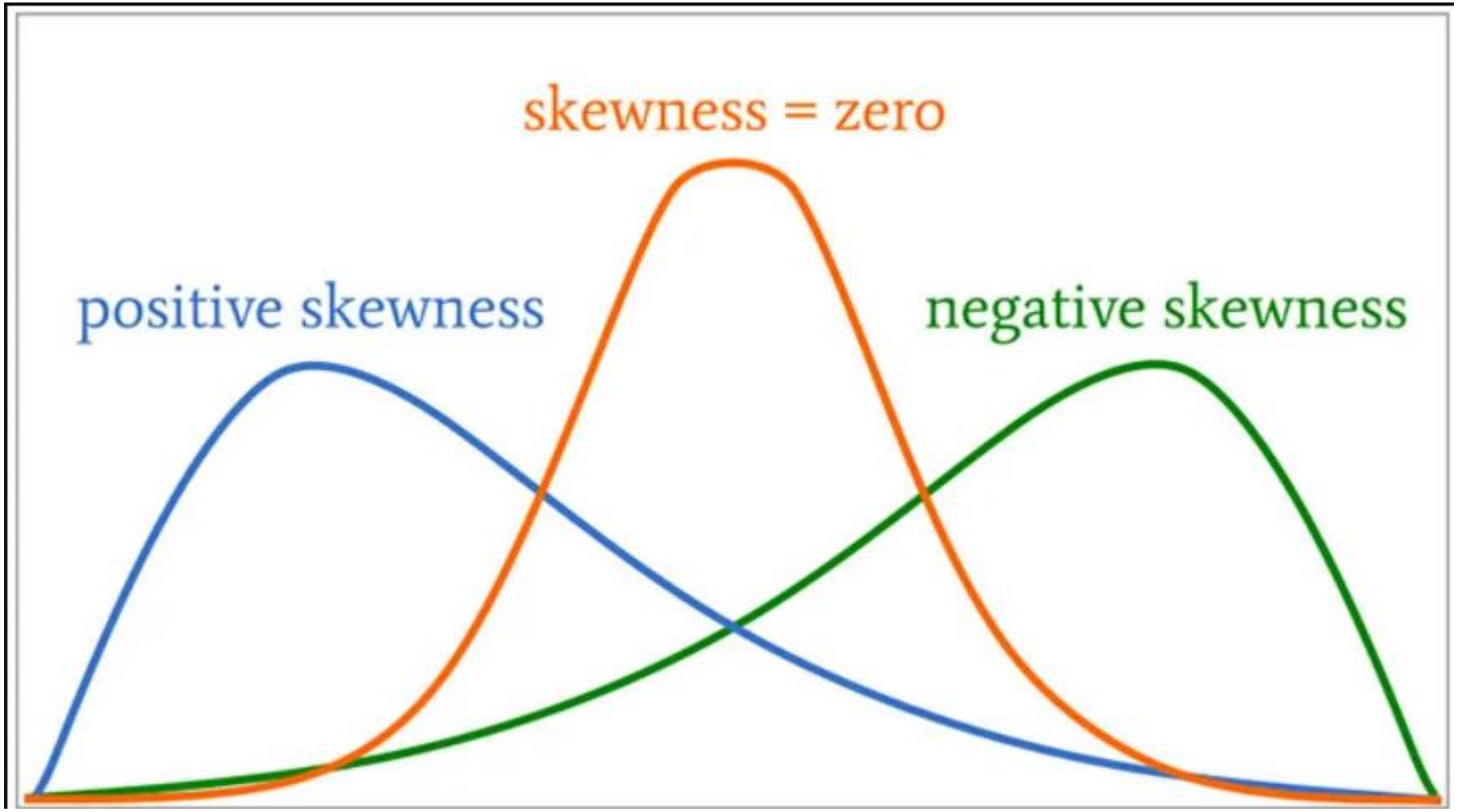
- Variačný koeficient (coefficient of variation)

$$V = \frac{s}{\bar{x}} \cdot 100\ \%$$

Opisné charakteristiky

- Šikmosť (skewness) Vyjadruje zošikmenie súboru, teda zjednodušene či prevažujú vysoké hodnoty alebo nízke hodnoty v súbore. Ak vyjde koeficient šikmosti nula, potom ide o symetrické rozdelenie. Ak je menší ako nula ide o vpravo zošikmené rozdelenie (viac väčších hodnôt a málo menších), ak je koeficient vyšší ako *0* ide o vľavo zošikmené rozdelenie (viac menších hodnôt a málo väčších).

