



ŠTATISTIKA

- náuka o získavaní, spracovaní a interpretácii informácií v nameraných dátach, prieskumoch...

Význam pojmu štatistika:

1. Vedná disciplína, ktorá sa zaoberá zberom, spracovaním a vyhodnocovaním štatistických údajov
2. Číselné i nečíselné údaje alebo súhrn údajov o hromadných javoch
3. Praktická činnosť, ktorá vedie k získaniu informácií – údajov o hromadných javoch

Základné delenie štatistiky:

Popisná (deskriptívna)

- Spracovanie hromadných empirických dát
- Niekoľkými číslami, tabuľkami, grafmi stručne vystihnúť dôležité
- Závery iba o daných dátach (nie je možné zovšeobecňovať)

Induktívna (matematická)

Na základe dát pomocou matematických metód (teória pravdepodobnosti) usudzujeme o vlastnostiach populácie (závery zovšeobecňujeme):

- Teória odhadu populačných parametrov (bodové a intervalové odhady)
- Testovanie štatistických hypotéz
- Štatistická predikcia – odhady budúceho vývoja sledovanej veličiny

Štatistika dnes – nástroj na zvyšovanie kvality rozhodovania



- Je nový spôsob výroby lepší ako predchádzajúci?
- Neokráda firma zákazníka – spĺňajú jej produkty deklarovanú kvalitu?
- Ktorá liečba je pre danú chorobu efektívnejšia?
- Je úroveň vzdelania na vybraných fakultách rovnaká?

Kde? Kedy? Prečo sa používa štatistika?

Skúmame zložitý systém, ktorý

- nie je možné jednoducho pochopiť (potrebujeme empirické skúsenosti)
- za rovnakých alebo podobných podmienok sa môže prejavovať odlišným spôsobom
- chceme odhaliť súvislosti, zákonitosti, závislosti, systematické chyby

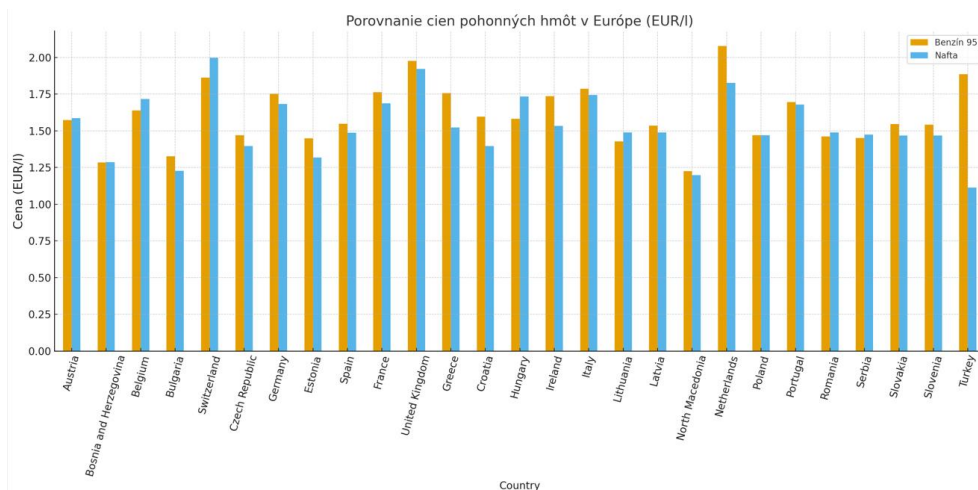
Oblasti aplikácie: prírodné vedy (chémia, fyzika, medicína, farmakológia, genetika...), ekonómia (makroekonomika, poisťovníctvo, bankovníctvo...), technické vedy (doprava, strojníctvo, priemyselná výroba - kontrola kvality, riadenie ...), spoločenské vedy...

Štatistika v reálnom živote:

EUROSTAT, FinStat, Štatistický úrad SR, OECD, WHO (epidemiológia, prenatálna diagnostika, fajčenie...), NCZI, Ústredný portál verejnej správy...
volebné prieskumy, správy v médiách (účelové zneužitie štatistiky...),

The screenshot shows the Eurostat website interface. At the top, there is a navigation bar with 'Home', 'Data', 'News', 'Publications', 'About us', 'Contact us', and 'Help'. Below this, a 'Statistical themes' section is visible, featuring a 'Main categories' sidebar with links like 'Special interest topics', 'EU policy indicators', 'General and regional statistics', 'Economy and finance', 'Population and social conditions', 'Industry, trade, and services', 'Agriculture, fisheries, and forestry', 'International trade', 'Transport', 'Environment and energy', and 'Science, technology, and digital society'. The main content area displays 'Special interest topics' with a link to 'Ukraine' and 'EU policy indicators' with links to 'Euro indicators', 'Sustainable development', 'European pillar of social rights', 'Circular economy', and 'Macroeconomic imbalance procedure (MIP)'. At the bottom, 'General and regional statistics' is listed with a link to 'Regions and cities'.

Porovnanie aktuálnych cien pohonných hmôt v Európe k 3.10.2025



Spracované podľa AI

Štatistika v živote študenta MTF:

DP, odborné články, pojmy p-hodnota, štatistická významnosť, interval spoľahlivosti...

Rebríčky

Firmy s najväčšími tržbami
Najziskovejšie firmy
Najväčší platitelia dane
Najväčší zamestnávateľia
Najlepšie hospodáriace obce
Najväčšie politické strany
Najväčšie univerzity

Online Štatistiky

Štatistika finančných výsledkov firiem 2022
Štatistika počtu vzniknutých a zaniknutých firiem 2023
Najväčší platcovia dane z príjmu 2022

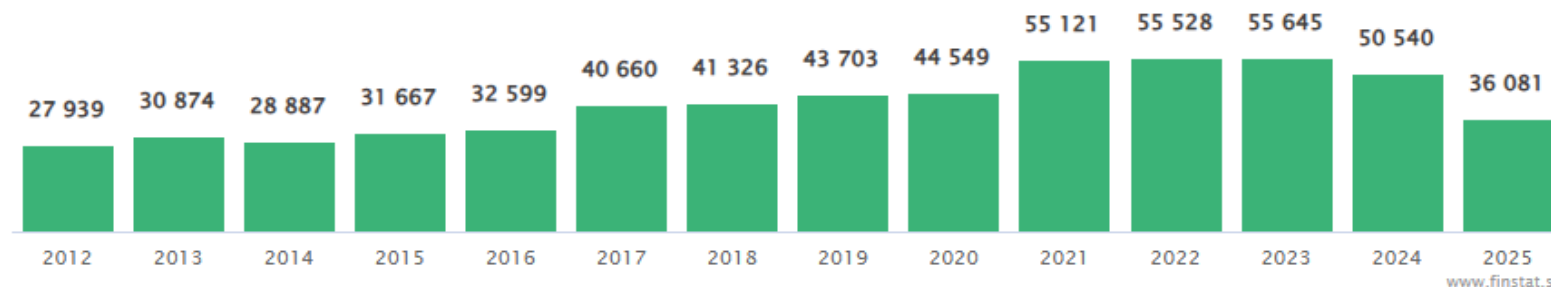
[Všetky Online štatistiky »](#)

Analýzy

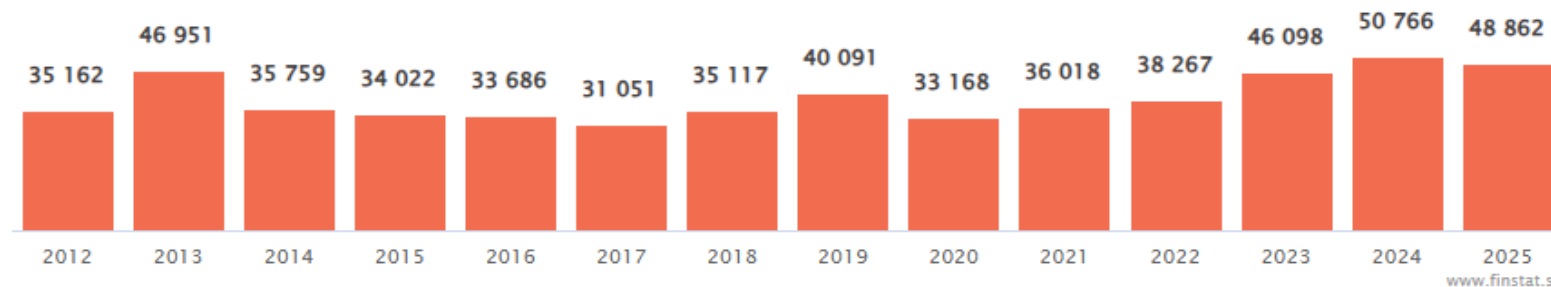
Ako hospodária politické strany?
Analýza hospodárenia maloobchodných reťazcov za rok ...
Analýza hospodárenia nefinančných podnikov za rok 2022
Analýza vzniku a zániku firiem v roku 2022
FinStat skóre – nový model hodnotenia rizika úpadku

[Všetky Analýzy »](#)

Počet vzniknutých živností po rokoch



Počet zaniknutých živností po rokoch





STATISTICAL
OFFICE
OF THE SLOVAK
REPUBLIC

THE ELECTIONS 2024

INFLATION 2024

Key indicators

Validation of inflation



Statistics

Databases

Products

Services

Metadata

Surveys

About us

 Search



SLOVENSKÁ
ŠTATISTIKA
a DEMOGRAFIA

SLOVAK STATISTICS
and DEMOGRAPHY

vedecký časopis/scientific journal

3/2025
ročník 35

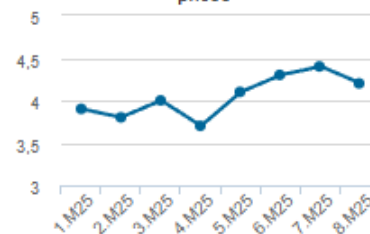
Slovak Statistics and Demography
3/2025

Current issue of journal Slovak Statistics and
Demography 3/2025. [More](#)

1 2 3 4 5

 Inflation rate

Year-on-year change,
calculated from consumer
prices



 Detail

 GDP growth rate

 Unemployment

 Average monthly wage

 Key indicators

 Gross domestic product 0,6 %
2 Q 2025 / 2 Q 2024

 News

Foreign trade in August 2025



09.10.2025 | Statistical Office of the SR | Foreign trade | Information Reports

Both imports and exports decreased in August, with a trade balance remaining in surplus for the fourth consecutive month.

 Read more

07.10.2025

News: The new edition of the Statistical Yearbook of Bratislava 2025 brings the latest data on life in the capital

<https://slovak.statistics.sk/wps/portal/ext/home/>

Tržby 2024 (EUR)

	Firma	Vznik	Tržby 2024 (EUR)
1.	Alza.sk s. r. o.	2004	539 442 883
2.	GymBeam s.r.o.	2011	165 765 253
3.	Allegro Slovakia s. r. o.	2005	43 395 154
4.	GAME PLATFORMA s. r. o.	2011	27 546 838
5.	MTBIKER community s. r. o.	2019	20 301 492
6.	ARAŠID spol. s r.o.	2002	20 121 428
7.	MTBIKER s. r. o.	2013	18 423 569
8.	eyerim s.r.o.	2014	17 423 554
9.	Wilsondo s. r. o.	2013	16 208 022
10.	4Home CEE, s. r. o.	2022	15 908 726

EBITDA 2024 (EUR)

	Firma	IČO	Vznik	EBITDA 2024 (EUR)
1.	Alza.sk s. r. o.	36562939	2004	10 072 976
2.	GymBeam s.r.o.	46440224	2011	8 792 236
3.	Bloom Cosmetics s.r.o.	51228637	2017	1 937 593
4.	Ksisters s. r. o.	53143230	2020	1 785 069
5.	J&J AUTOMOTIVE, s.r.o.	36612669	2006	1 401 288
6.	THE HOLSTERS COMPANY s.r.o.	44638205	1996	1 382 993
7.	Wilsondo s. r. o.	47429151	2013	1 115 871
8.	Kvety.sk s.r.o.	36641715	2005	1 010 883
9.	Bestblades s. r. o.	53611713	2021	863 119
10.	Allegro Slovakia s. r. o.	35950226	2005	758 200

Rast tržieb e-shopov 2021-2024

	Firma	Vznik	Tržby 2021 (EUR)	Tržby 2024 (EUR)	CAGR (%)
1.	Ksisters s. r. o.	2020	353 701	11 401 551	218
2.	SHAPEN s. r. o.	2018	385 246	3 382 368	106
3.	EXACT Invest s. r. o.	2017	980 641	8 537 890	106
4.	Alza.sk s. r. o.	2004	78 552 053	539 442 883	90
5.	Mia Dresses, s. r. o.	2015	365 765	2 382 482	87
6.	SportZoo s. r. o.	2013	752 249	3 130 441	61
7.	Garó & Co., s.r.o.	2014	1 278 508	5 082 411	58
8.	Bloom Cosmetics s.r.o.	2017	2 757 769	9 802 387	53
9.	GymBeam s.r.o.	2011	52 904 687	165 765 253	46
10.	DREVKO s. r. o.	2019	894 158	2 682 743	44

Svetová populácia

8 251 420 902	Súčasná svetová populácia
102 422 373	Narodenia tento rok
139 439	Narodenia dnes
48 261 902	Úmrtia tento rok
65 704	Úmrtia dnes
54 160 471	Čistý prírastok populácie tento rok

Vláda a Ekonomika

\$ 6 875 338 241	Celosvetové výdaje vlád na zdravotníctvo dnes
\$ 4 525 188 311	Celosvetové výdaje vlád na vzdelanie dnes
\$ 1 848 218 905	Celosvetové výdaje vlád na ozbrojené sily dnes
52 399 436	Vyrobené autá tento rok
108 021 388	Vyrobené bicykle tento rok
272 747 845	Predané počítače tento rok

https://www.worldometers.info/world-population

Vybrané citáty

„Štatistika je veda, ktorá vyrába nespoľahlivé fakty zo spoľahlivých údajov.“
Evan Esar

„Verím iba tým štatistikám, ktoré som sám sfalšoval.“

Winston Churchill

„Poznám tri druhy klamstva: Malé klamstvo, veľké klamstvo a štatistika.“
George Bernard Shaw

„Sedíš li jednou půlkou v ledu a druhou na rozpálených kamnech, je ti statisticky velmi příjemně.“

Murphyho zákon

„Smrt' jedného muža je tragédia, smrť miliónov je štatistika. “

J. Stalin

Základný štatistický súbor ZS (populácia) je súbor všetkých štatistických jednotiek.

všetci študenti MTF STU

Výberový štatistický súbor VS(vzorka) je množina vybraných štatistických jednotiek.

vybraná skupina študentov MTF STU

Štatistická jednotka je základný prvok, na ktorom skúmame konkrétny prejav určitého hromadného javu.

konkrétny študent MTF STU

Štatistický znak je vonkajší, postihnuteľný, merateľný prejav vlastností štatistickej jednotky.

počet bodov z testu MMPVE

výška

hmotnosť

počet súrodencov

vzdialenosť bydliska od školy

značka mobilného telefónu

farba očí

krvná skupina

štatistický znak



kvantitatívny

(číselné vyjadrenie)

diskrétny - nadobúda len niektoré číselné hodnoty, najčastejšie prirodzené čísla alebo celé nezáporné čísla (počet detí, počet členov domácnosti, počet porúch v sérii výrobkov...)

↓
jednoduché triedenie

↓
intervalové triedenie

spojitý - môže nadobúdať v rámci určitého intervalu ľubovoľné hodnoty (hrubá mzda, spotreba materiálu, zisk, pridaná hodnota...)

↓
intervalové triedenie

kvalitatívny

(slovné vyjadrenie)

alternatívny - môže nadobúdať len dve obmeny (pohlavie)

↓
asociačná tabuľka

množný - nadobúda viac ako dve obmeny (farba očí, krvná skupina)

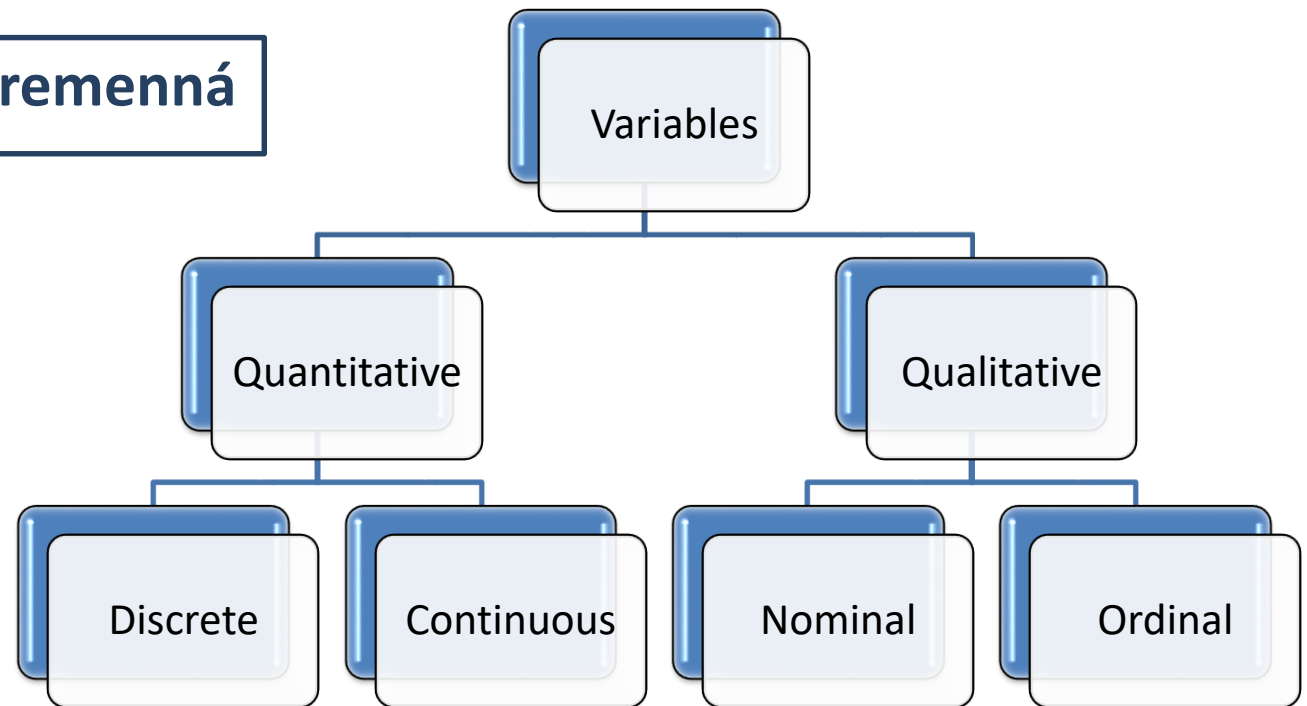
↓
kontingenčná tabuľka

štatistický znak



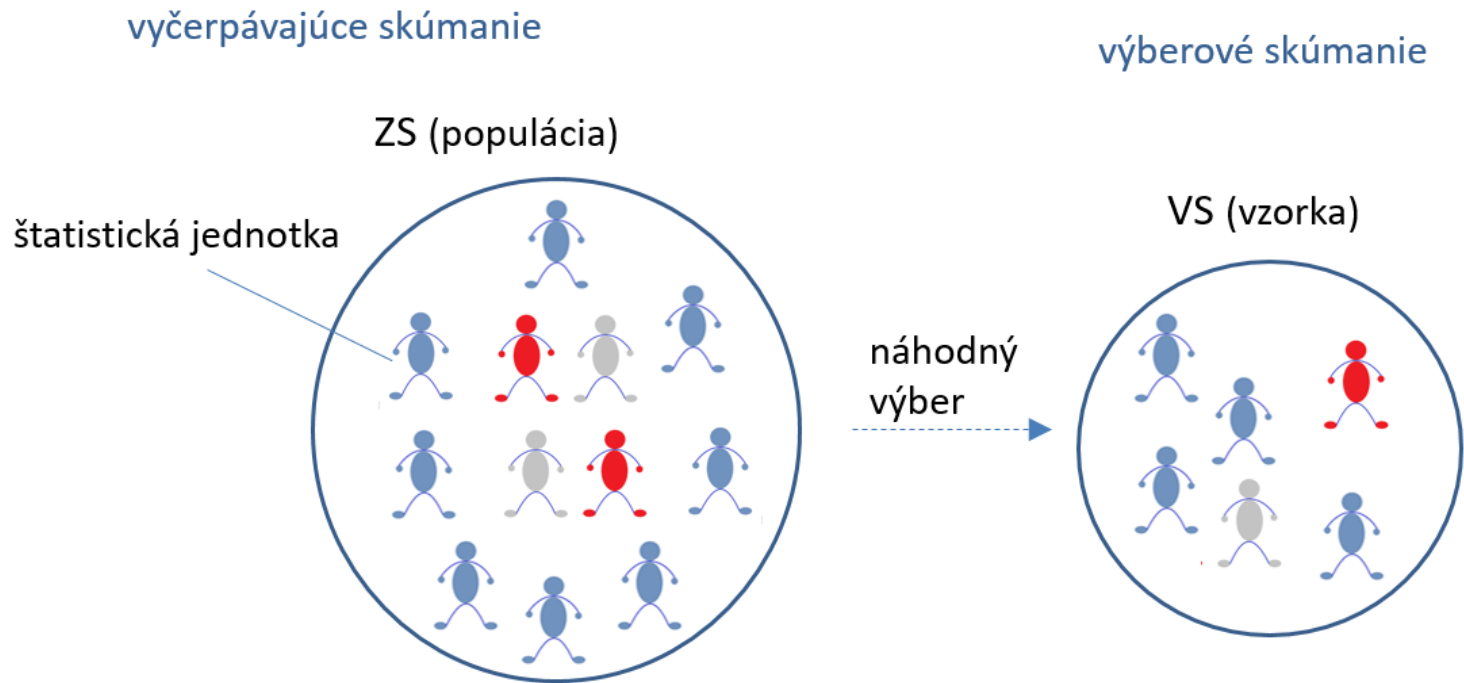
softvérové značenie „**variable**“

štatistická premenná



Druhy štatistického skúmania :

- **vyčerpávajúce** skúmanie - skúma sa celý základný súbor (každá štatistická jednotka),
- **výberové** skúmanie - skúma sa iba výberový súbor (vybrané štatistické jednotky).



Pri štatistickom skúmaní nás nezaujímajú konkrétne hodnoty sledovaného znaku na každej štatistickej jednotke, ale iba typické vlastnosti, ktoré nám poskytujú základné informácie o súbore.