Opisné charakteristiky

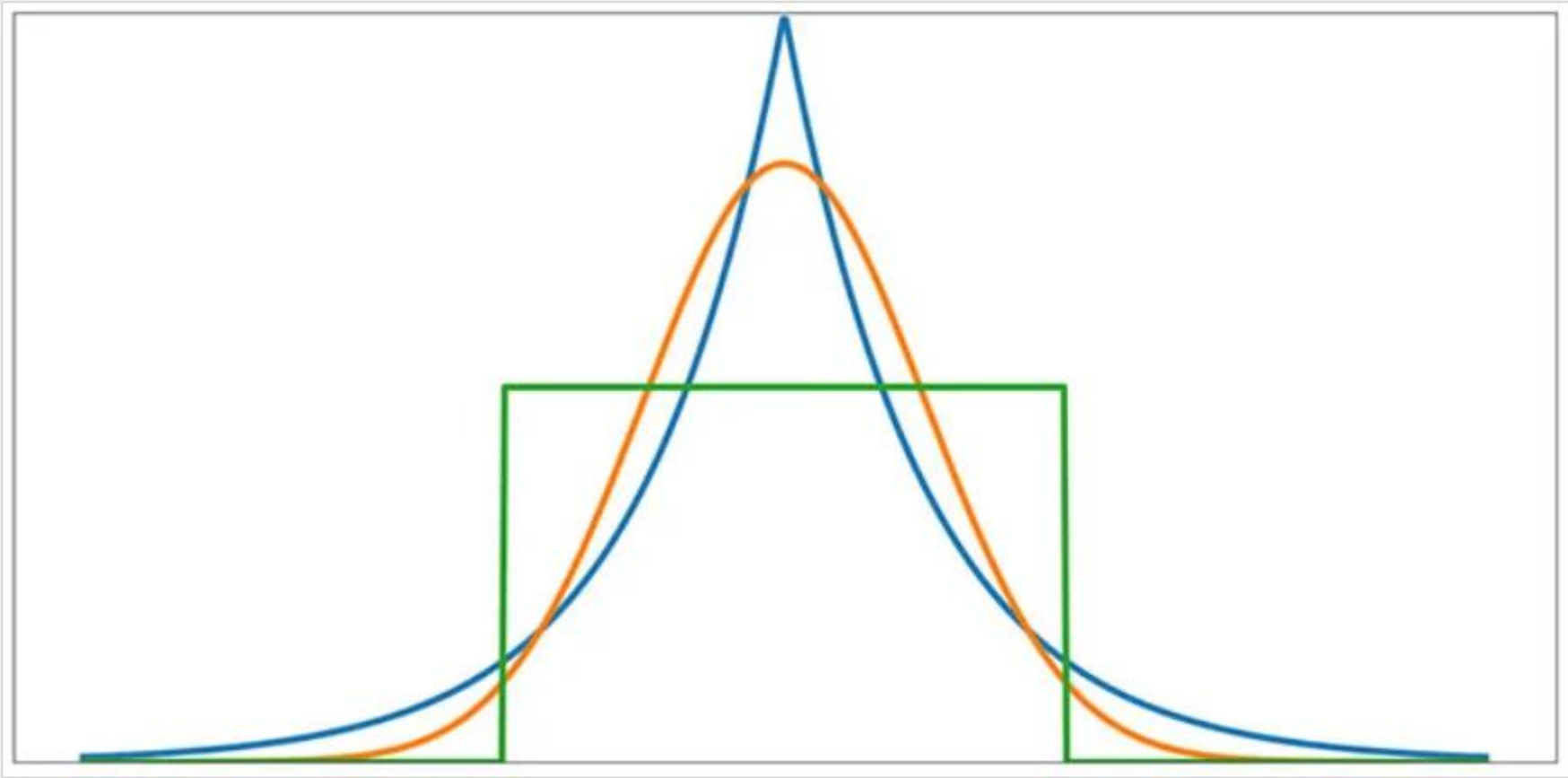


Opisné charakteristiky

- Špicatosť (kurtosis) Vyjadruje rozloženie dát v súbore. Ak je výsledok väčší ako *0* potom je to špicatejšie rozdelenie a ak je výsledok menší ako *0* potom je rozdelenie plochejšie. Špicaté rozdelenie prakticky znamená, že väčšina hodnôt v súbore je blízko priemeru. Plochejšie rozdelenie naopak znamená, že v súbore máme veľa nízkych aj vysokých hodnôt a nie sú blízko priemeru.

Hanák, R. (2016). Dátová analýza pre sociálne vedy. Bratislava : Ekonóm.

Opisné charakteristiky



Notice that kurtosis greater than or less than 3 corresponds to non-normal distribution shapes.

Kvartily

- V štatistike je kvartil typ kvantilu a sú to tri body, ktoré rozdeľujú zoradené dáta do štyroch rovnakých skupín (podľa počtu čísiel), z ktorých každá predstavuje štvrtinu vzorky dát.
- Existujú tri kvartily: prvý kvartil (Q1), druhý kvartil (Q2), a tretí kvartil (Q3).
- Prvý kvartil (dolný kvartil), sa rovná 25. percentilu dát.(oddelí najmeších 25% dát od najvyššých 75%)
- Druhý (stredný) kvartil (medián) súboru dát sa rovná 50. percentilu dát. (rozdelí zotriedené dáta na polovice)
- Tretí kvartil, tzv. horný kvartil je rovný 75. percentilu dát. (oddelí najmeších 75% dát od najvyššých 25%)

Kvartily

- Ako sa vypočíta kvartil?