VS podľa rozsahu:

- veľmi malé (do cca 20)
- malé (cca 30 až 50)
- veľké (rádovo stovky)
- veľmi veľké (rádovo tisíce a viac)

Etapy štatistického skúmania:

- štatistické **získovanie** (meranie),
- štatistické **spracovanie** (usporiadanie, triedenie, výpočet číselných charakteristík)
- štatistická **analýza**.

Náhodný výber

1. pravdepodobnosť zaradenia do VS pre všetky štatistické jednotky je rovnaká
2. štatistické jednotky sú vybrané do VS od seba nezávisle

náhodnosť je dôležitejšia ako veľkosť vzorky

Pred 300 rokmi britský predavač John Graunt – z malej vzorky extrapoláciou získal užitočné informácie o celej populácii (počet ľudí žijúcich v Londýne počas morovej epidémie)

Presnosť vzorkovania sa zvyšuje vďaka náhodnému výberu, nie iba rozsahom vzorky (náhodne zvolený VS 1100 respondentov na otázku s dvoma možnosťami odpovede áno-nie dobre reprezentuje celú populáciu)
→ od určitej fázy sa pri rastúcom počte pozorovaní neustále znižuje prírastok nových informácií, ktoré každé nové pozorovanie poskytuje.

TRIEDENIE

Pri spracovaní nameraných hodnôt štatistického znaku (VS) triedením určujeme **rad rozdelenia početností** výskytu hodnôt skúmanej premennej, ktorý:

- je empirickým ekvivalentom teoretického zákona rozdelenia pravdepodobnosti.
- špecifikuje ako často, s akou frekvenciou sa vyskytujú jednotlivé hodnoty štatistických znakov, resp. triedy hodnôt, intervaly hodnôt.

V Exceli triedime pomocou

- =FREQUENCY(údajové_pole;binárne_pole),
údajové_pole - všetky namerané hodnoty štatistického znaku;
binárne_pole - jednotlivé hodnoty štatistického znaku reprezentujúce danú triedu alebo hornú hranicu intervalu.
- analytické nástroje s ponukou **Histogram**
Údaje→Analýza dát→Histogram (Data→Data analysis→Histogram)...

TRIEDENIE

Zásady triedenia:

- **jednoznačnosť** – každá štatistická jednotka môže byť zapísaná iba do jednej triedy (skupiny)
- **úplnosť** – pre každú štatistickú jednotku existuje trieda, do ktorej ju možno zaradiť

Podľa typu štatistického znaku (typu náhodnej premennej), resp. počtu hodnôt, ktoré náhodná premenná v štatistickom súbore nadobúda používame:

jednoduché triedenie

intervalové triedenie



menej ako 15

Jednoduché triedenie

Triedenie nameraných hodnôt štatistického znaku v Exceli:

až po označení výstupnej oblasti voláme funkciu FREQUENCY

	A	B	C	D	E	F	G	H
1	5		Příklad.	Počet výrobkov, vyrobených jednotlivými pracovníkmi				
2	1							
3	5		xi	ni	Ni	fi	Fi	
4	2		1					
5	4		2					
6	5		4					
7	4		5					
8	2		9					
9	5							
10	1							
11	4							
12	4							
13	5							
14	5							
15	2							
16	5							
17	4							
18	9							
19	4							
20	5							

ako prvé označíme bunky,
do ktorých chceme
zapísať početnosti
jednotlivých tried

xi	ni	Ni	fi	Fi
=FREQUENCY(				
FREQUENCY(údajové_pole; binárne_pole)				
2				
4				
5				
9				

xi	ni	Ni	fi	Fi
=FREQUENCY(A1:A20;C4:C8)				
2				
4				
5				
9				

príkaz potvrdíme
trojkombináciou kláves
CTRL +SHIFT+ENTER

xi	ni	Ni	fi	Fi
1	2			
2	3			
4	6			
5	8			
9	1			

výstupom sú vypočítané
triedne početnosti

Intervalové triedenie

Ak náhodná premenná nadobúda **veľký počet rôznych hodnôt**, volíme triedenie do navzájom disjunktných intervalov.

Najbežnejší spôsob značenia je polouzavretý intervalu (DH, HH) , kde DH je dolná hranica, HH je horná hranica. Intervaly sú zvyčajne ekvidištantné (rovnakej veľkosti).

Existuje niekoľko spôsobov na stanovenie počtu k intervalov:

- $k \approx \sqrt{n}$,
- $k \leq 5 \log n$,
- Sturgessovo pravidlo $k \approx 1 + 3,322 \log n$,
- Yuleho vzorec $k \approx 2,5 \sqrt[4]{n}$,

alebo šírky intervalov h :

Scottovo pravidlo $h = \frac{3,5\hat{\sigma}}{\sqrt[3]{n}}$,

Freedman-Diaconisovo pravidlo $h = \frac{2(Q_{75}-Q_{25})}{\sqrt[3]{n}}, \dots$.

Platí pritom $R = kh$.

Prioritou je vždy vychádzať z prirodzeného, objektívneho skutočného predurčeného počtu tried!

Intervalové triedenie

Triedenie nameraných hodnôt štatistického znaku v Exceli:

=FREQUENCY (údajové_pole; binárne_pole)

všetky hodnoty
štatistického znaku

horné hranice
intervalov

	A	B	C	D	E	F
1	97		Příklad.	Dĺžka vybraných výrobkov v cm.		
2	103					
3	104		DH	HH	ni	
4	106		90	100		
5	106		100	110		
6	111		110	120		
7	111		120	130		
8	112		130	140		
9	112		140	150		
10	112		150	160		
11	113					
12	113					

	A	B	C	D	E	F	G
1	97		Příklad.	Dĺžka vybraných výrobkov v cm.			
2	103						
3	104		DH	HH	ni		
4	106		90	100	=FREQUENCY(A1:A40;D4:D10)		
5	106		100	110			
6	111		110	120			
7	111		120	130			
8	112		130	140			
9	112		140	150			
10	112		150	160			
11	113						
12	113						

	A	B	C	D	E	F
1	97		Příklad.	Dĺžka vybraných výrobkov v cm.		
2	103					
3	104		DH	HH	ni	
4	106		90	100	1	
5	106		100	110	4	
6	111		110	120	17	
7	111		120	130	12	
8	112		130	140	5	
9	112		140	150	0	
10	112		150	160	1	
11	113				0	

potvrdíme
CTRL+ SHIFT+ENTER

výstupom sú početnosti výskytu hodnôt štatistického znaku vo zvolených intervaloch

Rad rozdelenia početností, reprezentovaný frekvenčnou tabuľkou

	A	B	C	D	E	F	G	H
1	97		Příklad.	Dĺžka vybraných výrobkov v cm.				
2	103							
3	104		DH	HH	ni	Ni	fi	Fi
4	106		90	100	=FREQUENCY(A1:A40;D4:D10)	=E4	=E4/\$E\$11	=G4
5	106		100	110	=FREQUENCY(A1:A40;D4:D10)	=F4+E5	=E5/\$E\$11	=H4+G5
6	111		110	120	=FREQUENCY(A1:A40;D4:D10)	=F5+E6	=E6/\$E\$11	=H5+G6
7	111		120	130	=FREQUENCY(A1:A40;D4:D10)	=F6+E7	=E7/\$E\$11	=H6+G7
8	112		130	140	=FREQUENCY(A1:A40;D4:D10)	=F7+E8	=E8/\$E\$11	=H7+G8
9	112		140	150	=FREQUENCY(A1:A40;D4:D10)	=F8+E9	=E9/\$E\$11	=H8+G9
10	112		150	160	=FREQUENCY(A1:A40;D4:D10)	=F9+E10	=E10/\$E\$11	=H9+G10
11	113		Σ:		=SUM(E4:E10)		=SUM(G4:G10)	

DH	HH	ni	Ni	fi	Fi
90	100	1	1	0,025	0,025
100	110	4	5	0,1	0,125
110	120	17	22	0,425	0,55
120	130	12	34	0,3	0,85
130	140	5	39	0,125	0,975
140	150	0	39	0	0,975
150	160	1	40	0,025	1
Σ:		40		1	

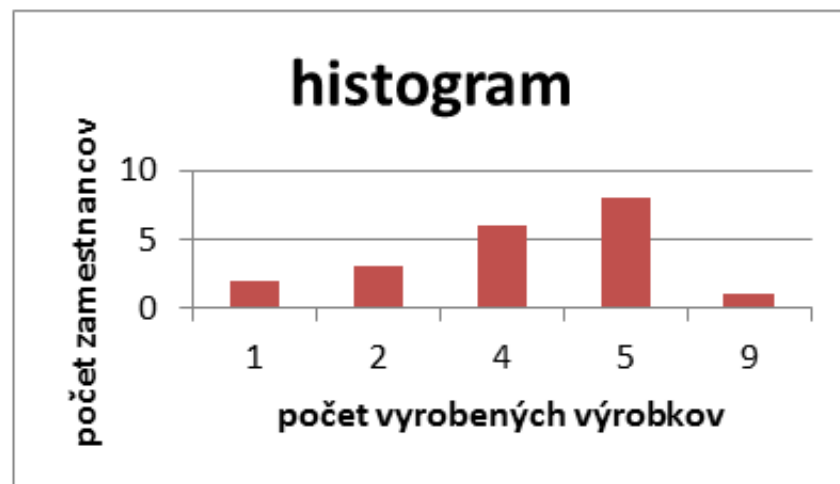
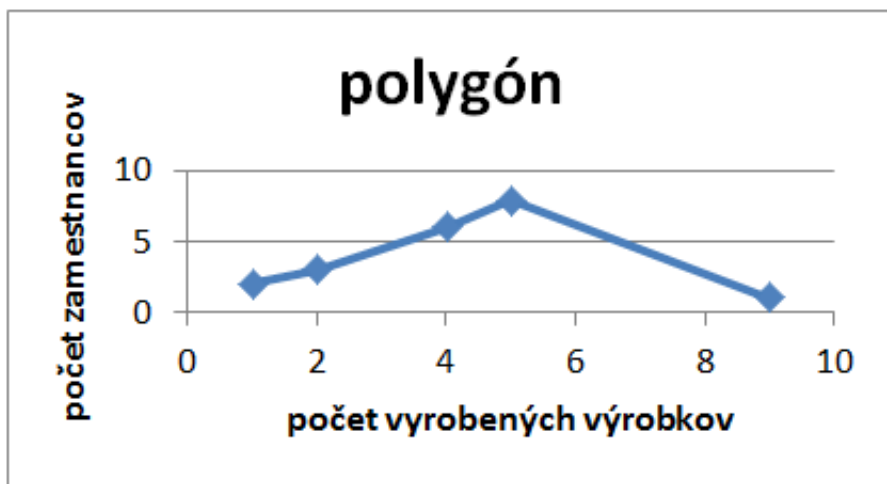
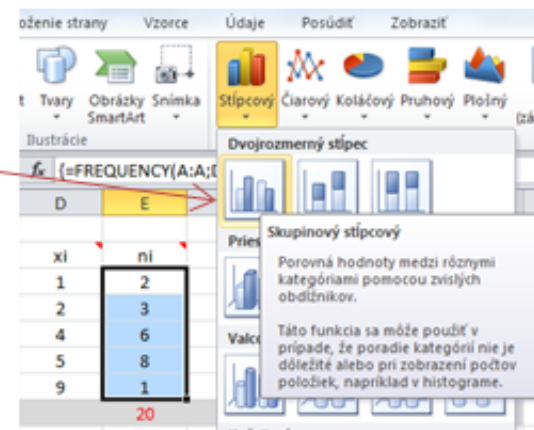
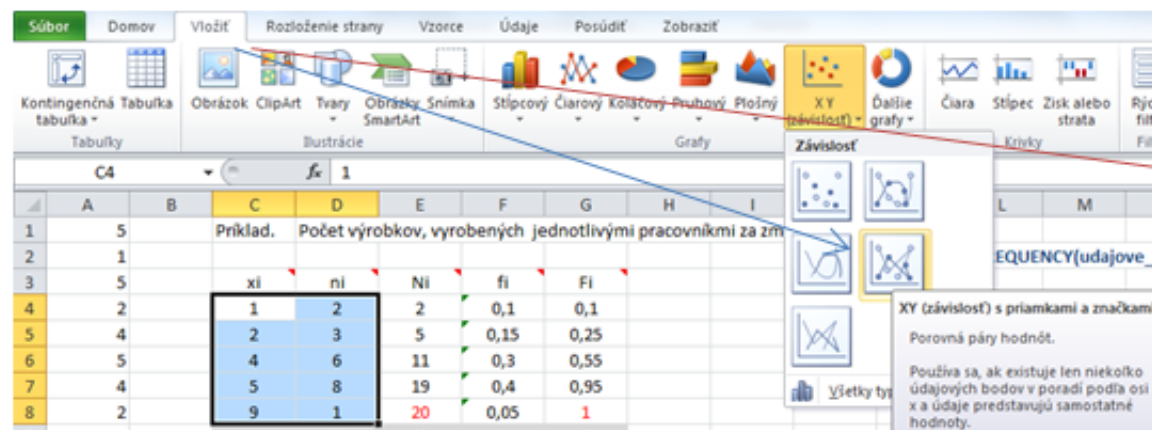
Príklad

Počty výrobkov vyrobených jednotlivými vybranými pracovníkmi za zmenu sú: 5, 1, 5, 2, 4, 5, 4, 2, 5, 1, 4, 4, 5, 5, 2, 5, 4, 9, 4, 5. Zostrojte frekvenčnú tabuľku a zobrazte graficky rozdelenia početností počtu vyrobených výrobkov.

Frekvenčná tabuľka

hodnoty štatistického znaku v i -tych triedach	absolútne početnosti	kumulatívne absolútne početnosti	relatívne početnosti	kumulatívne relatívne početnosti
x_i	n_i	N_i	f_i	F_i
1	2	2	0,1	0,1
2	3	5	0,15	0,25
4	6	11	0,3	0,55
5	8	19	0,4	0,95
9	1	20	0,05	1
súčtový riadok → $\Sigma:$	20	—	1	—

Grafické znázornenie v Exceli (histogram, polygón, ogiva, box plot):



Kontingenčná tabuľka (cross table, pivot table)

Vďaka kontingenčným tabuľkám vieme rýchlo organizovať a sumarizovať údaje rozsiahlych tabuliek a databáz.

Nástroje kontingenčných tabuliek umožňujú tieto údaje porovnávať, filtrovať, odhaľovať vzťahy a analyzovať trendy.

	A	B	C	D	E	F	G
1	Predaj číslo	Produkt	Cena	Kusov	Predaj celkom	Kraj	Dátum
2	1	Klávesnica	129 EUR	20	2580	Prešovský	23/05/2018
3	2	Myš	35 EUR	12	420	Nitriansky	15/12/2018
4	3	Tablet	845 EUR	15	12675	Žilinský	01/10/2019
5	4	Sluchadlá	35 EUR	17	595	Nitriansky	25/06/2018
6	5	Klávesnica	45 EUR	7	315	Žilinský	03/05/2018
7	6	Tablet	1.183 EUR	13	15379	Banskobystrický	22/04/2018
8	7	Notebook	1.075 EUR	8	8600	Prešovský	29/10/2019
9	8	Myš	50 EUR	9	450	Trnavský	04/02/2019
10	9	Kindle	1.281 EUR	18	23058	Trenčiansky	30/10/2019
11	10	Notebook	373 EUR	15	5595	Trnavský	10/07/2019
12	11	Tablet	1.424 EUR	5	7120	Nitriansky	01/04/2019

Polia kontingenčnej..

Vyberte polia, ktoré chcete pridať do zostavy:

Prehľadávať

☒ Produkt
☐ Cena
☐ Kusov
☒ Predaj celkom
☐ Kraj
☐ Dátum

Presuňte polia medzi nižšie uvedenými oblasťami:

Filtre

Stĺpce

Riadky

Hodnoty

Produkt

Súčet z Predaj c...

☐ Odložiť aktualizáciu rozlož... Aktualizovať

Year 2014 .Y

Drag fields between areas below:

FILTERS

COLUMNS

ROWS

VALUES

Revenue	Column	Q1	Q2	Q3	Q4
Row Labels					
Midwest			1,590	2,000	5,170
Northeast		35	184	660	
South		483	1,702	15,879	
West		19,263	3,292	2,212	1,740
Grand Total		19,780	6,768	20,751	6,910

Označenia riadkov

Kindle	
Klávesnica	
Myš	26533
Notebook	160397
Počítač	84439
Sluchadlá	18293
Tablet	183261
Televízor	158579
Celkový súčet	767925

Kontingenčná tabuľka obsahuje počty výskytov znakov, ich súčty a medzisúčty alebo namiesto súčtov je možné použiť výpočet percent, odchýlok, priemerov...

	A	B	C
1	Property Type	Sales price	
2	Apartment	\$328,484	
3	House	\$1,145,078	
4	Apartment	\$587,737	
5	Apartment	\$522,867	
6	Apartment	\$321,706	
7	Condo	\$486,653	
8	Apartment	\$482,425	
9	Townhouse	\$931,004	
10	Condo	\$321,095	
11	Townhouse	\$548,455	
12	Apartment	\$376,895	
13	Condo	\$709,642	
14	Apartment	\$277,876	
15	Townhouse	\$838,370	
16	Condo	\$227,156	

Count of Apartments Sum of sales price

	A	B	C
1	Property Type	COUNTA of Property Ty	SUM of Sales price
2	Apartment	7	\$2,897,990
3	Condo	4	\$1,744,546
4	House	1	\$1,145,078
5	Townhouse	3	\$2,317,829
6	Grand Total	15	\$8,105,443
7			
8			

	A	B	C	D	E	F
1	Order	Order Date	Customer	Customer Name	Address	City
4	30023	02/27/14	27	Company AA	789 27th Street	Los V
5	30025	02/27/14	27	Company AA	789 27th Street	Los V
6	30016	02/04/14	4	Company D	123 40th Street	New
7	30016	02/04/14	4	Company D	123 40th Street	New
8	30016	02/04/14	4	Company D	123 40th Street	New
9	30017	02/12/14	12	Company L	123 12th Street	Los V
10	30017	02/12/14	12	Company L	123 12th Street	Los V
11	30018	02/06/14	8	Company H	123 8th Street	Portl
12	30019	02/04/14	4	Company D	123 40th Street	
13	30020	02/29/14	29	Company CC	789 29th Street	
14	30021	02/09/14	9	Company C	123 3rd Street	
15	30022	02/06/14	6	Company F	123 6th Street	
16	30023	02/26/14	26	Company BB	789 26th Street	
17	30024	02/08/14	8	Company H	123 8th Street	
18	30025	02/10/14	10	Company J	123 10th Street	
19	30026	02/07/14	7	Company G	123 7th Street	
20	30027	02/10/14	10	Company J	123 10th Street	
21	30027	02/10/14	10	Company J	123 10th Street	
22	30027	02/10/14	10	Company J	123 10th Street	CHIC
23	30028	02/11/14	11	Company K	123 11th Street	Mian
24	30028	02/11/14	11	Company K	123 11th Street	Mian
25	30029	02/01/14	1	Company A	123 1st Street	Seatt
26	30029	02/01/14	1	Company A	123 1st Street	Seatt
27	30029	02/01/14	1	Company A	123 1st Street	Seatt
28	30030	02/28/14	28	Company BB	789 28th Street	Minn
29	30030	02/28/14	28	Company BB	789 28th Street	Minn
30	30031	02/09/14	9	Company I	123 9th Street	Salt L
31	30031	02/09/14	9	Company I	123 9th Street	Salt L
32	30032	02/06/14	6	Company F	123 6th Street	Milwa
33	30033	02/06/14	6	Company H	123 8th Street	Portl
34	30034	02/03/14	3	Company C	123 3rd Street	Los A
35	30034	02/03/14	3	Company C	123 3rd Street	Los A



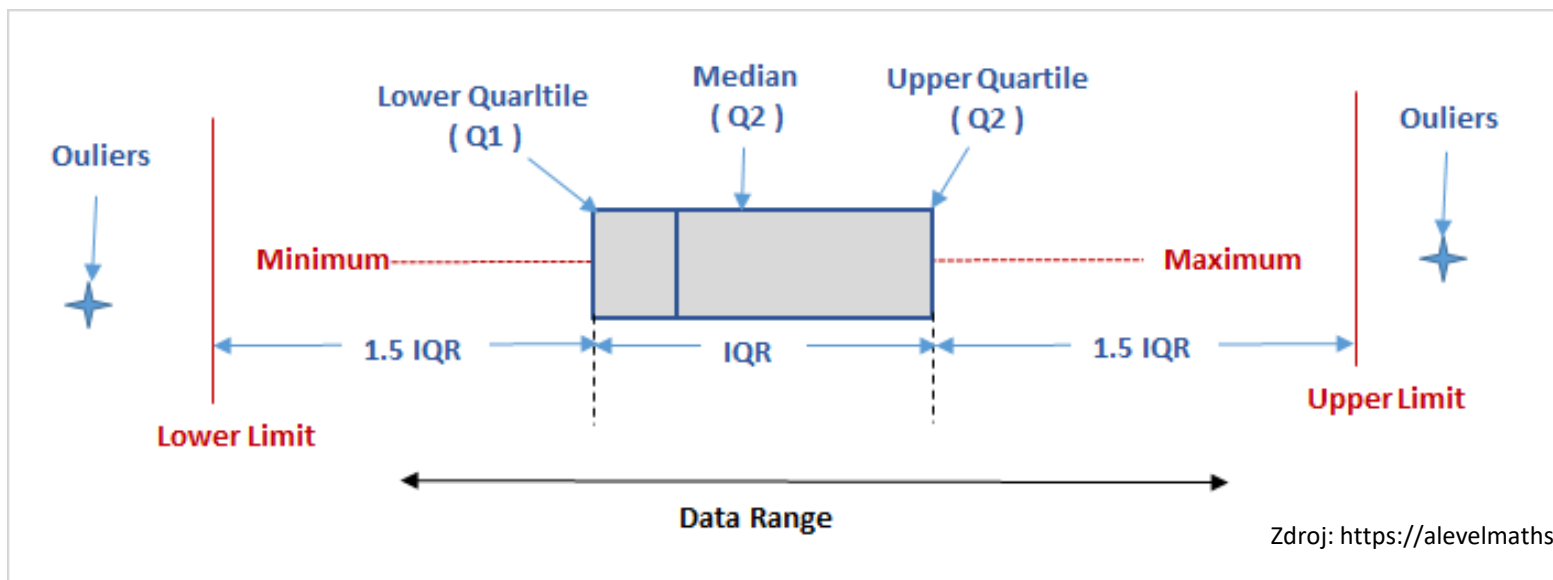
Zdrojové dáta možno ľahko vizualizovať pomocou kontingenčného grafu.

Krabicový graf (Box Plot, Box and Whisker Plot)

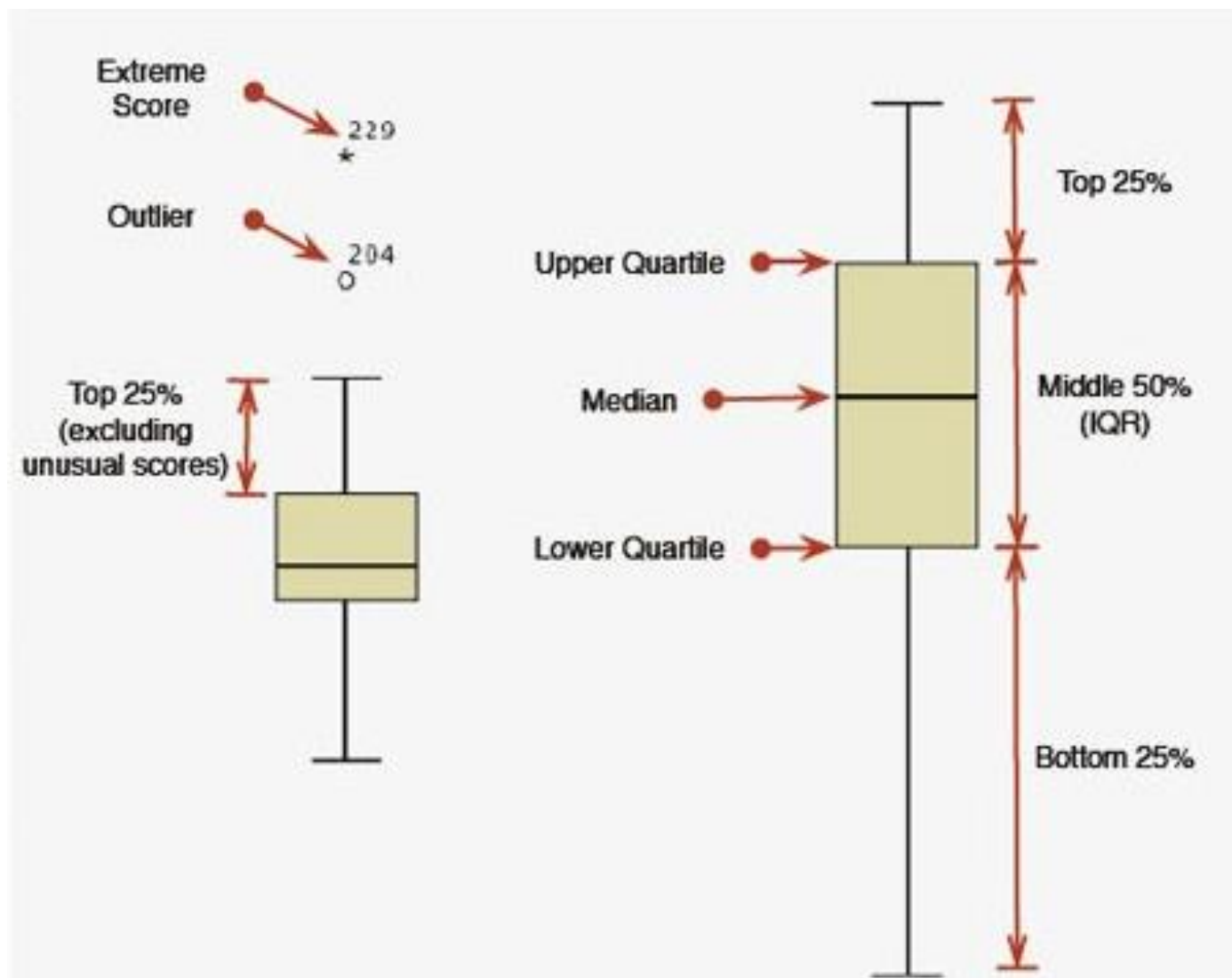
- patrí do exploračnej (predbežnej) analýzy údajov

Výhody:

- jednoduchý a rýchly pohľad na rozvrstvenie hodnôt v súbore (vidíme rozpätie medzi minimálnou i maximálnou hodnotu i rozpätie medzi jednotlivými kvartilmi – resp. ako je celkový počet jednotiek v súbore rozdelený na štvrtiny);
- poukazuje na prípadnú asymetriu - rozloženie dát pomocou kvartilov;
- zobrazuje odľahlé namerané hodnoty;
- histogram je ľahko pochopiteľný spôsob sumarizácie hodnôt, ale musíme dobre zvoliť veľkosť tried, do ktorých zoskupujeme hodnoty.

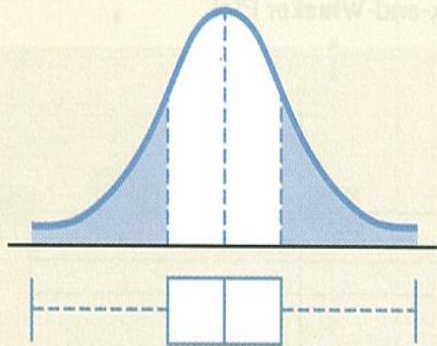


Krabicový graf (Box Plot, Box and Whisker Plot)

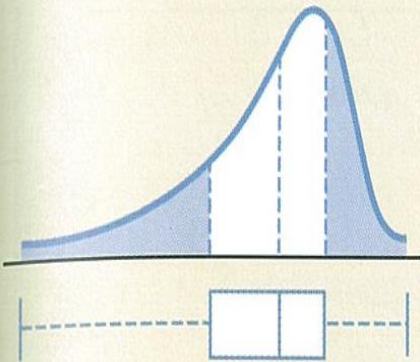


Box plot a k nemu prislúchajúci polygón

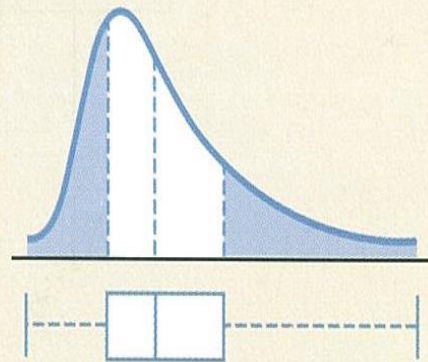
Plocha pod každým polygónom je rozdelená na štyri kvartily korešpondujúce s piatimi hodnotami príslušného box plotu.



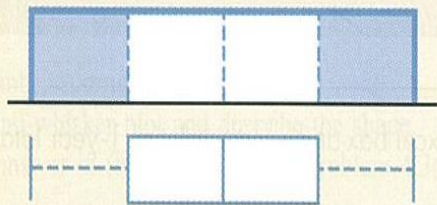
(a) Bell-Shaped Distribution



(b) Left-Skewed Distribution



(c) Right-Skewed Distribution

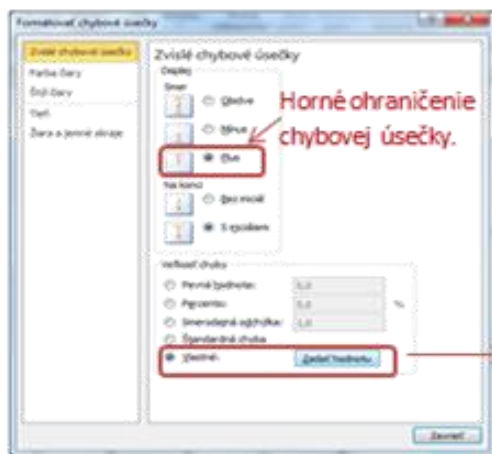


(d) Rectangular Distribution

- malý obdĺžnik a krátke úsečky znamenajú malú variabilitu;
- obdĺžnik posunutý doprava a dlhá úsečka vľavo znamená záporné zošikmenie
- obdĺžnik posunutý doľava a dlhá úsečka vpravo znamená kladné zošikmenie

Zobrazenie box plotu

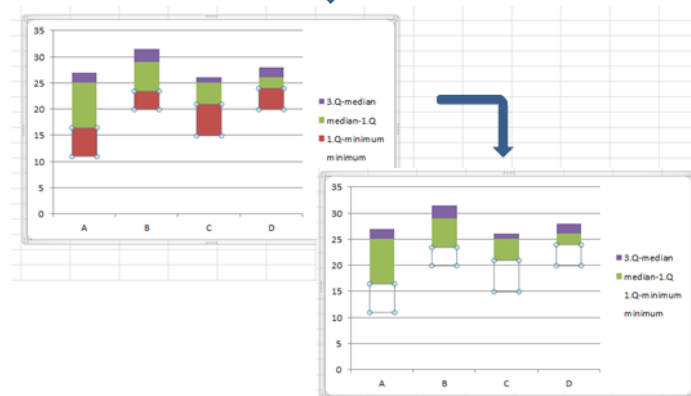
V staršej verzii Excelu pomocou pomocnej tabuľky, dvojrozmerného skladaného stĺpcového grafu a chybových úsečiek



Dialogové okná pri zadávaní vlastnej hodnoty veľkosti chyby:



	Výrobná linka			
	A	B	C	D
minimum	11	20	15	20
1.Q-minimum	5,5	3,5	6	4
median-1.Q	8,5	5,5	4	2
3.Q-median	2	2,5	1	2
maximum-3.Q	3	2,5	2	1



Excel ponúka zobrazenie tohto typu grafu ako škatuľový graf.

Vložiť graf

Odporúčané grafy Všetky grafy

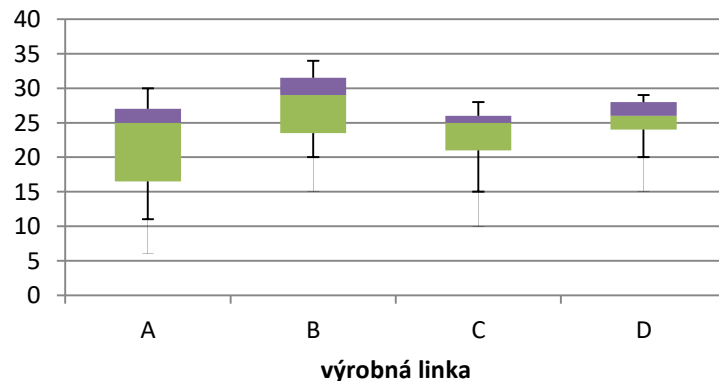
- Posledné
- Šablóny
- Stĺpcový
- Čiarový
- Koláčový
- Pruhový
- Plošný
- X Y (závislosť)
- Burzový
- Povrchový
- Radarový
- Stromová mapa
- Lúčový
- Histogram
- Škatuľový
- Vodopádový
- Kombinovaný

Škatuľový



počet kusov vyrobených výrobkov v
tisícoch

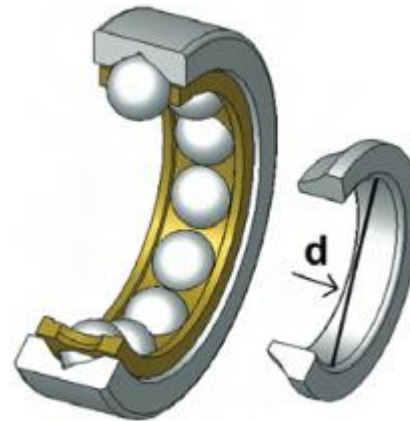
Porovnanie počtu vyrobených výrobkov na vybraných linkách



Príklad

Strojársená firma rozbieha výrobu guľčkových ložísk na troch strojoch, ktoré sú od rôznych výrobcov – stroje ALFA, BETA a GAMA. Podľa normy, vnútorný priemer ložiska d má byť 3,1 cm. Je povolená kladná odchýlka do hodnoty 3,19 cm. Menší byť nemôže, lebo potom by sa doň nezmestil ďalší komponent ložiska.

Matematický zápis: $d \in \langle 3,1 \text{ cm}; 3,19 \text{ cm} \rangle$



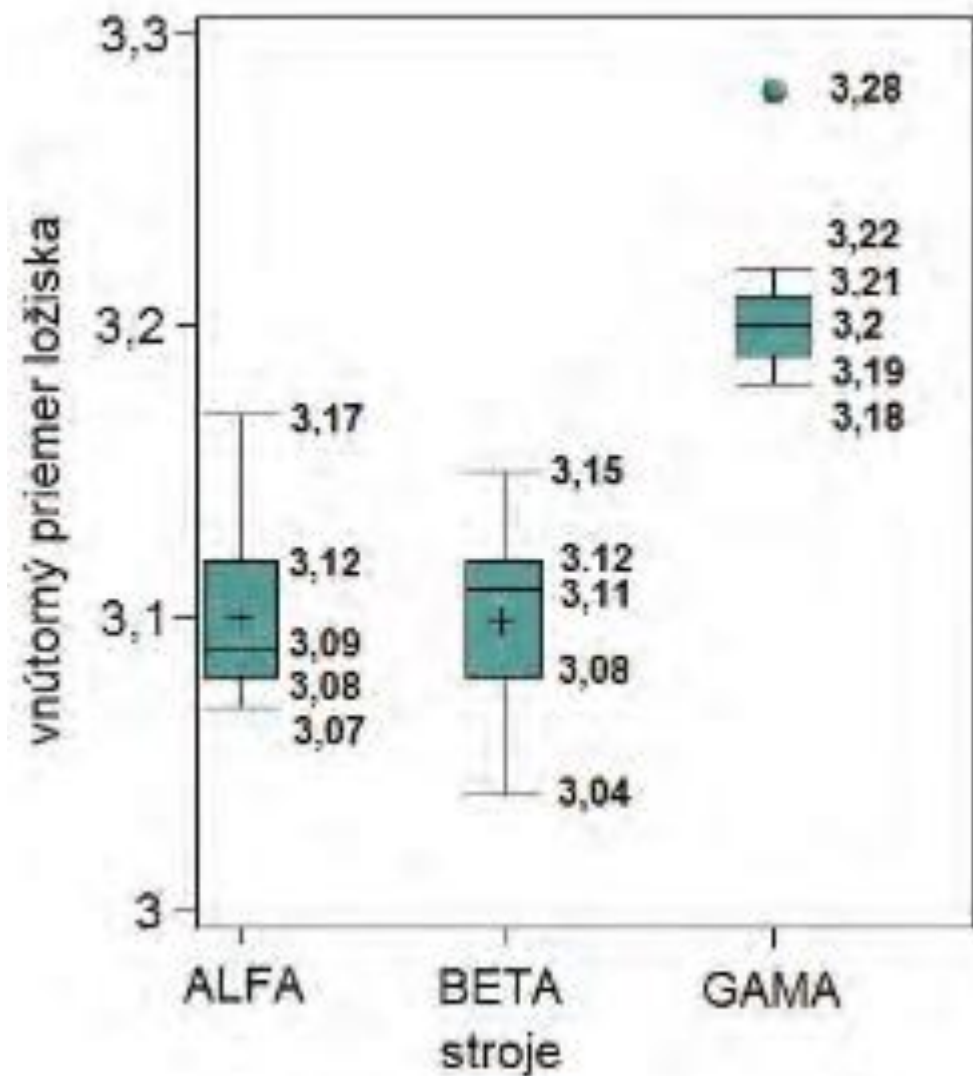
Zdroj: <https://sk.wikipedia.org/wiki>

Po prvej hodine skúšobnej prevádzky chcel kontrolór zistiť, na koľko dané stroje pri výrobe ložísk spĺňajú normu. Ak by zistil porušenie normy, stroj sa musí „preladiť“, čiže znova nastaviť. Preto náhodne vybral z každého stroja po 15 vyrobených kusov ložísk a zmeral ich vnútorný priemer.

Namerané hodnoty:

ALFA	3,14	BETA	3,1	GAMA	3,18
ALFA	3,1	BETA	3,12	GAMA	3,2
ALFA	3,12	BETA	3,05	GAMA	3,21
ALFA	3,17	BETA	3,04	GAMA	3,2
ALFA	3,11	BETA	3,12	GAMA	3,18
ALFA	3,17	BETA	3,07	GAMA	3,21
ALFA	3,08	BETA	3,11	GAMA	3,2
ALFA	3,09	BETA	3,13	GAMA	3,19
ALFA	3,07	BETA	3,11	GAMA	3,2
ALFA	3,08	BETA	3,12	GAMA	3,19
ALFA	3,08	BETA	3,08	GAMA	3,21
ALFA	3,07	BETA	3,15	GAMA	3,18
ALFA	3,08	BETA	3,11	GAMA	3,22
ALFA	3,09	BETA	3,08	GAMA	3,28
ALFA	3,08	BETA	3,12	GAMA	3,19
priemer:	3,10	priemer:	3,10	priemer:	3,20

požadovaný vnútorný priemer ložiska: $d \in \langle 3,1 \text{ cm}; 3,19 \text{ cm} \rangle$



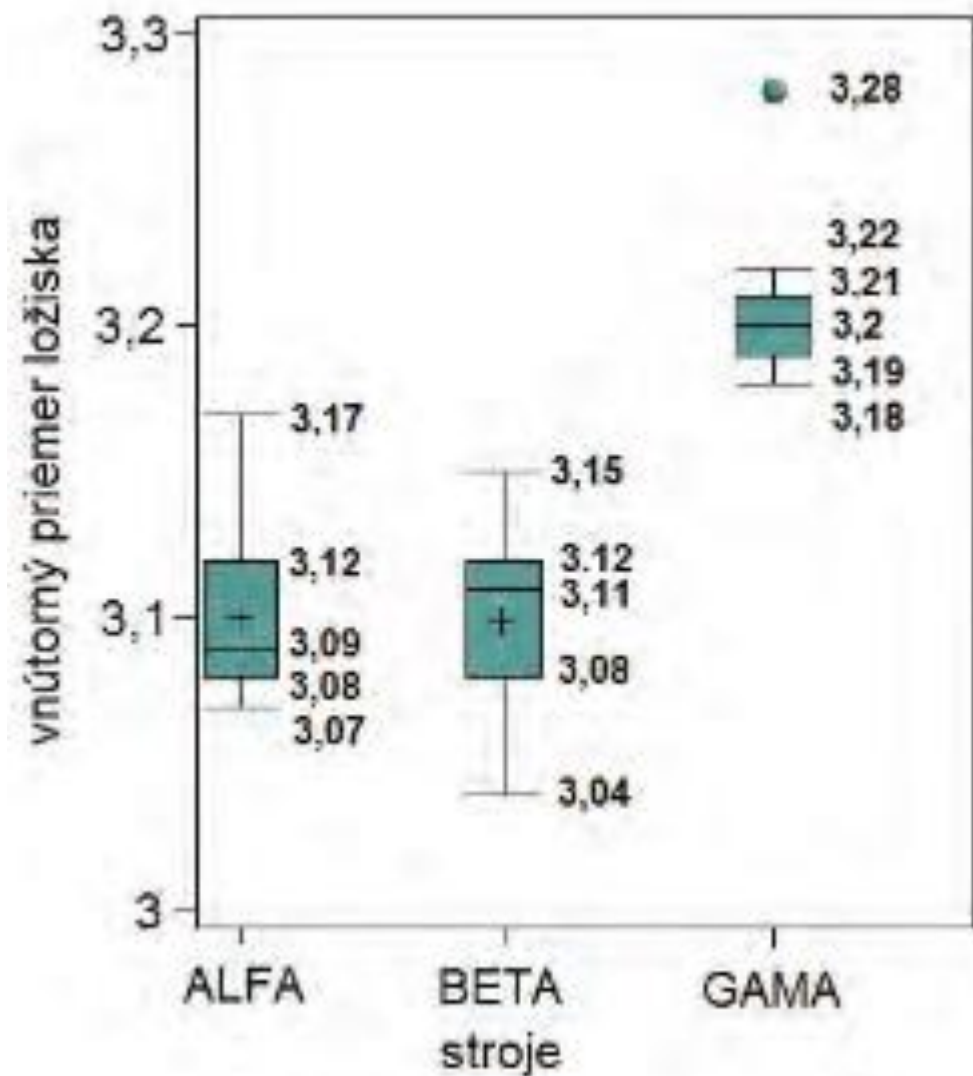
Stroj **ALFA**:

Kedže medián 3,09 cm je menší ako aritmetický priemer, väčšina ložísk má menší vnútorný priemer, ako je minimálna povolená hranica.

Ak sa pozrieme na namerané údaje, vidíme, že len 6 hodnôt z 15 vyhovuje norme (čo predstavuje len 40% náhodne vybratých ložísk).

Vzdialenosť medzi minimálnou hodnotou a mediánom (0,02 cm) je menšia ako vzdialenosť medzi mediánom a maximálnou hodnotou (0,08 cm). Z tohto zistenia a zo vzájomnej polohy mediánu a priemeru vidíme kladné zošikmenie údajov.

požadovaný vnútorný priemer ložiska: $d \in \langle 3,1 \text{ cm}; 3,19 \text{ cm} \rangle$

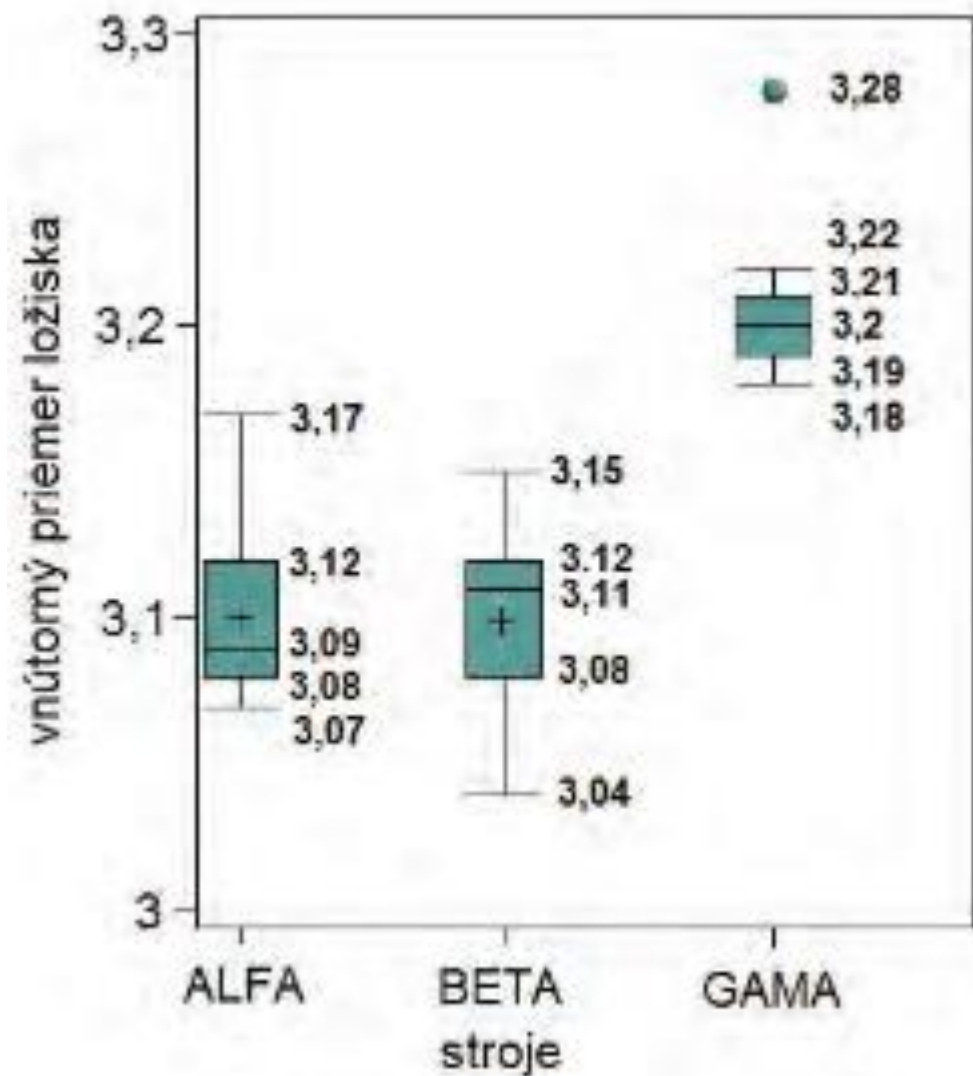


Stroj **BETA**:

Medián 3,11 cm je väčší ako aritmetický priemer, preto môžeme tvrdiť, že väčšina ložísk vyrobených na tomto stroji spĺňa normu. Zo zistených údajov naozaj vidíme, že 10 z 15 vybratých ložísk (2/3 výberu) je v žiadanom intervale.