- Zotriedime súbor dát s číslami s n prvkami a vyberieme n/4-tý prvok ako kvartil Q_1, n/2-tý prvok ako kvartil Q_2 a 3n/4-tý prvok zotriedeného súboru dát ako kvartil Q_3. Ak indexy n/4, n/2 alebo 3n/4 nie sú prirodzené čísla, potom použijeme priemer medzi okolitými prvkami (vtedy kvartil už nepatrí do množiny vstupných dát).
- Napríklad pre n=100 prvkov, je prvý kvartil Q_1 rovný 25-tému prvku zotriedených dát, kvartil Q_2 rovný 50-tému prvku zotriedených dát a kvartil Q_3 rovný 75-tému prvku zotriedených dát.
- Nultý kvartil Q_0 by bol minimálny prvok a štvrtý kvartil Q_4 by bol maximálny prvok dát, avšak v štatistike sa tieto krajné kvartily volajú minimum resp. maximum.

https://www.socscistatistics.com/

Induktívna štatistika – testy štatistických hypotéz

- **Test štatistickej hypotézy** je pravidlo, ktoré na základe získaných údajov dovoľuje jednoznačne rozhodnúť, či danú hypotézu prijímame, alebo zamietame.
- Test štatistickej hypotézy nazývame aj **testom štatistickej významnosti**.
- Ak H_0 na zvolenej hladine zamietneme, resp. nezamietneme, tak povieme, že **výsledok testu je štatisticky významný**, resp. **nevýznamný na danej hladine**.

Induktívna štatistika – testy štatistických hypotéz

- Štatistický test sa vykonáva tak, že z údajov získaných meraním alebo pozorovaním vypočítame **hodnotu testovej charakteristiky (kritéria)**, ktorú porovnávame s kritickou hodnotou.
- pri rozdieloch menších ako kritická hodnota H_0 nezamietame.
- pri rozdieloch väčších ako kritická hodnota H_0 zamietame.