Pri vzájomnom porovnaní strojov ALFA a BETA pomocou predchádzajúcich faktov konštatujeme, že aj keď odchýlky vnútorných priemerov ložísk od optimálnej hodnoty (3,1 cm) aj ich aritmetické priemery sa rovnajú, predsa stroj BETA je lepší – viac jeho výrobkov spĺňa normu. U oboch strojov kontrolór odporúča znížiť variabilitu a nastaviť stroje na o niečo vyššiu hranicu ako je 3,1 cm.

požadovaný vnútorný priemer ložiska: $d \in \langle 3,1 \text{ cm}; 3,19 \text{ cm} \rangle$



Stroj **GAMA**:

Z box plotu vidíme, že jedno číslo sa vymyká z „normálu“ – hodnota 3,28 cm je vyznačená ako extrémna. Vzhľadom na malé odchýlky ostatných hodnôt od priemeru (rozdiel medzi maximálnou a minimálnou hodnotou v box plote je len 0,4 cm) možno predpokladať, že v tomto prípade ide o chybu merania. Preto túto hodnotu v ďalšej analýze nebudeme brať do úvahy a vylúčime ju zo súboru.

Paradoxne, aj keď podľa aritmetického priemeru výrobky tohto stroja nevyhovujú norme, tento stroj vyrába ložiská najpresnejšie – s najmenšou variabilitou. Kontrolór odporučí zmenšiť hranicu nastavenia.

Výpočet výberových charakteristík

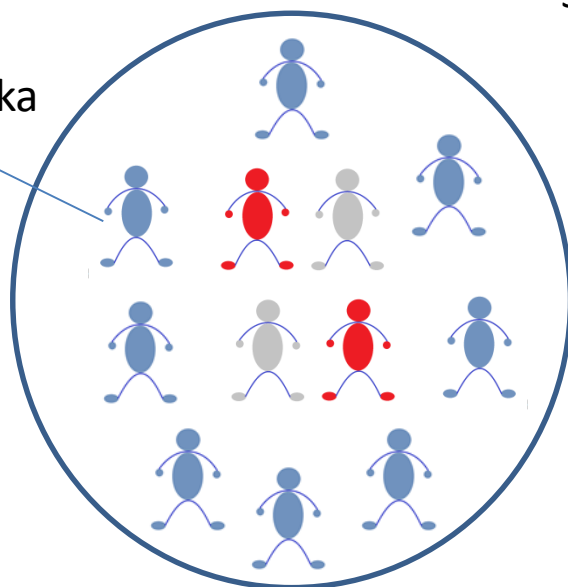
vyčerpávajúce skúmanie

výberové skúmanie

ZS (populácia)

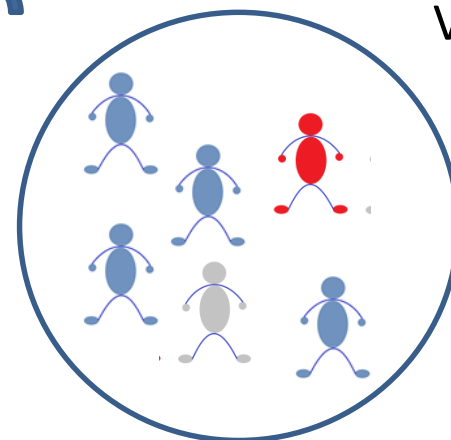
štatistická indukcia

štatistická jednotka



náhodný
výber

VS (vzorka)



Parametre ZS:

M resp. (μ) – stredná hodnota;

σ^2 – rozptyl

σ – smerodajná odchýlka;

Q – ľubovoľný parameter;

Charakteristiky VS:

$x_1, x_2, \dots, x_n$ namerané hodnoty,

n je rozsah výberového súboru;

$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ – aritmetický priemer
(stredná hodnota);

$s_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ – rozptyl
výberového súboru...

Parametre ZS:

M resp. (μ) – stredná hodnota $\mu = \frac{1}{n} \sum_{i=1}^n x_i$;

σ^2 – rozptyl $\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$;

σ – smerodajná odchýlka $\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2}$;

Na výpočet uvedených parametrov je potrebné poznať hodnoty štatistického znaku X na každej štatistickej jednotke ZS.

ZS $\longrightarrow$ veľký počet VS

Vzhľadom na konkrétny VS je jeho **(výberová) charakteristika** konštantná veličina, ale z hľadiska skúmania ZS je to náhodná premenná s vlastným zákonom (výberového) rozdelenia.

Každé výberové rozdelenie závisí od:

- rozdelenia znaku X ;
- funkcie výberovej charakteristiky;
- rozsahu VS.

Výberové charakteristiky používame pri odhadovaní parametrov ZS, testovaní hypotéz o vlastnostiach týchto parametrov ZS ...

Výberové rozdelenie:

Výberové rozdelenie je rozdelenie ako každé iné – popisujeme ho funkciami a charakteristikami.

Z charakteristík výberových rozdelení nás zaujíma predovšetkým stredná hodnota a rozptyl; napr. pri rozdelení výberového priemeru $\bar{X}$ je $E(\bar{X})$ stredná hodnota výberového priemeru a $D(\bar{X})$ rozptyl výberového priemeru.

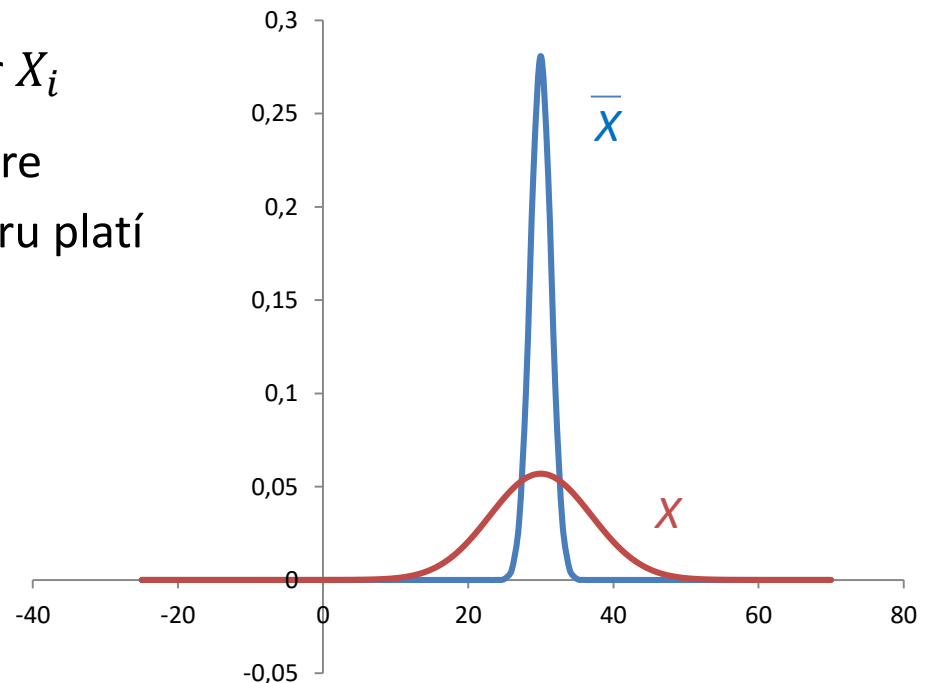
Ak $X_1, X_2, \dots, X_n$ tvoria náhodný výber X_i

a $E(X_i) = \mu$ a $D(X_i) = \sigma^2$, potom pre strednú hodnotu výberového priemeru platí

$$E(\bar{X}) = \mu$$

rozptyl výberového priemeru platí

$$D(\bar{X}) = \frac{\sigma^2}{n}$$



Výberové číselné charakteristiky

Výberový začiatkový moment k-teho stupňa

$$\tilde{z}_k = \frac{1}{n} \sum_{i=1}^n x_i^k \cdot n_i \text{ pre triedený súbor (vážený začiatkový moment)}$$

$$\tilde{z}_k = \frac{1}{n} \sum_{i=1}^n x_i^k \text{ pre netriedený súbor}$$

$$k = 1: \tilde{z}_1 = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i - \text{výberový aritmetický priemer ;}$$

Výberový centrálny moment k-teho stupňa

$$\tilde{c}_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k \cdot n_i \text{ pre triedené údaje (vážený centrálny moment)}$$

$$\tilde{c}_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k \text{ pre netriedené údaje}$$

$$k = 1: \tilde{c}_1 = 0$$

$$k = 2: \tilde{c}_2 = s_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \text{ je rozptyl VS}$$

Poznámka:

$$s_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \text{ je výberový rozptyl}$$

Výberové charakteristiky polohy

1. aritmetický priemer $\tilde{z}_1 = \bar{x}$; =AVERAGE

- vyjadruje, aký objem hodnôt štatistického znaku pripadá v priemere na jednu jednotku súboru

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

$$\bar{x} = \frac{1}{N} \sum_{j=1}^m x_j n_j \quad (\text{vážený aritmetický priemer pri triedenom súbore})$$

Vlastnosti aritmetického priemeru:

1. je citlivý na extrémne hodnoty
2. súčet odchýlok hodnôt znaku od jeho aritmetického priemeru je nulový;
3. ak pripočítame ku všetkým hodnotám ľubovoľnú konštantu, potom aritmetický priemer sa líši od pôvodného priemeru o túto konštantu;
4. ak všetky hodnoty znaku násobíme nenulovou konštantou, potom i aritmetický priemer z týchto hodnôt získame násobením pôvodného priemeru touto konštantou
5. súčet štvorcov odchýlok premennej od určitej konštanty je najnižší, ak daná konštantá je rovná priemeru tejto premennej.

Výberové charakteristiky polohy

2. Modus $Mo = \hat{x}$ - najčastejšie sa vyskytujúca hodnota vo VS =MODE
3. Medián $Me = \tilde{x}$ - tá hodnota, ktorá delí usporiadaný VS na dve rovnako početné časti
=MEDIAN

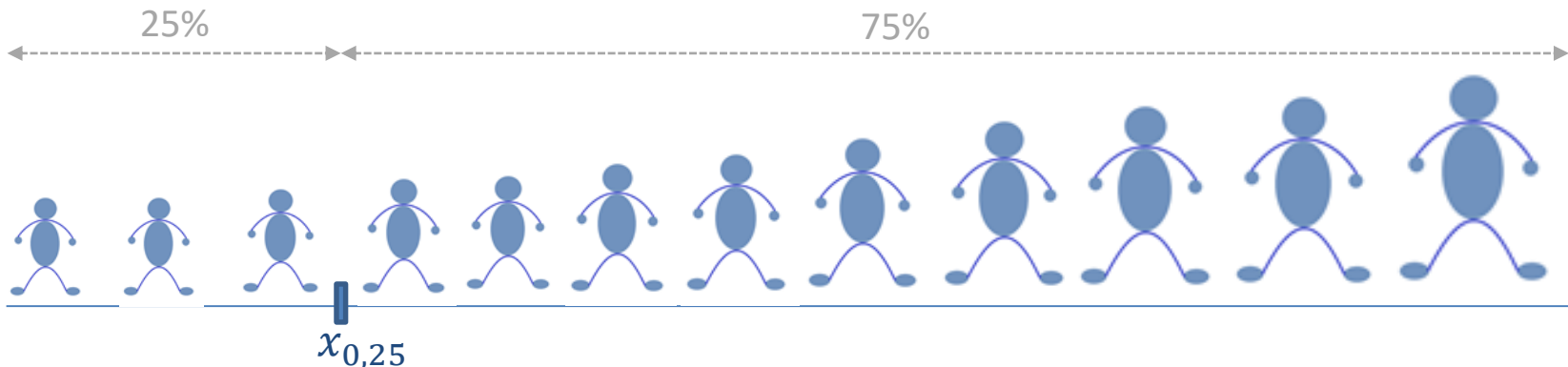
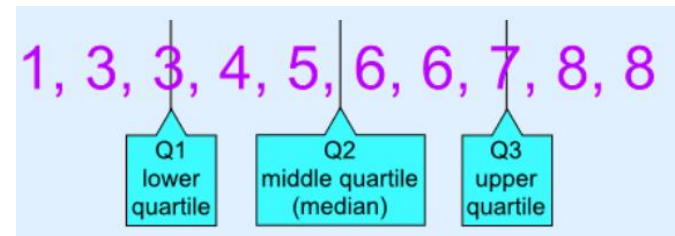
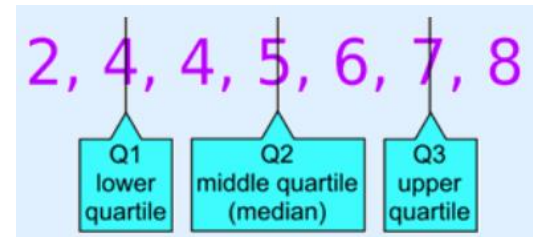
4. Kvantily

VS delíme na k rovnako početných častí

$k = 4$ — =QUARTILE

$k = 100$ — =PERCENTILE

k-ty percentil je hodnota, pod ktorou leží $k\%$ menších hodnôt v súbore usporiadanom podľa veľkosti vzostupne



Výberové charakteristiky variability

- **absolútne miery**

1. variačné rozpätie $R = x_{max} - x_{min}$; =RANGE

2. kvantilové rozpätie

- kvartilové rozpätie $IQR = R_{Q4} = \tilde{x}_{0,75} - \tilde{x}_{0,25}$

- decilové rozpätie $R_{Q10} = \tilde{x}_{0,9} - \tilde{x}_{0,1}$

- kvartilové odchýlka $Q = \frac{\tilde{x}_{0,75} - \tilde{x}_{0,25}}{2}$

- priemerná absolútna chyba $\bar{d} = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}| n_i$ =AVEDEV

3. rozptyl VS

$$s_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 n_i \quad =\text{VAR.P}$$

smerodajná (štandardná) odchýlka $s_n = \sqrt{s_n^2}$ =STDEV.P

4. výberový rozptyl (neskreslený bodový odhad rozptylu ZS)

$$s_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 n_i \quad =\text{VAR.S}$$

výberová smerodajná (štandardná) odchýlka $s_{n-1} = \sqrt{s_{n-1}^2}$ =STDEV.S

Výberové charakteristiky variability

- **relatívne miery**

1. variačný koeficient $v_k = \frac{s_{n-1}}{\bar{x}} \cdot 100(\%)$

čím je jeho hodnota väčšia, tým je súbor nerovnorodejší

2. pomerná priemerná odchýlka $\bar{d}_p = \frac{\bar{d}}{\bar{x}} \cdot 100(\%)$

Miera šikmosti (asymetrie)

Pearsonova miera šikmosti $S_P = \frac{\bar{x} - M_o}{s_{n-1}}$

Kvartilová miera šikmosti $S_Q = \frac{(\tilde{x}_{0,75} - \tilde{x}_{0,5}) - (\tilde{x}_{0,5} - \tilde{x}_{0,25})}{(\tilde{x}_{0,75} - \tilde{x}_{0,5}) + (\tilde{x}_{0,5} - \tilde{x}_{0,25})}$

Momentová miera šikmosti (koeficient šikmosti) $\alpha = \frac{\tilde{c}_3}{s^3} = \text{SKEW}$

Miera špicatosti

meria hustotu chvostov rozdelenia premennej, teda charakterizuje výskyt extrémne vysokých a extrémne nízkych hodnôt.

Koeficient špicatosti $\beta = \frac{\tilde{c}_4}{s^4} - 3$ =KURT

Špicatosť rozdelenia sa porovnáva so špicatosťou normálneho rozdelenia.

<https://www.viemematiku.sk/rozhodovacka-statistika-pojmy-3-uroven/7546>

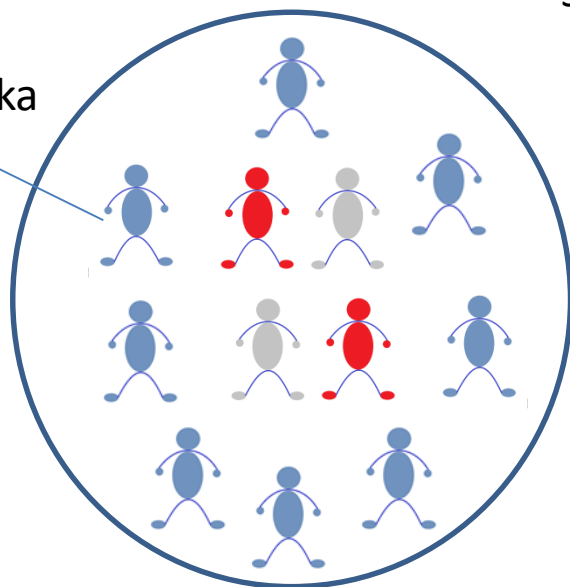
vyčerpávajúce skúmanie

výberové skúmanie

ZS (populácia)

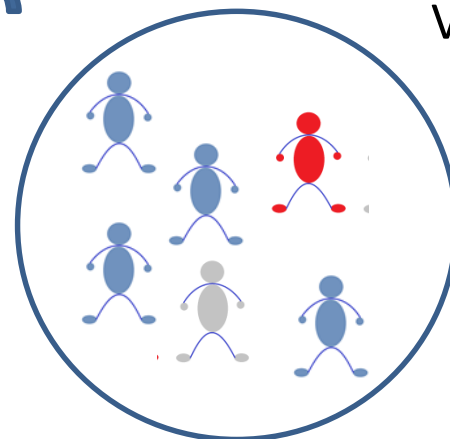
štatistická indukcia

štatistická jednotka



náhodný
výber

VS (vzorka)



Parametre ZS:

M resp. (μ) – stredná hodnota;

σ^2 – rozptyl

σ – smerodajná odchýlka;

Q – ľubovoľný parameter;

Charakteristiky VS:

$x_1, x_2, \dots, x_n$ namerané hodnoty,

n je rozsah výberového súboru;

$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ – aritmetický priemer
(stredná hodnota);

$s_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ – rozptyl
výberového súboru...

u_n – ľubovoľná charakteristika

Induktívna štatistika

Na základe výberového skúmania robíme odhady parametrov ZS:



Bodový odhad (estimátor)

parametra ZS odhadneme jedným číslom (bodom) vypočítaným z VS (pomocou výberových charakteristík)

$$estM = \bar{x},$$

$$est \sigma^2 = s_{n-1}^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2.$$



Intervalový odhad parametra Q ZS

robíme s pravdepodobnosťou $1 - \alpha$. Pravdepodobnosť, že Q leží v intervale (q_1, q_2) zapisujeme:

$$P(q_1 < Q < q_2) = 1 - \alpha,$$

$1 - \alpha$ je spoľahlivosť intervalového odhadu ;

α sa nazýva riziko intervalového odhadu; náhodné premenné q_1, q_2 sú medze intervalu;

(q_1, q_2) je interval spoľahlivosti – konfidenčný interval.

Výhoda intervalových odhadov oproti bodovým odhadom je v tom, že nesú informáciu o neistote (nepresnosti) odhadu.

Bodovým odhadom parametra Q rozumieme výberovú charakteristiku (štatistiku) u_n , ktorej hodnoty kolíšu okolo parametra Q .

Zo všetkých charakteristík vyberáme takú, ktorá najlepšie aproximuje skutočnú hodnotu parametra Q pomocou:

1. **nestrannosti** – $E(u_n) = Q$

(bodový odhad skutočnú hodnotu parametra Q systematicky nenadhodnocuje ani nepodhodnocuje; nestranný odhad = nevychýlený, neskreslený odhad; $E(u_n) - Q$ nazývame vychýlenie odhadu u_n)

2. **výdatnosti** – u_n^* je najlepší (výdatnejší) nestranný odhad parametra Q ak platí
 $D(u_n^*) < D(u_n)$

(ak máme k dispozícii viac nestranných odhadov, vyberieme ten s najmenším rozptylom)

3. **konzistencie** – u_n nazveme konzistentným odhadom Q , ak pre každé $\varepsilon > 0$ platí $\lim_{n \rightarrow \infty} P(|u_n - Q| < \varepsilon) = 1$

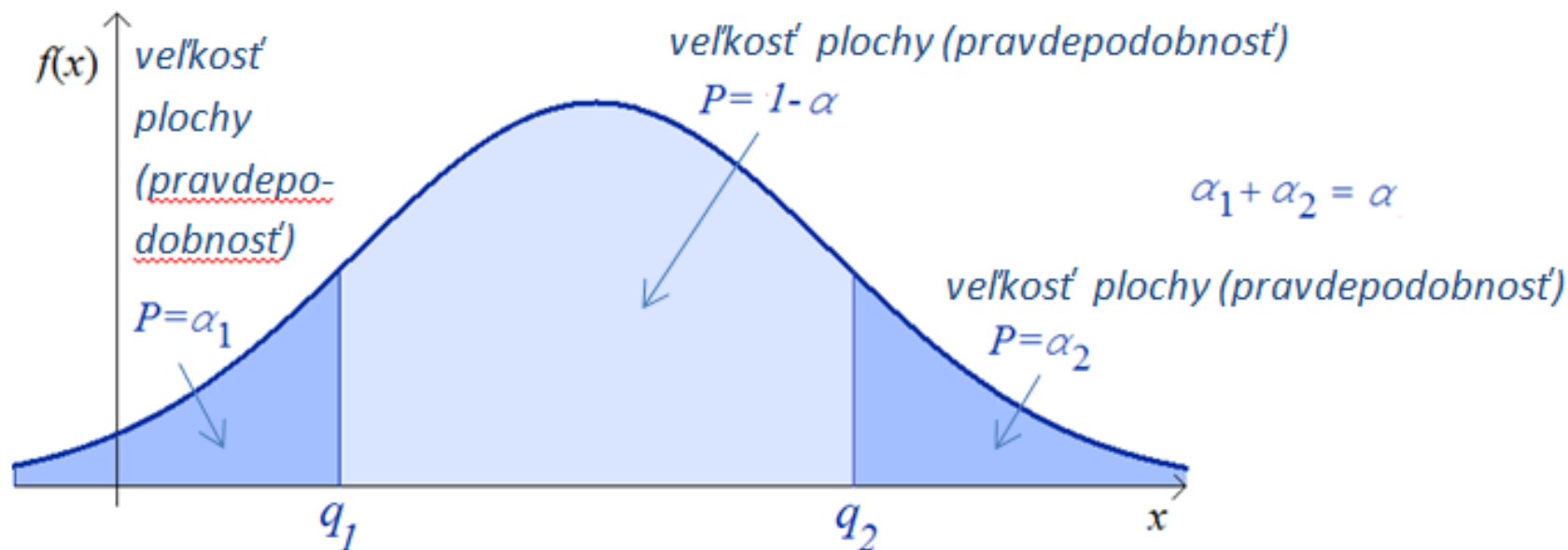
(s rozsahom výberu sa zväčšuje pravdepodobnosť, že bodový odhad bude blízky Q)

4. **presnosti** – posudzujeme pomocou výberovej smerodajnej chyby odhadu SE

napr. smerodajná chyba odhadu strednej hodnoty $\widehat{SE}(\bar{X}) = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{s_{n-1}}{\sqrt{n}}$

Intervalový odhad parametra Q sa nazýva odhad pomocou číselného intervalu, v ktorom sa odhad parametra nachádza s určitou pravdepodobnosťou.

$$P(q_1 < Q < q_2) = 1 - \alpha$$



Pre obojstranný interval spoľahlivosti je $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$ obojstranne symetrický.

100(1- α)% obojstranný interval spoľahlivosti pre strednú hodnotu

$$X \sim N(M, \sigma^2)$$

Utvorme náhodnú premennú $t = \frac{\bar{x} - M}{\frac{s_{n-1}}{\sqrt{n}}} \sim t(n-1)$

Zo symetrie hustoty rozdelenia pravdepodobnosti vyplýva:

$t_{n-1}\left(\frac{\alpha}{2}\right) = -t_{n-1}\left(1 - \frac{\alpha}{2}\right)$, kde $t_{n-1}\left(1 - \frac{\alpha}{2}\right)$ je $\left(1 - \frac{\alpha}{2}\right) \cdot 100\%$ -ný kvantil.

$$P\left(-t_{n-1}^{kv}\left(1 - \frac{\alpha}{2}\right) < \frac{\bar{x} - M}{\frac{s_{n-1}}{\sqrt{n}}} < t_{n-1}^{kv}\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha$$

$$P\left(\bar{x} - \frac{s_{n-1}}{\sqrt{n}} t_{n-1}^{kv}\left(1 - \frac{\alpha}{2}\right) < M < \bar{x} + \frac{s_{n-1}}{\sqrt{n}} t_{n-1}^{kv}\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha$$

Krátko zapísané: $\bar{x} \pm \frac{s_{n-1}}{\sqrt{n}} t_{n-1}^{kv}\left(1 - \frac{\alpha}{2}\right)$

$\Delta = \frac{s_{n-1}}{\sqrt{n}} t_{n-1}^{kv}\left(1 - \frac{\alpha}{2}\right)$ je prípustná chyba odchýlky $SE = \frac{s_{n-1}}{\sqrt{n}}$ je výberová chyba

$100(1-\alpha)\%$ pravostranný interval spoľahlivosti pre strednú hodnotu M

$$P\left(-\infty < M < \bar{x} + \frac{s_{n-1}}{\sqrt{n}} t_{n-1}^{kv}(1 - \alpha)\right) = 1 - \alpha$$

$100(1-\alpha)\%$ ľavostranný interval spoľahlivosti pre strednú hodnotu M

$$P\left(\bar{x} - \frac{s_{n-1}}{\sqrt{n}} t_{n-1}^{kv}(1 - \alpha) < M < \infty\right) = 1 - \alpha$$

100(1- α)% obojstranný interval spoľahlivosti pre rozptyl σ^2

$$X \sim N(M, \sigma^2)$$

Utvorme náhodnú premennú $\chi^2 = \frac{n-1}{\sigma^2} s_{n-1}^2 \sim \chi^2(n-1)$.

$$P\left(\chi_{n-1}^{2\ kv}\left(\frac{\alpha}{2}\right) < \frac{n-1}{\sigma^2} s_{n-1}^2 < \chi_{n-1}^{2\ kv}\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha$$

$$P\left(\frac{(n-1) \cdot s_{n-1}^2}{\chi_{n-1}^{2\ kv}\left(1 - \frac{\alpha}{2}\right)} < \sigma^2 < \frac{(n-1) \cdot s_{n-1}^2}{\chi_{n-1}^{2\ kv}\left(\frac{\alpha}{2}\right)}\right) = 1 - \alpha$$

100(1- α)% obojstranný interval spoľahlivosti pre smerodajnú odchýlku σ

$$P\left(\sqrt{\frac{(n-1)s_{n-1}^2}{\chi_{n-1}^{2\ kv}\left(1 - \frac{\alpha}{2}\right)}} < \sigma < \sqrt{\frac{(n-1)s_{n-1}^2}{\chi_{n-1}^{2\ kv}\left(\frac{\alpha}{2}\right)}}\right) = 1 - \alpha$$

$$\text{kde } s_{n-1}^2 = \text{est } \sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

100(1- α)% pravostranný interval spoľahlivosti pre rozptyl σ^2

$$P\left(0 < \sigma^2 < \frac{(n-1) \cdot s_{n-1}^2}{\chi_{n-1}^{2\ kv}(\alpha)}\right) = 1 - \alpha$$

100(1- α)% ľavostranný interval spoľahlivosti pre rozptyl σ^2

$$P\left(\frac{(n-1) \cdot s_{n-1}^2}{\chi_{n-1}^{2\ kv}(1-\alpha)} < \sigma^2 < \infty\right) = 1 - \alpha$$

$$\text{kde } s_{n-1}^2 = \text{est } \sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Príklad

Určte rozptyl, smerodajnú odchýlku, výberový rozptyl a výberovú smerodajnú odchýlku z hodnôt 1,2,5,6,7,8,8,9.

$$\rightarrow 1, 2, 5, 6, 7, 8, 8, 9 \quad ; \quad n = 8$$

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i \Rightarrow \bar{X} = \frac{1}{8} (1+2+5+\dots+9) = 5,75$$

= AVERAGE

rozptyl vs

$$s_x^2 = \frac{1}{n} \sum (x_i - \bar{X})^2 \Rightarrow s_x^2 = \frac{1}{8} \left(\underbrace{(1-5,75)^2 + (2-5,75)^2 + \dots}_{= \text{DEVSQ}} + \underbrace{(9-5,75)^2}_{= \text{VAR.P}} \right) = 7,4375$$

$$s_x^2 = \overline{x^2} - \bar{X}^2$$

$$\overline{x^2} = \frac{1}{n} \sum x_i^2 \Rightarrow \frac{1}{8} (1^2 + 2^2 + \dots + 9^2) = 40,5 \quad \left. \vphantom{\overline{x^2}} \right\} s_x^2 = 40,5 - 5,75^2 = 7,4375$$

smerodajná odchýlka

$$s_x = \sqrt{s_x^2} \Rightarrow s_x = \sqrt{7,4375} = \text{STDEV.P}$$

výberový rozptyl

$$s_{n-1}^2 = \frac{1}{n-1} \sum (x_i - \bar{X})^2 \Rightarrow s_{n-1}^2 = \frac{1}{7} \left(\underbrace{(1-5,75)^2 + \dots}_{= \text{DEVSQ}} + \underbrace{(9-5,75)^2}_{= \text{VAR.S}} \right) = 8,5$$

výberová smerodajná odchýlka

$$s_{n-1} = \sqrt{s_{n-1}^2} \Rightarrow s_{n-1} = \sqrt{8,5} \approx 2,91548$$

Príklad

Mliekareň dosiahla pri balení mlieka a masla na dvoch automatických linkách tieto kontrolné výsledky: pre mlieko je $\bar{x} = 1,036$ l a $s_n(x) = 0,284$ l, pre maslo je $\bar{y} = 247,9$ g a $s_n(y) = 42,6$ g. Ktorá linka bola v rámci kontroly presnejšia?

→ mlieko : $\bar{x} = 1,036$ l ; $s_{n-1}(x) = 0,284$ l

maslo : $\bar{y} = 247,9$ g ; $s_{n-1}(y) = 42,6$ g

variálny koeficient $\sigma_x = \frac{s_{n-1}(x)}{\bar{x}} = \frac{0,284}{1,036} = 0,274$

$\sigma_y = \frac{s_{n-1}(y)}{\bar{y}} = \frac{42,6}{247,9} = 0,172$

} $\Rightarrow$ príčina presnosti
dosiahla linka
na balení masla

(Ak menší je variálny koeficient, tým je súbor homogénnejší;
 $\sigma > 0,5$ zn. veľa odlišných hodnôt)

Príklad

Pri zisťovaní spotreby benzínu určitého typu automobilu pri rýchlosti 90 km/h boli u 20 náhodne vybraných automobiloch zistené tieto hodnoty (v l/100 km):

6,5 6,8 6,7 6,0 5,6 6,6 5,5 6,4 5,5 6,5 6,3 6,2 6,3 5,9 5,8 6,4 6,5 6,3 5,7 6,1.

Predpokladáme, že spotreba benzínu je náhodná premenná s normálnym rozdelením.

- Odhadnite priemernú spotrebu benzínu u sledovaného typu automobilu a danej rýchlosti.
- Určte, v akých medziach je možné so spoľahlivosťou 95% očakávať priemernú spotrebu benzínu u tohto typu automobilu.
- S určte 95% spoľahlivosťou určte priemernú spotrebu, ktorá nebude prekročená.

→ X - spotreba benzínu ; $X \sim N(\mu, \sigma^2)$

$$\bar{x} = \frac{1}{20}(6,5 + 6,8 + \dots + 6,1) = 6,18 ; s_{n-1}^2 = \frac{1}{19}((6,5 - 6,18)^2 + \dots + (6,1 - 6,18)^2) = 0,15958$$

a) $\text{odhad } \mu = \bar{x} = 6,18$

$$s_{n-1} = 0,39947$$

b) 95% IS pre μ : $\bar{x} \pm \Delta \Rightarrow 6,18 \pm \frac{0,39947}{\sqrt{20}} \cdot 2,093 = 6,18 \pm 0,186959277$

$$\Delta = \frac{s_{n-1}}{\sqrt{n}} \cdot t_{n-1} \left(1 - \frac{\alpha}{2}\right)$$

$$P(5,993 < \mu < 6,367) = 0,95$$

c) $P(\mu < \bar{x} + t_{n-1} \left(1 - \frac{\alpha}{2}\right) \cdot \frac{s_{n-1}}{\sqrt{n}}) = 1 - \alpha$

$$\mu < 6,18 + 1,729 \cdot \frac{0,399}{\sqrt{20}} = 6,334$$

Príklad

Výrobca dodáva výrobky odberateľom v balíkoch s predpísanou hmotnosťou 250 g. Za účelom posúdenia presnosti baliaceho automatu bolo náhodne vybraných 15 balíkov a zistená ich hmotnosť (v g): 256 266 240 260 235 263 259 234 258 253 265 234 245 237

Na základe uskutočneného merania posúďte, či baliaci automat balí dostatočne presne, t.j. či variabilita balíčkov vyjadrená smerodajnou odchýlkou neprekračuje stanovenú hodnotu 15 g. Predpokladáme, že hmotnosť balíkov je náhodná premenná s normálnym rozdelením.

→ X - hmotnosť balíku ; $X \sim N(\mu, \sigma^2)$

$n = 15$, predpísaná hmotnosť = 250 ; $s_{n-1}^2 = 150,838$; $\chi_{14}^{2,0.95} = 6,570631$
odchýlka < 15

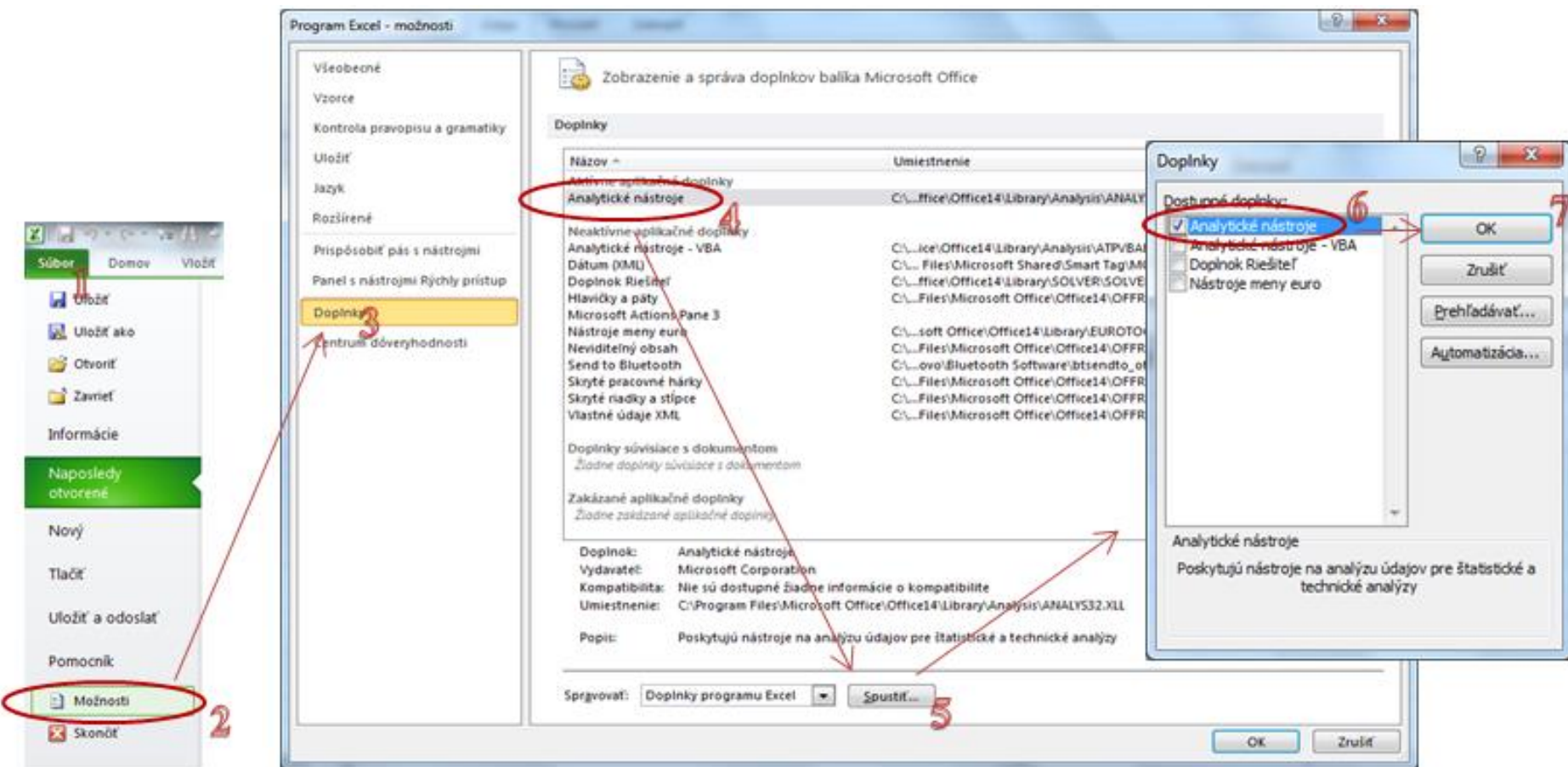
$$P\left(\sigma^2 < \frac{(n-1) \cdot s_{n-1}^2}{\chi_{n-1}^2(\alpha)}\right) = 1 - \alpha$$

$$0 < \sigma^2 < \frac{14 \cdot 150,838}{6,570631} = 321,42 \Rightarrow 95\% \text{ pre } \sigma^2: (0; 321,42)$$

$$\text{--} \sigma : (0; 17,928) \Rightarrow$$

balenie je horšie ako požaduje predpis

Využitie analytických nástrojov a vybraných štatistických funkcií v Exceli -
nástroje na analýzu údajov sú špeciálnym doplnkom programu Excel, ktorý je potrebné pred prvým použitím aktivovať.

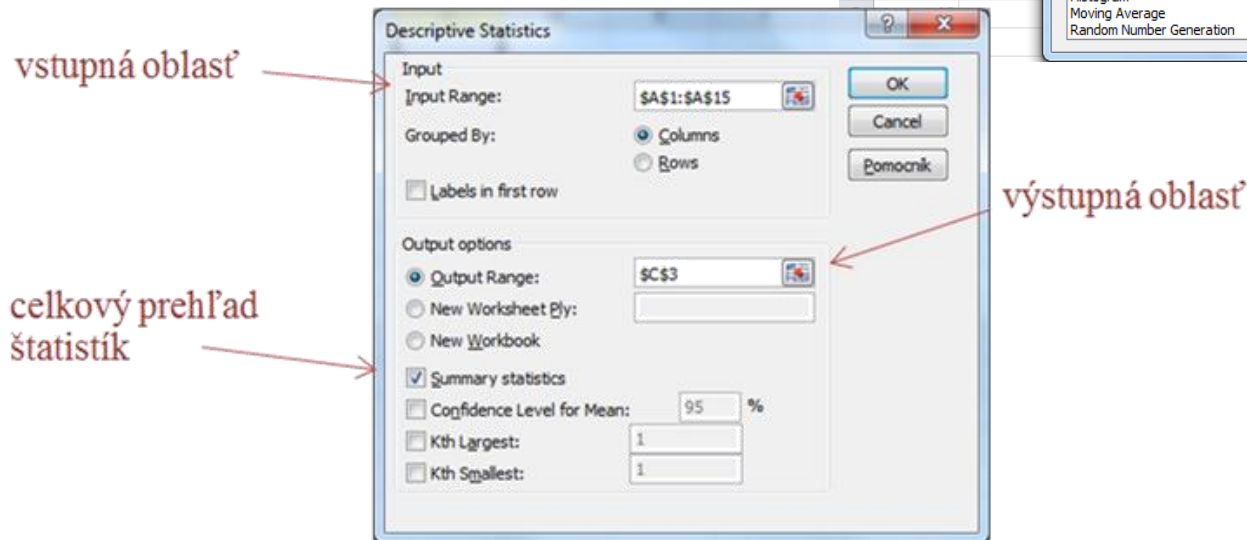
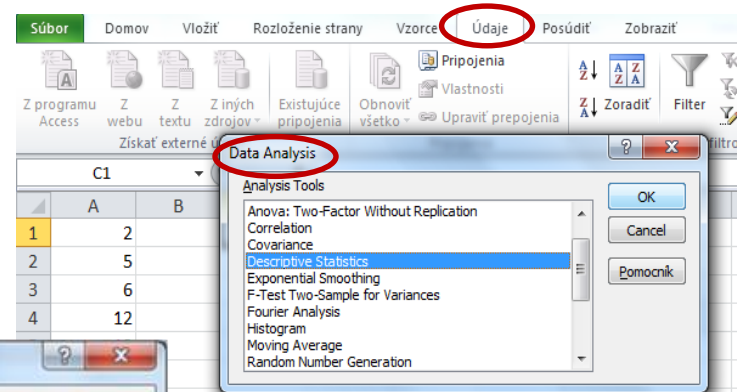


Príklad

Pomocou nástroja Opisná štatistika (Descriptive Statistics) urobte analýzu údajov výberového súboru VS: 2, 5, 6, 12, 12, 12, 20, 21, 22, 22, 23, 23, 25, 25, 26. Odhadnite parametre základného súboru ZS.

Riešenie.

Údaje → Analýza dát (Data → Data analysis)



Opisná štatistika v Exceli (Descriptive Statistics)

Výstupná tabuľka v Exceli obsahuje niekoľko parametrov popisujúcich VS a niekoľko výberových parametrov (náhodných premenných), ktoré slúžia na bodový odhad parametrov popisujúcich ZS:

	A	B	C	D
1	2		Príklad. Popisná štatistika	
2	5			
3	6		Column1	
4	12			
5	12		Mean	17,06667
6	12		Standard Error	2,094134
7	20		Median	21
8	21		Mode	12
9	22		Standard Deviation	8,110546
10	22		Sample Variance	65,78095
11	23		Kurtosis	-1,02712
12	23		Skewness	-0,69009
13	25		Range	24
14	25		Minimum	2
15	26		Maximum	26
16			Sum	256
17			Count	15

Charakteristiky (parametre) výberového súboru VS sú:

rozsah súboru $n = 15$ (Count),

súčet všetkých hodnôt štatistického znaku vo VS = 256(Sum),

Charakteristiky strednej hodnoty VS:

aritmetický priemer $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 17,06667$ (Mean),

medián $Me = 21$ (Median),

modus $Mo = 12$ (Mode).

Charakteristiky variability VS:

maximálna hodnota $x_{max} = 26$ (Maximum),

minimálna hodnota $x_{min} = 2$ (Minimum),

rozpätie $R = x_{max} - x_{min} = 24$ (Range),

rozptyl (tzv. **vychýlený rozptyl**) $s_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = 61,39556$ nie je súčasťou tabuľky, musíme ho dopočítať pomocou funkcie =VAR.P(číslo1;číslo2;...),

smerodajnú odchýlku $s_n = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = 7,835532$ vypočíta funkcia =STDEV.P(číslo1;číslo2;...).

Tvar rozdelenia popisujú (tzv. štandardizovaný) koeficient špicatosti $\beta = -1,02712$ (Kurtosis) a (tzv. štandardizovaný) koeficient šikmosti $\alpha = -0,69009$ (Skewness).