Obojstranné resp. jednostranné testy

- H0 – nulová (základná) hypotéza (Null Hypothesis).
- H1 – alternatívna hypotéza (Alternative Hypothesis).
- **obojstranný test**:

$$H0 : Q = Q_0$$

$$H1 : Q \neq Q_0$$

- resp**. jednostranné testy**:

$$H1 : Q > Q_0 \text{ pravostranný}$$

$$H1 : Q < Q_0 \text{ ľavostranný}$$

Druhy štatistických testov významnosti

- **Neparametrický test**
  - Netýka sa len základných parametrov súboru.
  - Nemá toľko podmienok ako parametrický test, nemusí byť známy typ rozdelenia.
  - Univerzálnejšie, ale štatisticky menej účinné (menšia schopnosť rozoznať aj malé odchýlky od nulovej hypotézy).

Druhy štatistických testov významnosti

- **Parametrický test**
  - Týka sa niektorého parametru rozdelenia náhodnej veličiny.
  - Zvyčajne vyžadujú normálne rozdelenie náhodnej veličiny.
  - Sú štatisticky účinnejšie ako neparametrické testy.

Parametrické metódy (PM)

- *Párový Studentov t-test*
- *Studentov t-test pre nezávislé vzorky*
- *Z-test*
- *Analýza rozptylu (ANOVA)*
  - *Jednofatkorová ANOVA (one-way ANOVA)*
  - *Viacfaktorová ANOVA (multifactor ANOVA)*
- *Analýza roztylu pre opakované merania*
  - *Jednofatkorová ANOVA (one-way ANOVA for repeated measures)*
  - *Viacfaktorová ANOVA (multifactor ANOVA for repeated measures)*
- *Pearsonov korelačný koeficient*

Neparametrické metódy (NPM)

- *χ^2 - test dobrej zhody (Chi-square test)*
- *Znamienkový test (Sign test)*
- *Wilcoxonov test (Wilcoxon sign-rank test)*
- *Mann – Whitneyov U test (Mann – Whitney U test)*
- *Friedmanov test (The Friedman test for repeated measures)*
- *Kruskal – Wallisov test (Kruskal – Wallis test for independent measures)*
- *Spearmanov korelačný koeficient*

Neparametrické metódy (NPM)

Klasifikácia vybraných neparametrických testov:

-- jednovýberové testy
- test extrémnych hodnôt Dixonov test
- Wilcoxonov test
- Znamienkový test

– dvojvýberové testy pre nezávislé výbery
- Wilcoxonov test (Mann – Whitneyov U-test)

– neparametrická analýza rozptylu
- Kruskal – Walisov test
- Friedmanov test

– testy náhodnosti
- test založený na bodoch zvratu

– testy nezávislosti
- Hoeffdingov test
- Kendallov koeficient a test nezávislosti
- Spearmanov korelačný koeficient

– a iné....

Nonparametric test	Parametric Alternative
1-sample sign test	One-sample Z-test, One sample t-test
1-sample Wilcoxon Signed Rank test	One sample Z-test, One sample t-test
Friedman test	Two-way ANOVA
Kruskal-Wallis test	One-way ANOVA
Mann-Whitney test	Independent samples t-test
Mood's Median test	One-way ANOVA
Spearman Rank Correlation	Correlation Coefficient

Chi-kvadrát test

- **χ^2 - štatistika má široké použitie:**
  - používa sa na overenie predpokladu o závislosti (asociácii) dvoch kvalitatívnych znakov
  - používa sa aj na overenie predpokladu o rozdelení skúmaného znaku v ZS; či sú dáta rozdelené podľa nejakého známeho teoretického rozdelenia, napr. normálneho, binomického, Poissonovho a pod. **(test dobrej zhody)**
- **Všeobecný postup:**
  - Formulovanie hypotézy
  - Výpočet teoretických početností v triednych intervaloch
  - Zistenie skutočných početností v triednych intervaloch
  - Výpočet testovacej **χ^2-** štatistiky
  - Nájdenie kritickej hodnoty a rozhodnutie o výsledku

χ^2 - test dobrej zhody

Hodnota	1	2	3	4	5	6	$\sum_{i=1}^{k}$
Skutečné četnosti	5	14	4	10	14	13	60
Očekávané četnosti	10	10	10	10	10	10	60
χ^2	2.5	1.6	3.6	0	1.6	0.9	10.2