Bodové odhady parametrov základného súboru ZS

Bodový odhad strednej hodnoty M ZS – výberový priemer:

$$est M = \bar{x} = 2,094134.$$

Na jeho určenie môžeme použiť výsledné údaje z výstupu Opisnej štatistiky z Analýzy dát alebo použiť funkciu =AVERAGE(číslo1;číslo2;...).

Bodový odhad rozptylu σ^2 ZS – výberový rozptyl (nevychýlený rozptyl):

$$est \sigma^2 = s_{n-1}^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2 = 65,78095.$$

Jeho číselná hodnota je výstupom Opisnej štatistiky z Analýzy dát, resp. použijeme funkciu =VAR.S(číslo1;číslo2;...).

Medzi vychýleným a nevychýleným rozptylom platí vzťah $s_{n-1}^2 = \frac{n}{n-1} s_n^2$.

Bodový odhad smerodajnej odchýlky σ – výberová smerodajná odchýlka:

$$est \sigma = s_{n-1} = \sqrt{\frac{1}{n-1} \sum_i (x_i - \bar{x})^2} = 8,110546.$$

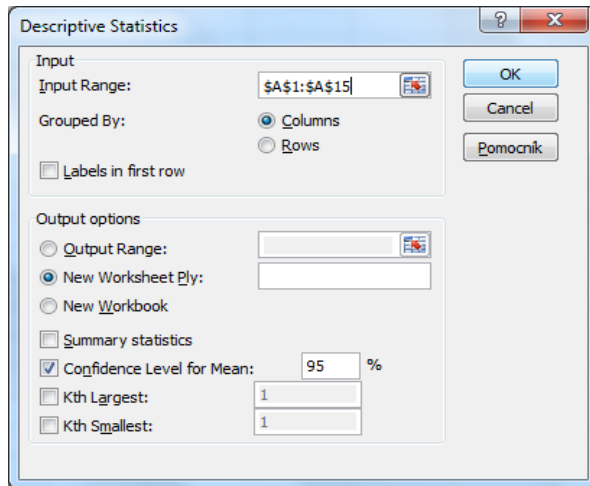
Jeho číselná hodnota je výstupom Opisnej štatistiky z Analýzy dát, resp. možno použiť funkciu =STDEV.S(číslo1;číslo2;...).

Intervalové odhady parametrov základného súboru ZS

Pre intervalový odhad strednej hodnoty M ZS platí

$$P(\bar{x} - \Delta < M < \bar{x} + \Delta) = 1 - \alpha.$$

Excel umožňuje priamy výpočet prípustnej chyby odchýlky Δ pomocou funkcie `=CONFIDENCE.T(alfa;smerodajná_odch;veľkosť)` alebo pomocou analytického nástroja Popisnej štatistiky a voľbou konfidenčného intervalu so zadáním požadovanej spoľahlivosti $1 - \alpha$ vyjadrenej v percentách,



$$\Delta = \text{CONFIDENCE.T}(0,05; D9; D17) = 4,491471$$

$P(17,06667 - 4,491471 < M < 17,06667 + 4,491471) = 1 - 0,05$, t.j. s 95%-nou spoľahlivosťou sa stred rozdelenia ZS nachádza v intervale (12,5752; 21,55814).

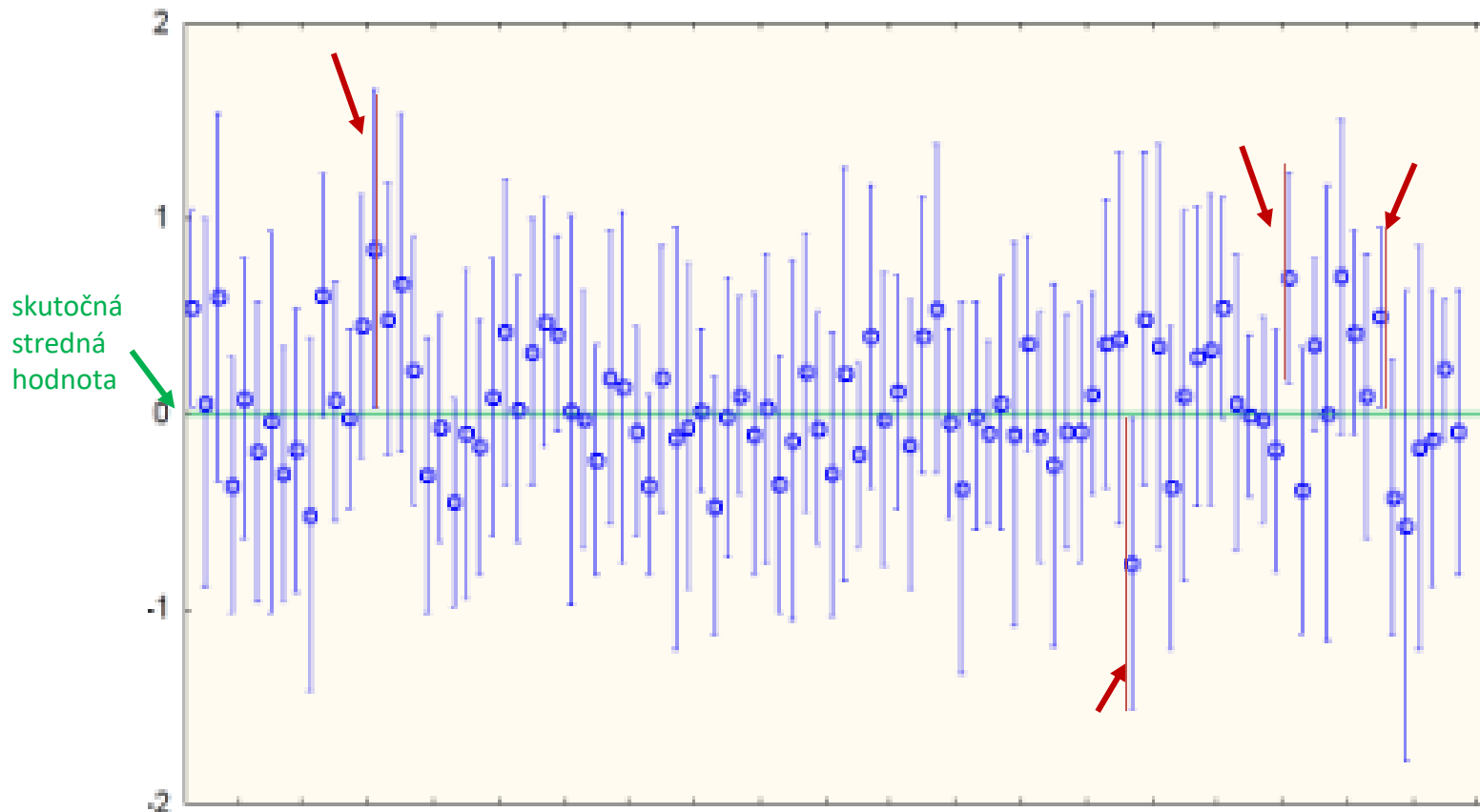
Čo znamená, že spoľahlivosť intervalového odhadu strednej hodnoty je 95%?

$$X \sim N(0,1)$$

Realizujeme 100 náhodných výberov z normovaného normálneho rozdelenia, každý s rozsahom 10 a pre každý výber urobíme intervalový odhad strednej hodnoty.

Každá čiara predstavuje jeden 95% interval spoľahlivosti pre jeden výber.

Vidíme, že približne 95% intervalov pokrýva skutočnú strednú hodnotu.



Intervalové odhady parametrov základného súboru ZS

Pre intervalový odhad rozptylu σ^2 ZS platí

$$P\left(\frac{(n-1)s_{n-1}^2}{\chi_{n-1}^{2,kv}\left(1-\frac{\alpha}{2}\right)} < \sigma^2 < \frac{(n-1)s_{n-1}^2}{\chi_{n-1}^{2,kv}\left(\frac{\alpha}{2}\right)}\right) = 1 - \alpha.$$

Hodnoty chí-kvadrát kvantilov sú

$$\chi_{n-1}^{2,kv}\left(1 - \frac{\alpha}{2}\right) = \text{CHISQ.INV}(0,975; 14) = 26,11895 \text{ a}$$

$$\chi_{n-1}^{2,kv}\left(\frac{\alpha}{2}\right) = \text{CHISQ.INV}(0,025; 14) = 5,628726.$$

$P(35,2592 < \sigma^2 < 163,6131) = 1 - 0,05$, t.j. s 95%-nou spoľahlivosťou sa rozptyl ZS nachádza v intervale (35,2592; 163,6131).

Intervalové odhady parametrov základného súboru ZS

Pre intervalový odhad smerodajnej odchýlky σ ZS platí:

$$P\left(\sqrt{\frac{(n-1)s_{n-1}^2}{\chi_{n-1}^{2,kv}\left(1-\frac{\alpha}{2}\right)}} < \sigma < \sqrt{\frac{(n-1)s_{n-1}^2}{\chi_{n-1}^{2,kv}\left(\frac{\alpha}{2}\right)}}\right) = 1 - \alpha, \text{ t.j. } P\left(\sqrt{35,2592} < \sigma < \right)$$

Induktívna štatistika: intervalové odhady vs. testovanie hypotéz

Intervalové odhady (confidence intervals)

- umožňujú odhadnúť neistotu v odhade parametra náhodnej premennej,
- používame vtedy, keď chceme u náhodnej premennej určiť veľkosť parametra, resp. veľkosť efektu (rozdielu, resp. pomeru parametrov dvoch náhodných premenných).

Testovanie hypotéz (hypothesis testing)

- umožňuje posúdiť, či experimentálne získané dáta nepopierajú predpoklad, ktorý sme urobili **pred** začiatkom testovania,
- používame, ak chceme overiť platnosť dopredu definovanej hypotézy (s dopredu určenou hladinou významnosti).

Rôzne druhy štatistických testov

Štatistické testy rozlišujeme:

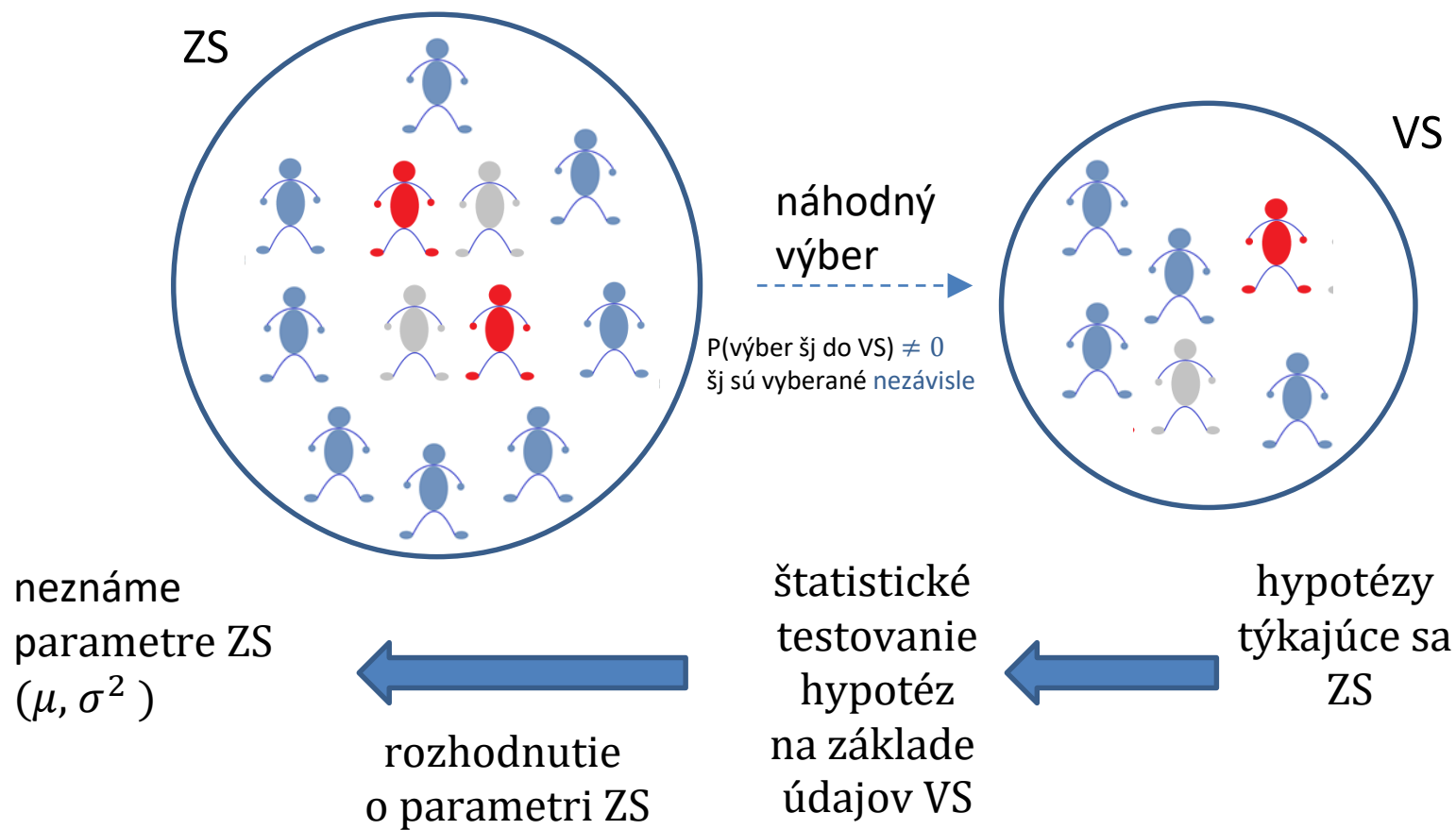
- obojstranné a jednostranné (**Two-tailed or one-tailed**)
- párové a pre jednoduché vzorky (**For paired or independent samples**)
- parametrické a neparametrické (**Parametric or nonparametric**)



TESTOVANIE HYPOTÉZ

- rozhodovací proces, v ktorom proti sebe stojí nulová a alternatívna hypotéza

Testovanie hypotéz



Štatistická hypotéza – predpoklad (tvrdenie) o parametroch ZS, o jeho rozdelení, prípadne o parametroch viacerých súborov.

Zdrojom štatistických hypotéz môžu byť predchádzajúce skúsenosti, teória, požiadavky na kvalitu produktu, dohady založené na pozorovaní...

Každý test hypotézy je zameraný na overenie určitej základnej (nulovej) hypotézy. Pri testovaní proti sebe vždy stoja nulová a alternatívna hypotéza.

$$H_0 \text{ } >< \text{ } H_1$$

Nulová hypotéza H_0 - tvrdenie, že efekt je nulový, resp. že neexistuje závislosť, že dáta majú určitý typ rozdelenia, ...

Alternatívna hypotéza $H_A(H_1)$ - tvrdenie popierajúce nulovú hypotézu (obvykle to, čo chceme dokázať)

Hypotéza H_0 vyjadruje predpoklad o rovnosti parametra Q ZS s konštantou Q_0 :

$$H_0: Q = Q_0 \text{ resp. } H_0: Q - Q_0 = 0$$

Alternatívne tvrdenie, ktoré prijímame v prípade zamietnutia H_0 :

obojsstranná alternatíva $H_1: Q \neq Q_0$

jednostranná $H_1: Q > Q_0$ (pravostranná)

$H_1: Q < Q_0$ (ľavostranná)

Každý test štatistickej hypotézy je zameraný na overenie platnosti nulovej hypotézy H_0 . Možné závery testu štatistickej hypotézy:

- H_0 zamietame
- H_0 nezamietame (prijímame)

Neexistuje matematický postup, ktorým by sme dokázali platnosť hypotézy H_0 !!!

Testovanie hypotéz



Súd musí rozhodnúť či:
Obžalovaný je nevinný
alebo
Obžalovaný je vinný

Súd nepozná skutočnosť
absolútne presne

		SKUTOČNOSŤ	
		NEVINNÝ	VINNÝ
ROZHODNUTIE SÚDU	OSLOBODI Ť	správne	Chyba II. druhu (β)
	ODSÚDI Ť	Chyba I. druhu (α)	správne

Obžalovaný zostáva nevinným,
pokiaľ súd nemá evidentné dôkazy o
jeho vine.

PRINCÍP PREZUMPCIE NEVINY

POSTUP SÚDU:

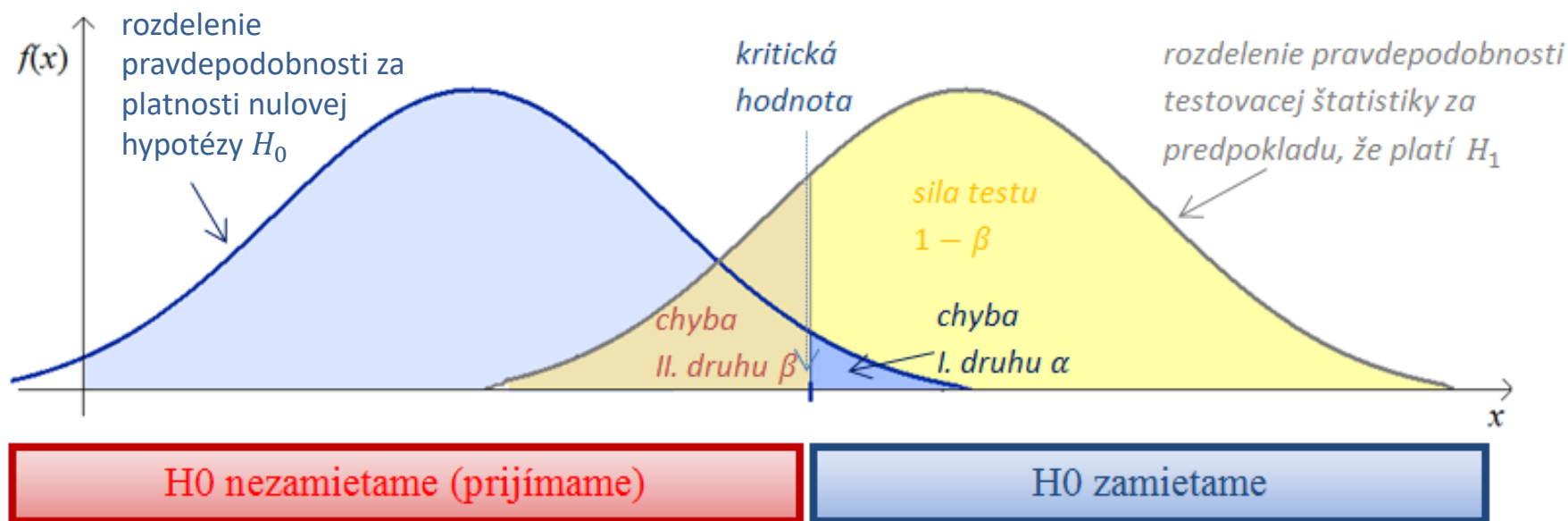
1. Vychádzať z prezumpcie neviny
2. Zvoliť prijateľnú úroveň rizika omylu, že odsúdi nevinného (chyba I. druhu)
3. Nájsť správny zákon a interpretovať ho
4. Porovnať dôkazy so zákonom
5. Ak dôkazy postačujú odsúdiť, ak nie oslobodiť

Ak súd neodsúdi obžalovaného to nemusí znamenať, že je nevinný. Súd môže mať iba málo dôkazov(chyba II.druhu)

Pri testovaní $H_0 >< H_1$ robíme úsudky na základe údajov z VS.

Môžeme sa dopustiť chýb:

- **chyby I. druhu** – zamietneme H_0 a v skutočnosti H_0 je pravdivá (zamietame málo pravdepodobné prípady, ktoré môžu nastať aj pri platnosti H_0)
- **chyby II. druhu** – **nezamietneme** H_0 aj keď v skutočnosti H_0 **ne**platí (aj keď sme blízko „optimálnej“ hodnote pre platnosť H_0 , môže platiť H_1)



Rozhodnutia o prijatí resp. zamietnutí H_0 pri testovaní hypotéz

		SKUTOČNOSŤ	
		H_0 je pravdivá	H_0 je nepravdivá
ROZHODNUTIE O H_0	H_0 nezamietame (prijímame)	správne rozhodnutie (spoľahlivosť testu) pravdepodobnosť: $1-\alpha$	nesprávne rozhodnutie (chyba II. druhu) pravdepodobnosť: β
	H_0 zamietame (neprijímame)	nesprávne rozhodnutie (chyba I. druhu) pravdepodobnosť: α	správne rozhodnutie (sila testu) pravdepodobnosť: $1-\beta$

H_0 : You are not pregnant.

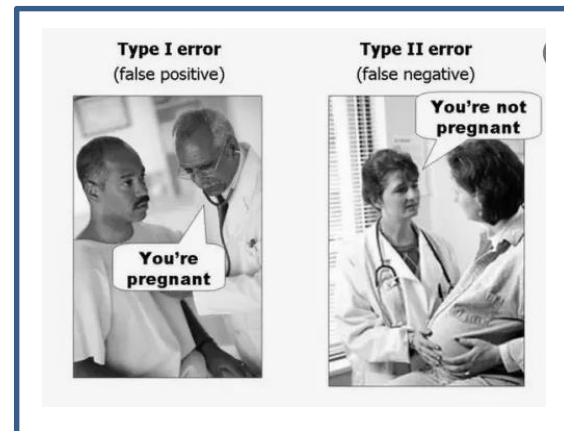
H_1 : You are pregnant.

false positive (FP):

α = P(zamietnutie H_0 / H_0 je pravdivá)

false negative (FN):

β = P(nezamietnutie H_0 / H_0 je nepravdivá)



Sila testu $1-\beta$ (Power W)

Sila testu $1-\beta$ je pravdepodobnosť zamietnutia H_0 , keď H_0 neplatí (pravdepodobnosť odhalenia neplatnosti H_0).

Doporučená hodnota $1-\beta$: 0,7-0,8 (čím vyššia, tým lepšie)

Silu testu ovplyvňujú:

- typ zvoleného testu: test je silnejší, keď pri rovnakej α má väčšie W ;
- veľkosť VS (*sample size*): väčší súbor (viac info) $\Rightarrow$ väčšie $1-\beta \rightarrow$ Na začiatku volíme taký rozsah VS, aby bolo zaručené minimálne $1-\beta$.
- hladina významnosti α : menšie α (ťažšie zamietnutie H_0) $\Rightarrow$ menšie $1-\beta$, väčšie $\alpha \Rightarrow$ väčšie $1-\beta$ (vyššie riziko FP výsledku);
- veľkosť pozorovaného rozdielu (*effect size*): ľahšie identifikujeme „veľký“ rozdiel (výška muž-žena vs. Bratislavčan-Trnavčan);
- variabilita sledovanej premennej: väčší rozptyl dát $\Rightarrow$ ťažší odhad sledovaného parametra (potrebných viac dát) $\Rightarrow$ ťažšie rozhodnutie o H_0 , menší rozptyl dát $\Rightarrow$ väčšie $1-\beta$;

Sila testu $1-\beta$ je dôležitá vtedy, ak H_0 nezamietame. Ak je $1-\beta$ malá, je oprávnené zamietnutie H_0 málo pravdepodobné a teda je malá pravdepodobnosť prijatia H_1 .

Všeobecný postup pri testovaní hypotéz pomocou kritickej oblasti

1. Sformulujeme nulovú a alternatívnu hypotézu.
2. Zvolíme hladinu významnosti α - maximálnu pravdepodobnosť chyby I. druhu
3. Vyberieme vhodnú testovaciu štatistiku a vypočítame jej hodnotu na základe hodnôt VS.
4. O prijatí, resp. zamietnutí nulovej hypotézy možno rozhodnúť pomocou:
 - **kritickej oblasti** – určíme kritické hodnoty, ktoré vymedzujú oblasť zamietnutia, resp. prijatia nulovej hypotézy na základe toho, či hodnota testovacej štatistiky leží alebo neleží v danej kritickej oblasti;
 - **p-hodnoty** – porovnávame so zvolenou α ;
 - **$(1 - \alpha) \cdot 100\%$ IS** – nepotrebuje test. štatistiku.
5. Formulujeme záver o prijatí, resp. zamietnutí nulovej hypotézy na zvolenej hladine významnosti α .

Parameter ZS Q nepoznáme, iba ho odhadujeme na základe výberovej charakteristiky u_n .

Výber testovacej štatistiky T (testovacieho kritéria), t. j. funkciu

$T = f(Q, u_n)$ ovplyvňujú:

- rozdelenie sledovaného znaku (náhodnej premennej) X v ZS;
- rozsah VS.

Testovacia štatistika je náhodná premenná, ktorá môže nadobúdať hodnoty z oblasti prijatia hypotézy H_0 a z oblasti zamietnutia hypotézy H_0 - tzv. **kritickej oblasti** W . Obe množiny sú disjunktné – hranice, ktoré ich oddeľujú nazývame **kritické hodnoty**.

Vypočítanú hodnotu testovacej štatistiky pre daný VS nazývame **hodnotou (realizáciou) výberovej štatistiky** – označujeme t_0, F_0, χ_0^2

Testovanie pomocou

1. kritického oboru W – oboru zamietnutia H_0

$W = \left(t_{min}, Kvantil \left(\frac{\alpha}{2} \right) \right) \cup \left(Kvantil \left(1 - \frac{\alpha}{2} \right), t_{max} \right)$, kde použitá testovacia štatistika je z oboru hodnôt $\langle t_{min}, t_{max} \rangle$

2. p hodnoty:

ak $p < \alpha$, H_0 zamietame,

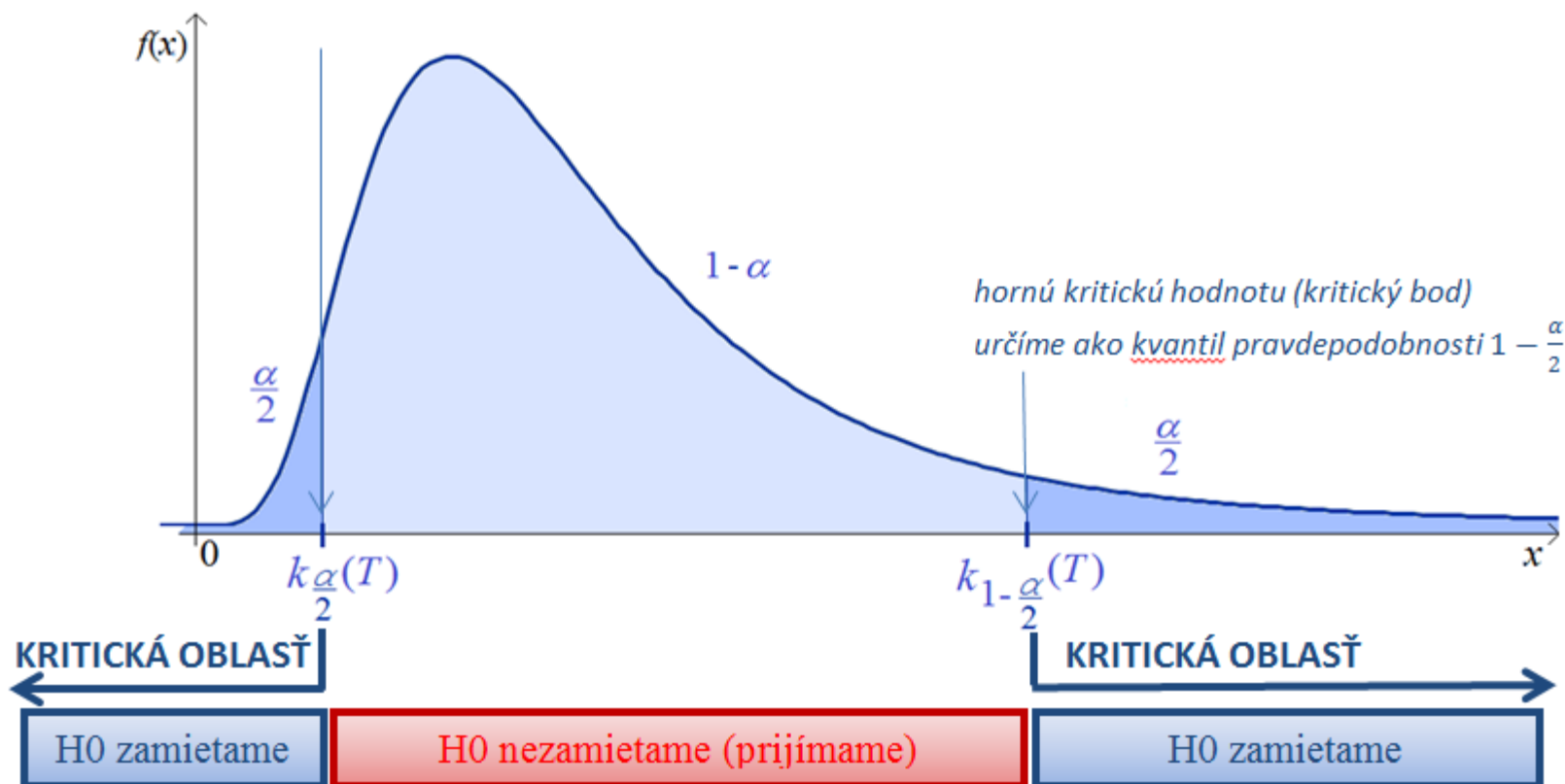
$p \geq \alpha$, H_0 nezamietame (prijímame).

$p = 2 \cdot \min\{F(t_0); 1 - F(t_0)\}$, kde t_0 je vypočítaná hodnota test. štatistiky

3. $(1-\alpha)\%$ intervalu spoľahlivosti IS:

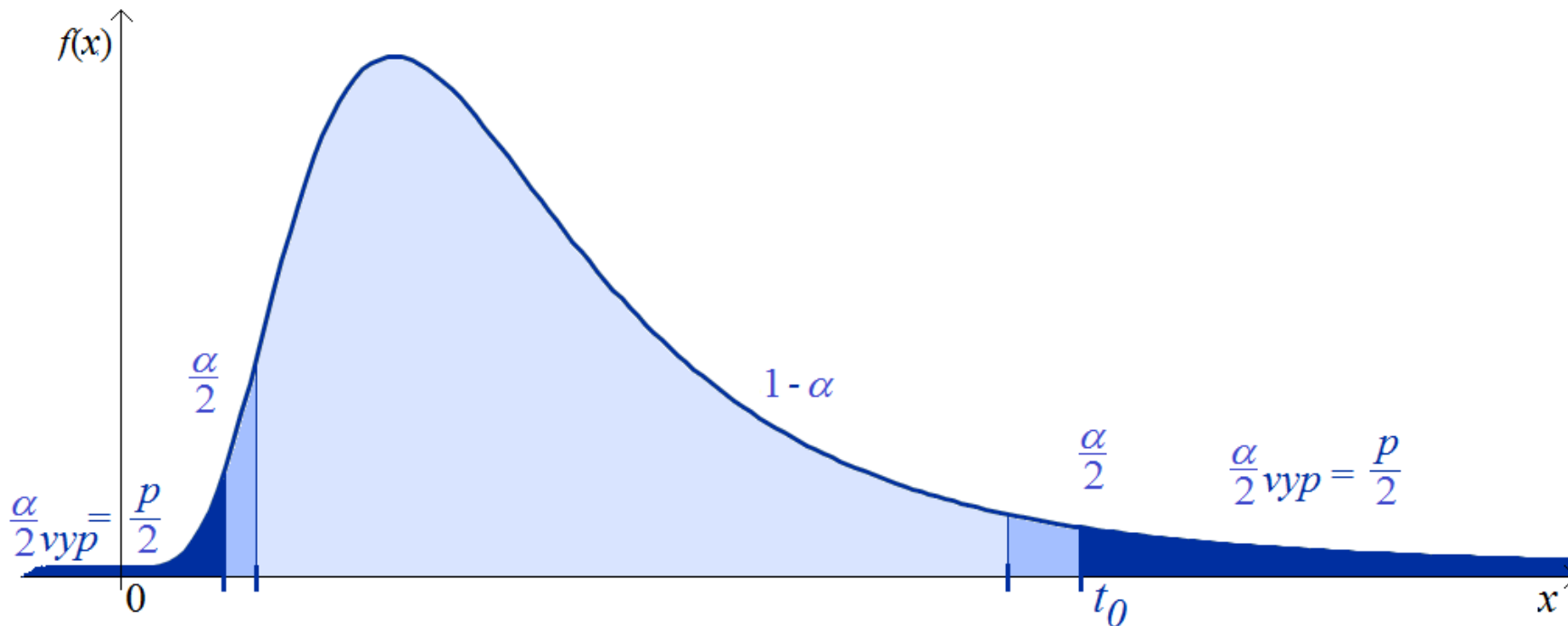
ak konštanta $Q_0 \notin IS$ ($\mu_0 \notin IS$ resp. $\sigma_0^2 \notin IS$), H_0 zamietame

dolnú kritickú hodnotu (kritický bod) určíme ako kvantil pravdepodobnosti $\frac{\alpha}{2}$

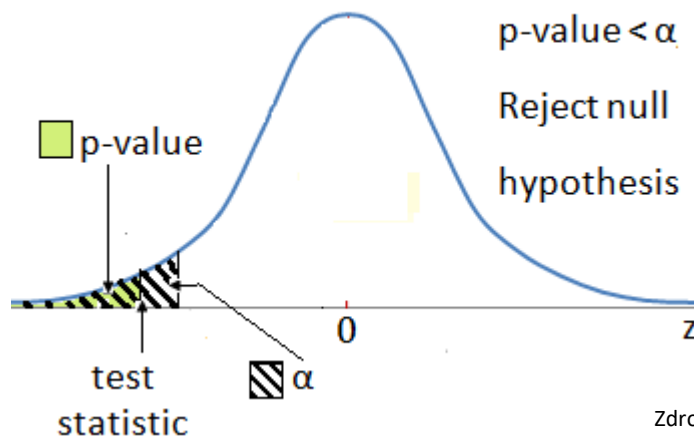
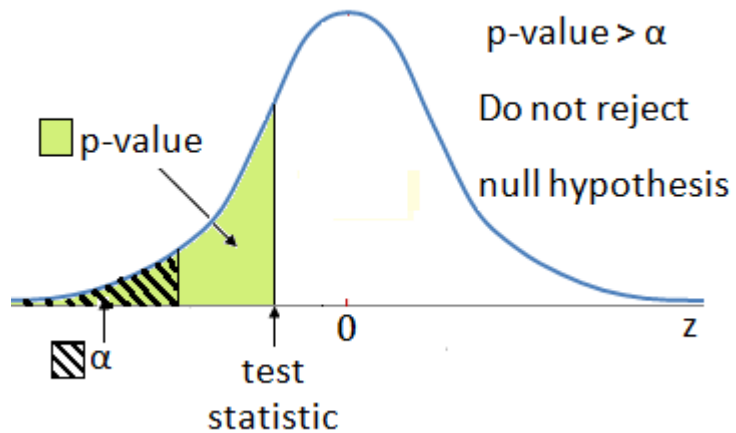


P hodnota (*p value, Signifikance*) je najnižšia hladina na zamietnutie hypotézy H_0 . Je to pravdepodobnosť, že pri platnosti H_0 by sme vzhľadom k jednostrannosti alebo obojstrannosti testu získali rovnakú alebo ešte extrémnejšiu (menej pravdepodobnú) hodnotu testovacej štatistiky.

Čím je p-hodnota menšia, tým viac sme presvedčení, že nulová hypotéza nie je správna a mala by byť zamietnutá.



Testovanie pomocou p hodnoty (jednostranné alternatívy)



$p < 0.05$



Zdroj: <https://mhisymmohammad.files.wordpress.com/2012/02/fred-mercury.jpg>

Jednovýberové testy o parametroch ZS s rozdelením $N(\mu, \sigma^2)$

Test hypotézy o zhode strednej hodnoty μ so známou konštantou μ_0

$$H_0: \mu = \mu_0 \quad >< \quad H_1: \mu \neq \mu_0$$

testovacia štatistika

- ak poznáme rozptyl σ^2 ZS: $z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$
- nepoznáme rozptyl σ^2 ZS: $t = \frac{\bar{x} - \mu_0}{\frac{s_{n-1}}{\sqrt{n}}} \sim t(n-1)$

Test hypotézy o zhode rozptylu σ^2 so známou konštantou σ_0^2

$$H_0: \sigma^2 = \sigma_0^2 \quad >< \quad H_1: \sigma^2 \neq \sigma_0^2$$

testovacia štatistika

$$\chi^2 = \frac{(n-1)s_{n-1}^2}{\sigma_0^2} \sim \chi^2(n-1)$$

Pre jednostranné alternatívne hypotézy platí

ak $H_1: \mu < \mu_0$ (resp. $H_1: \sigma^2 < \sigma_0^2$)

$$W = (t_{min}; Kvantil(\alpha))$$

$$p = F(t_0)$$

ak $H_1: \mu > \mu_0$ (resp. $H_1: \sigma^2 > \sigma_0^2$)

$$W = (Kvantil(1 - \alpha); t_{max})$$

$$p = 1 - F(t_0)$$

Dvojvýberové testy

Test rovnosti rozptylov dvoch normálnych rozdelení

$$H_0: \sigma_1^2 = \sigma_2^2 >< H_1: \sigma_1^2 \neq \sigma_2^2 \text{ (resp. } H_1: \sigma_1^2 > \sigma_2^2)$$

testovacia štatistika

$$F = \frac{s_{n_1-1}^2}{s_{n_2-1}^2} \sim F(n_1-1, n_2-1)$$

Poznámka:

homoskedasticita (homogenita rozptylov)– rozptyly v ZS sú zhodné

heteroskedaticita (nehomogenita rozptylov)– rozptyly v ZS nie sú zhodné

Dvojvýberové testy

Test zhody dvoch stredných hodnôt za predpokladu homoskedasticity

$$H_0: \mu_1 = \mu_2 \quad >< \quad H_1: \mu_1 \neq \mu_2$$

testovacia štatistika

$$t = \frac{\bar{x} - \bar{y}}{s} \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}} \sim t(n_1 + n_2 - 2), \text{ kde } s^2 = \frac{(n_1 - 1)s_{n_1-1}^2 + (n_2 - 1)s_{n_2-1}^2}{n_1 + n_2 - 2}$$

Test zhody dvoch stredných hodnôt za predpokladu heteroskedasticity

$$H_0: \mu_1 = \mu_2 \quad >< \quad H_1: \mu_1 \neq \mu_2$$

testovacia štatistika

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_{n_1-1}^2}{n_1} + \frac{s_{n_2-1}^2}{n_2}}} \sim t(v), \text{ kde } v \approx \frac{\left(\frac{s_{n_1-1}^2}{n_1} + \frac{s_{n_2-1}^2}{n_2}\right)^2}{\frac{1}{n_1-1} \left(\frac{s_{n_1-1}^2}{n_1}\right)^2 + \frac{1}{n_2-1} \left(\frac{s_{n_2-1}^2}{n_2}\right)^2} \text{ je zaokrúhlené}$$

dolu na najbližšie celé číslo

Dvojvýberové testy

Test zhody dvoch stredných hodnôt pre závislé výbery (párový test)

$$H_0: \mu = 0 \quad >< \quad H_1: \mu \neq 0$$

testovacia štatistika

$$t = \frac{\bar{D} - 0}{s_{n_D - 1}} \sqrt{n} \sim t(n - 1)$$

Poznámka:

Párový test je jednovýberovým testom aplikovaným na diferencie μ .

Príklad 1

Zo ZS s normálnym rozdelením s parametrami μ a $\sigma^2 = 0,04$ bolo náhodne vybraných 10 hodnôt: 2; 1,8; 2,1; 2,4; 1,9; 2,1; 2; 1,8; 2,3; 2,2.

Na hladine významnosti $\alpha = 0,05$ otestujte, či je stredná hodnota $\mu = 1,95$.

→ ZS s rozdelením $N(\mu; \sigma^2 = 0,04)$

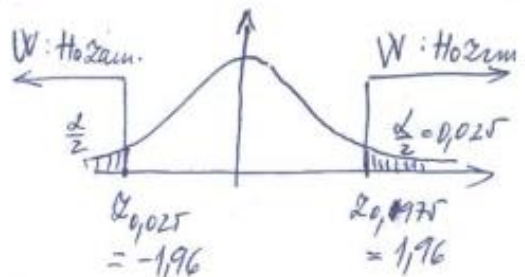
VS: 2; 1,8; 2,1; 2,4; 1,9; 2,1; 2; 1,8; 2,3; 2,2

$\alpha = 0,05$; $\mu = 1,95$?

$H_0: \mu = \mu_0 = 1,95$ $\neq$ $H_1: \mu \neq 1,95$

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

$$Z_0 = \frac{2,06 - 1,95}{\frac{0,2}{\sqrt{10}}} = 1,74$$



$$\bar{x} = 2,06$$

$$n = 10$$

$$Z_{0,025} = \text{NORM.S.INV}(0,025)$$

$$Z_{0,975} = \text{NORM.S.INV}(0,975)$$

a) pomocou kritického oboru: W

$$W = (-\infty; -1,96) \cup (1,96; \infty)$$

$\alpha_0 = 1,74 \notin W \Rightarrow H_0: \mu = 1,95$ nezamietame (prijmame)

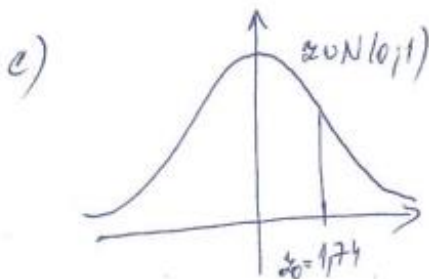
$H_1: \mu \neq 1,95$ zamietame

b) 95% IS pre μ : $(\bar{x} - \Delta; \bar{x} + \Delta) = 2,06 \pm 0,123959 \rightarrow (1,936; 2,184)$

$$\Delta = \text{CONFIDENCE.NORM}(0,05; 0,2; 10) = 0,123959$$

$1,95 \in (1,936; 2,184) \Rightarrow H_0: \mu = 1,95$ nezamietame

$H_1: \mu \neq 1,95$ zamietame



$$p = 2 \cdot \min \left\{ \underset{0,959}{F(1,74)}; \underset{0,04}{1 - F(1,74)} \right\} = 0,08$$

$p = 0,08 > 0,05 = \alpha \Rightarrow H_0: \mu = 1,95$ nezamietame

$H_1: \mu \neq 1,95$ zamietame

Ke blízke významnosti $\alpha = 0,05$ sa nám nepodarilo zamietnuť hypotézu, ne sledujeme hodnotu μ rovnú 1,95.

Príklad 2

Meraním dĺžky 10 valčekov boli získané hodnoty

5,37; 5,36; 5,35; 5,40; 5,41; 5,34; 5,29; 5,43; 5,42; 5,32.

Na hladine významnosti $\alpha = 0,05$ otestujte hypotézu, že priemerná dĺžka valčekov sa neodchyľuje od normy 5,4 mm. Predpokladáme, že sledovaná náhodná premenná (namerané dáta) pochádza z normálneho rozdelenia.

→ 10 valčekov

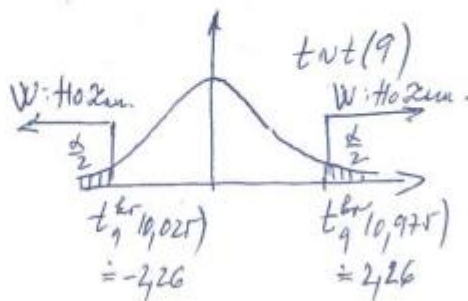
5,37; 5,36; 5,35; 5,4; 5,41; 5,34; 5,29; 5,43; 5,42; 5,32

$\alpha = 0,05$; $\mu_0 = 5,4$ mm

$H_0: \mu = 5,4$ vs $H_1: \mu \neq 5,4$

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \sim t(n-1)$$

$$t_0 = \frac{5,364 - 5,4}{\frac{0,056804}{\sqrt{10}}} \approx -2,00413$$



$n = 10$

$\bar{x} = 5,364$

$t_{\alpha/2}^{cr}(9, 0,025) = T.INV(0,975; 9)$

$A_{n-1} = 0,056804$

a) pomocou kritických oblastí W

$$W = (-\infty, -2,26) \cup (2,26, \infty)$$

$$t_0 = -2,00413 \notin W \Rightarrow H_0: \mu = 5,4 \text{ rokov}$$

$$H_1: \mu \neq 5,4 \text{ rokov}$$

b) 95% IS pre μ : $(\bar{x} - \Delta; \bar{x} + \Delta) = 5,364 \pm 0,040635$

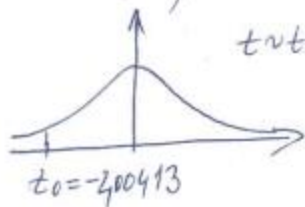
$$\Delta = \text{CONFIDENCE.T}(0,05; 9,056804; 10) = 0,040635$$

$$\Delta = \frac{s_{n-1}}{\sqrt{n}} \cdot t_{n-1}^{1-\frac{\alpha}{2}}$$

$$\mu_0 = 5,4 \in 95\% \text{ IS pre } \mu = (5,32; 5,4046...) \Rightarrow H_0: \mu = 5,4 \text{ rokov}$$

$$H_1: \mu \neq 5,4 \text{ rokov}$$

c) pomocou p-hodnoty



$$p = 2 \cdot \min \left\{ F(-2,00413); 1 - F(-2,00413) \right\} = 0,076$$
$$= 0,038 \quad = 0,96$$

$$p = 0,076 > 0,05 = \alpha \Rightarrow H_0: \mu = 5,4 \text{ rokov}$$

$$H_1: \mu \neq 5,4 \text{ rokov}$$

Na hladine významnosti $\alpha = 0,05$ nemôžeme zamietnuť hypotézu, že priemerná dĺžka života sa neodchyľuje od normy.