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - Np_i)^2}{Np_i}$$

	p					
ν	**0.100**	**0.050**	**0.025**	**0.010**	**0.005**	**0.001**
1	2.7055	3.8415	5.0239	6.6349	7.8794	10.8276
2	4.6052	5.9915	7.3778	9.2103	10.5966	13.8155
3	6.2514	7.8147	9.3484	11.3449	12.8382	16.2662
4	7.7794	9.4877	11.1433	13.2767	14.8603	18.4668
5	9.2364	11.0705	12.8325	15.0863	16.7496	20.5150

Hodnota kritéria na hladine významnosti 0,05: 11,07.

H_0 nezamietame

Znamienkový test

- Test na medián (X_{Me})

  $P(X < X_{Me}) = P(X > X_{Me}) = 0,5$

- **Formulovanie hypotézy:**

  $H_0: X_{Me} = X_0$

  $H_1: X_{Me} \neq X_0$

- nech NP **Y** = počet kladných rozdielov predpokladanej mediánovej hodnoty (X_0) a nameraných hodnôt x_i

- **Y má Bi(0,5;n) a platí:**

  $P(Y < k_1) = P(Y > k_2) = \alpha/2$

Znamienkový test

i	1	2	3	4	5	6	7	8	9
x_i	10,4	9,2	7,1	10,8	7,2	4,9	8,2	4,2	10,1
y_i	6,3	5,2	5,8	1,5	6,8	7,2	5,3	5,4	7,9
z_i	4,1	4,0	1,3	9,3	0,4	-2,3	2,9	-1,2	2,2

9	0	.0020	9		2	.0037	13		5	.0318	14
	1	.0195	8		3	.0176	12		6	.0835	13
	2	.0898	7		4	.0592	11		7	.1796	12
	3	.2539	6		5	.1509	10		8	.3238	11
	4	.5000	5		6	.3036	9		9	.5000	10
10	0	.0010	10		7	.5000	8	20	0	.0000	20
	1	.0107	9	16	0	.0000	16		1	.0000	19
	2	.0547	8		1	.0003	15		2	.0002	18
	3	.1719	7		2	.0021	14		3	.0013	17
	4	.3770	6		3	.0106	13		4	.0059	16
	5	.6230	5		4	.0384	12		5	.0207	15

k = 6 $k_1 = 1$ $k_2 = 8$

H_0 nezamietame

Wilcoxonov test

- Ako znamienkový, podmienka, že dáta sú aspoň poradové
- Je účinnejší ako znamienkový, keďže sa do úvahy berie aj poradie dát

Wilcoxonov test

i	1	2	3	4	5	6	7	8	9
x_i	10,4	9,2	7,1	10,8	7,2	4,9	8,2	4,2	10,1
y_i	6,3	5,2	5,8	1,5	6,8	7,2	5,3	5,4	7,9
d_i	4,1	4,0	1,3	9,3	0,4	-2,3	2,9	-1,2	2,2
R_i	8	7	3	9	1	5	6	2	4

$$T_+ = 8 + 7 + 3 + 9 + 1 + 6 + 4 = 38; \quad T_- = 5 + 2 = 7.$$

$T = \min(T_+, T_-) = 7$

	alpha values						
n	0.001	0.005	0.01	0.025	0.05	0.10	0.20
5	--	--	--	--	--	0	2
6	--	--	--	--	0	2	3
7	--	--	--	0	2	3	5
8	--	--	0	2	3	5	8
9	--	0	1	3	5	8	10
10	--	1	3	5	8	10	14

H_0 nezamietame

Z-test vs T-test

- Dôležitým rozdielom medzi t-testom a z-testom je veľkosť vzorky pre každý typ testu.
- T-test sa primárne používa na výskum s obmedzenou veľkosťou vzorky, zatiaľ čo z-test sa používa na testovanie hypotéz, ktoré vyžadujú, aby sa výskumníci pozreli na veľkosť populácie, ktorá je väčšia ako 30.