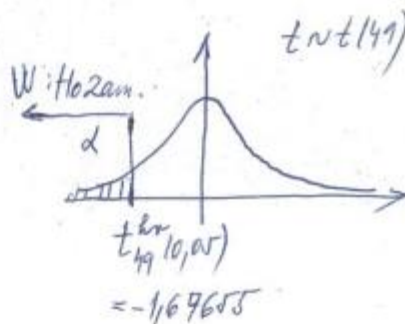
Príklad 3

Výrobca tvrdí, že ním vyrábané žiarovky majú v priemere životnosť 1000 hodín. Potrebujete sa uistiť, že ich životnosť v priemer nie je nižšia. Na základe údajov o životnosti 50 náhodne vybraných žiaroviek sa pokúste na hladine významnosti $\alpha = 0,05$ dokázať, že to tak je. Výberové charakteristiky sú $\bar{x} = 997,080$; $s_{49} = 104,709$.

→ $n = 50$; $\bar{x} = 997,080$; $s_{49} = 104,709$; $\alpha = 0,05$
overíme, či životnosť žiaroviek je väčšia ako 1000 hodín
 $H_0: \mu = 1000 \quad H_1: \mu < 1000$

test. stat. $t = \frac{\bar{x} - \mu_0}{\frac{s_{n-1}}{\sqrt{n}}} \sim t(n-1)$

$$t_0 = \frac{997,080 - 1000}{\frac{104,709}{\sqrt{50}}} = -0,197$$



a) $W = (-\infty; -1,67655)$

$t_0 = -0,197 \notin W \Rightarrow H_0: \mu = 1000$ nezamietame (priji'mame)

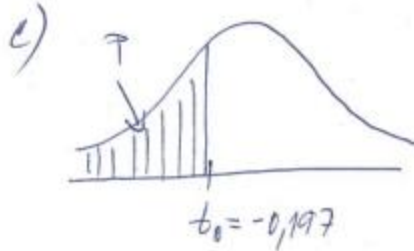
$H_1: \mu < 1000$ zamietame

t) $P(-\infty < \mu < \bar{x} + \frac{s-1}{\sqrt{n}} t_{n-1}^{kr}(1-\alpha)) = 0,95$

jednosmerný 95% IS pre μ : $(-\infty; 997,080 + \frac{104,709}{\sqrt{50}} \cdot t_{49}^{kr}(0,95))$

$\mu_0 = 1000 \in (-\infty; 1021,97) \Rightarrow H_0: \mu = 1000$ nezamietame

$H_1: \mu < 1000$ zamietame



$P = P(t_0) = T.DIST(-0,197; 49; 1) = 0,422 > 0,05 = \alpha$

$H_0: \mu = 1000$ nezamietame

$H_1: \mu < 1000$ zamietame

Az hľadíme významnosť $\alpha = 0,05$ az nám reprodukovalo zamietnutie H_0 ,
t.j. az spoľahlivosť 95% má aj mŕané hodnoty, az primerne!
Množstvo je menšie ako 1000 hodín.

Príklad 4

Plniaca linka je nastavená správne, ak plní minerálku do fliaš na priemernú hodnotu 7 dcl s priemernou odchýlkou menšou ako 0,3 dcl. Na hladine $\alpha = 0,05$ otestujte, či je linka nastavená správne, ak sme náhodne vybrali 10 naplnených fliaš s minerálkou, v ktorých bolo nasledujúce množstvo vody v dcl:

7,1 7,0 6,9 6,8 7,2 7,1 7,1 6,9 7,0 7,1.

→ VS: 7,1, 7, 6,9, 6,8, 7,2, 7,1, 7,1, 6,9, 7, 7,1

$\alpha = 0,05$; $\mu_0 = 7$ dcl? ; $\sigma < 0,3$?

$H_0: \mu = 7$ vs $H_1: \mu \neq 7$

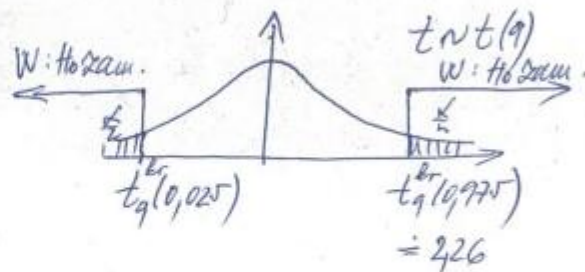
$$t = \frac{\bar{X} - \mu_0}{\frac{s_{n-1}}{\sqrt{n}}} \sim t(n-1)$$

$$t_0 = \frac{7,02 - 7}{0,122927} \cdot \sqrt{10} = 0,514496$$

$$\bar{X} = 7,02$$

$$s_0 = 0,122927$$

$$n = 10$$



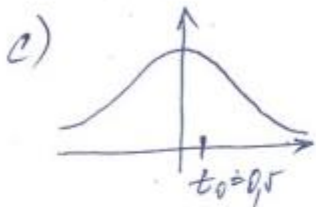
$$a) W = (-\infty; -2,26) \cup (2,26; \infty)$$

$t_0 = 0,514496 \notin W \Rightarrow H_0: \mu = 7$ nezamietame (prijmame)
 $H_1: \mu \neq 7$ zamietame

b) 95% IS für μ : $(7,02 - 0,087937; 7,02 + 0,087937) = (6,932; 7,1079)$

$7 \in (6,932; 7,108) \Rightarrow H_0: \mu = 7$ rezanidisme

H_7 : $\rho_n \neq 7\%$ Kümüleme



$$p = 2 \cdot \min \{ F(t_0); 1 - F(t_0) \} = 2 \cdot \min \{ 0,69; 0,30965 \} = 0,62 > 0,05 = \alpha$$

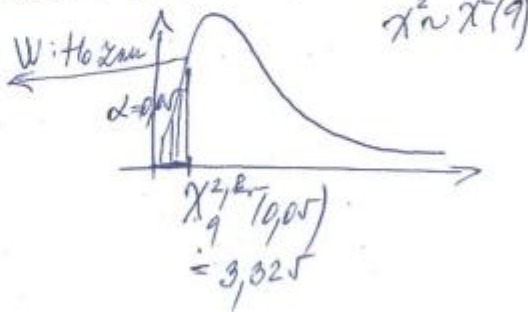
$\Rightarrow H_0$ hexamidine, H_1 kanamidine

$t_0 = 95$
 Bolr sled. prechádzaní, že pľmáca lírka pľmí minútllku sa predpľozný objem
 7dll. Tlra možná konfídencíal' sa spľlchlllroslá 95%. $\chi^2 \sim \chi^2(9)$

$$H_0: \sigma = 0,3 \quad \text{vs} \quad H_1: \sigma < 0,3$$

$$\chi^2 = \frac{(n-1) \cdot s_{n-1}^2}{\sigma_0^2} \sim \chi^2(n-1)$$

$$\chi^2 = \frac{9.0122927^2}{0.3^2} = 9.51$$



$$a) W = (0, 3, 325)$$

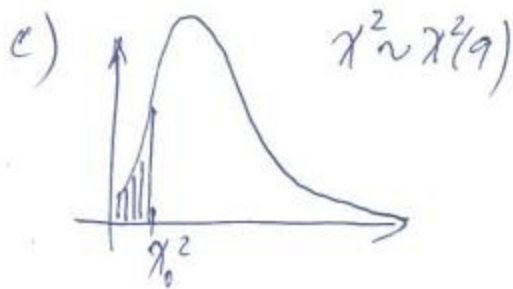
$\chi_0^2 = 1,51 \in (0, 3,325) = W \Rightarrow H_0: \tilde{\sigma} = 0,3$ samide me
 $H_1: \tilde{\sigma} < 0,3$ resamide me (pripišame)

$$d) 95\% \text{ IS pre } \tilde{\sigma}^2: \left(0; \frac{(n-1) \cdot s_{n-1}^2}{\chi_{n-1}^2, \alpha} \right) \Rightarrow (0; 0,040901)$$

$$95\% \text{ IS pre } \tilde{\sigma}: (0; 0,2) \quad \begin{array}{|c|c|c|} \hline 0 & \text{-----} & 0,2 \\ \hline & & 0,3 \end{array}$$

$$0,3 \notin (0; 0,2) \Rightarrow H_0: \tilde{\sigma} \leq 0,3 \text{ zamietame}$$

$$H_1: \tilde{\sigma} < 0,3 \text{ nezamietame}$$



$$P = F(\chi_0^2) = \text{CHISQ.DIST}(1,51; 9; 1) = 0,002937 < 0,05 = \alpha \Rightarrow$$

$$H_0: \tilde{\sigma} = 0,3 \text{ zamietame}$$

$$H_1: \tilde{\sigma} < 0,3 \text{ nezamietame}$$

Na hladine významnosti $\alpha = 0,5$ máme rozhodnúť, či plocha línky
je či fláče s priemerom odčítanou menšiu ako 0,3 dol.

Induktívna štatistika: intervalové odhady vs. testovanie hypotéz

Intervalové odhady (confidence intervals)

- umožňujú odhadnúť neistotu v odhade parametra náhodnej premennej,
- používame vtedy, keď chceme u náhodnej premennej určiť veľkosť parametra, resp. veľkosť efektu (rozdielu, resp. pomeru parametrov dvoch náhodných premenných).

Testovanie hypotéz (hypothesis testing)

- umožňuje posúdiť, či experimentálne získané dáta nepopierajú predpoklad, ktorý sme urobili **pred** začiatkom testovania,
- používame, ak chceme overiť platnosť dopredu definovanej hypotézy (s dopredu určenou hladinou významnosti).

Metódy overenia normálneho rozdelenia NP

Testy normality

Testy založené na momentoch:

- Test nulovej šikmosti a nulovej špicatosti NP
- Kombinovaný test - Jarque–Bera test (JB test)
- D'Agostinov test šikmosti
- Obdĺžnikový test šikmosti a špicatosti....
- Chi kvadrát test dobrej zhody

Testy založené na empirických distribučných funkciách:

- Kolmogorov Smirnov test
- Lillieforsov test
- Andersonov-Darlingov test

Orientačná kontrola normality dát:

aritmetický priemer s mediánom by sa nemali líšiť o viac ako 10%

Test nulovej šikmosti a nulovej špicatosti rozdelenia NP

sa skladá z dvoch častí:

1. $H_0: \alpha(X) = 0 >< H_1: \alpha(X) \neq 0$

Testovacia štatistika $u_3 = \frac{\alpha(X)}{\sqrt{D(\alpha(X))}} \sim N(0,1)$, kde $D(\alpha(X)) = \frac{6(n-2)}{(n+1)(n+3)}$

Kritická oblasť $W = \left\{ u_3; |u_3| \geq u^{kv} \left(1 - \frac{\alpha}{2} \right) \right\}$

Použitie: $n > 200$

2. $H_0: \beta(X) = 0 >< H_1: \beta(X) \neq 0$

Testovacia štatistika $u_4 = \frac{\beta(X) + \frac{6}{n+1}}{\sqrt{D(\beta(X))}} \sim N(0,1)$, kde $D(\beta(X)) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$

Kritická oblasť $W = \left\{ u_4; |u_4| \geq u^{kv} \left(1 - \frac{\alpha}{2} \right) \right\}$

Použitie: $n > 500$

Hypotézu, že VS pochádza z normálneho rozdelenia nezamietame, ak nezamietame súčasne hypotézu o nulovej šikmosti i nulovej špicatosti.

Ak zamietame aspoň jednu z oboch H_0 , potom s rovnakou spoľahlivosťou zamietame tvrdenie, že výber pochádza z normálneho rozdelenia.

Kombinovaný C-test nulovej šikmosti a nulovej špicatosti NP

Test je založený na myšlienke: súčet k štvorcov nezávislých normovaných normálnych NP má Pearsonovo χ^2 rozdelenie s k stupňami voľnosti.

H_0 : NP má normálne rozdelenie. $>< H_1$: NP nemá normálne rozdelenie.

Testovacia štatistika $C = u_3^2 + u_4^2 \sim \chi^2(2)$

Kritická oblasť $W = \{C; C \geq \chi_2^{2, kv}(1 - \alpha)\}$

Použitie: $n > 200$

χ^2 test dobrej zhody

Univerzálny test k overeniu toho, či náhodný výber pochádza z nejakého diskrétného alebo spojitého rozdelenia.

Východisková myšlienka: porovnávame relatívne početnosti $f_i = \frac{n_i}{n}$ s hypotetickými pravdepodobnosťami $\hat{\pi}_i$

H_0 : NP má rozdelenie daného typu. $>< H_1$: NP nemá rozdelenie daného typu.

Hodnoty VS triedime do k tried alebo disjunktných intervalov, pričom $n_i = 1, \dots, k$ sú k nim prislúchajúce absolútne početnosti. Odhad hypotetickej pravdepodobnosti $\hat{\pi}_i$, že NP nadobudne hodnoty z i -tej triedy počítame z podmienky, že NP má predpokladané rozdelenie.

Testovacia štatistika $\chi^2 = \sum_{i=1}^k \frac{(n_i - n\hat{\pi}_i)^2}{n\hat{\pi}_i} \sim \chi^2(k - c - 1)$, kde c je počet odhadovaných parametrov overovaného rozdelenia

Kritická oblasť $W = \{\chi^2; \chi^2 \geq \chi_v^{2, kv}(1 - \alpha)\}$

Použitie: veľké n

V praxi sa vyžaduje, aby vo všetkých triedach boli teoretické početnosti väčšie ako 5. Pri nesplnení sa triedy resp. intervaly zlučujú.

Kolmogorov – Smirnov test (test dobrej zhody)

Tento test je možné použiť aj pre výbery malého rozsahu, t.j. ak χ^2 nie je vhodný. Potrebná je znalosť všetkých parametrov uvažovaného spojitého rozdelenia.

H_0 : Náhodný výber pochádza z určitého rozdelenia so spojitou distribučnou funkciou $F(x)$. $>< H_1$: Náhodný výber nepochádza z daného rozdelenia.

Testovacia štatistika $D_n = \sup |F_n(x) - F(x)|$

Kritická oblasť $W = \{\chi^2; D_n \geq D_n(1 - \alpha)\}$

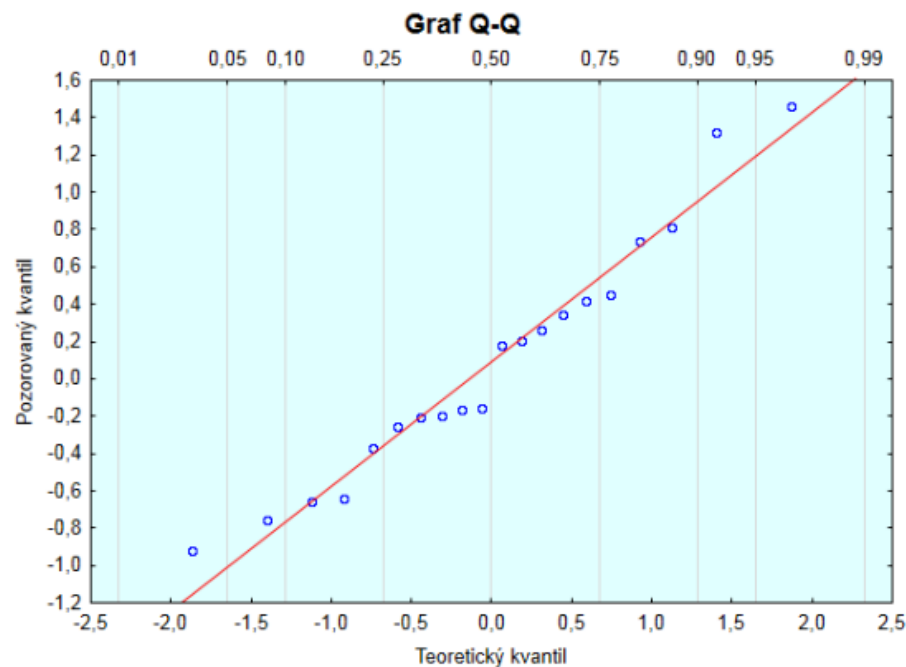
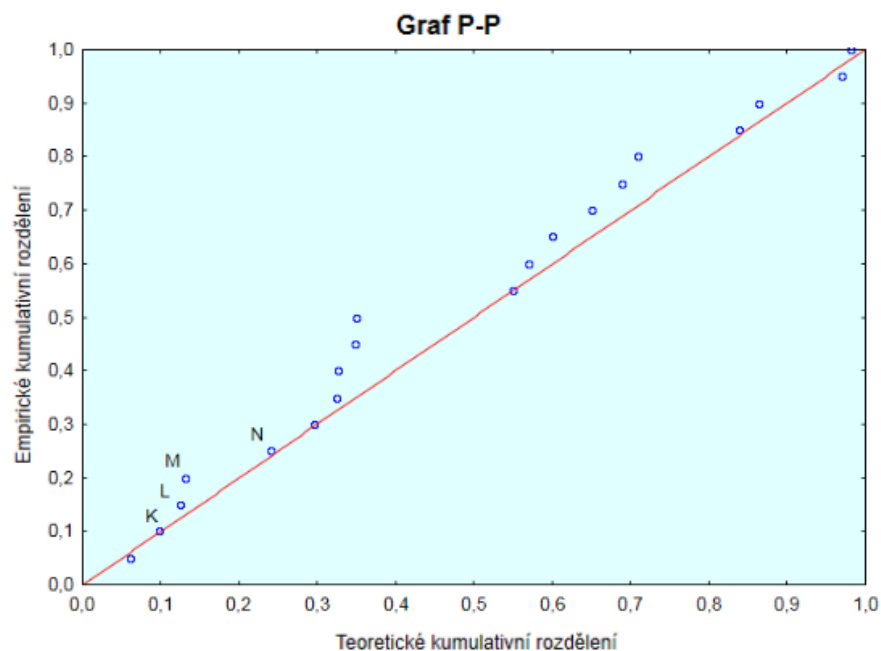
Použitie: malé n

$D_n(1 - \alpha) \cong \sqrt{\frac{1}{2n} \ln \frac{2}{\alpha}}$ pre veľké n , inak tabuľkové hodnoty

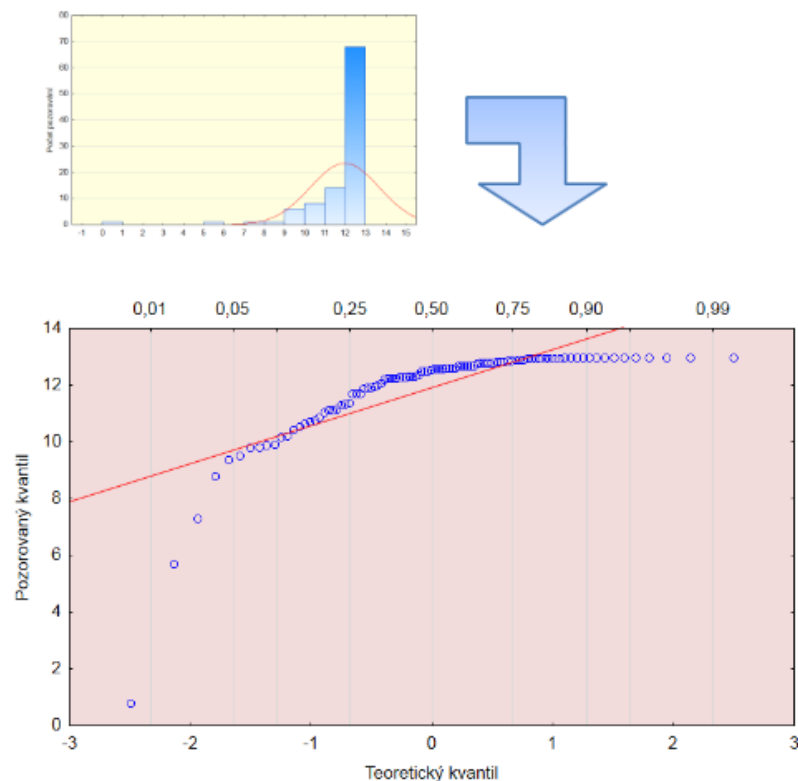
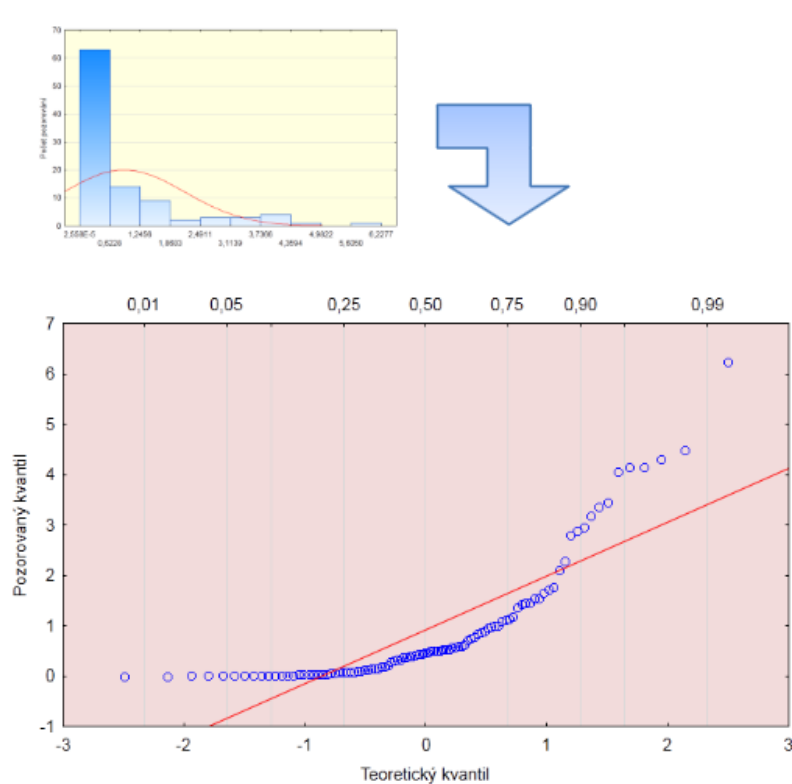
Metódy overenia normálneho rozdelenia NP

Grafické metódy

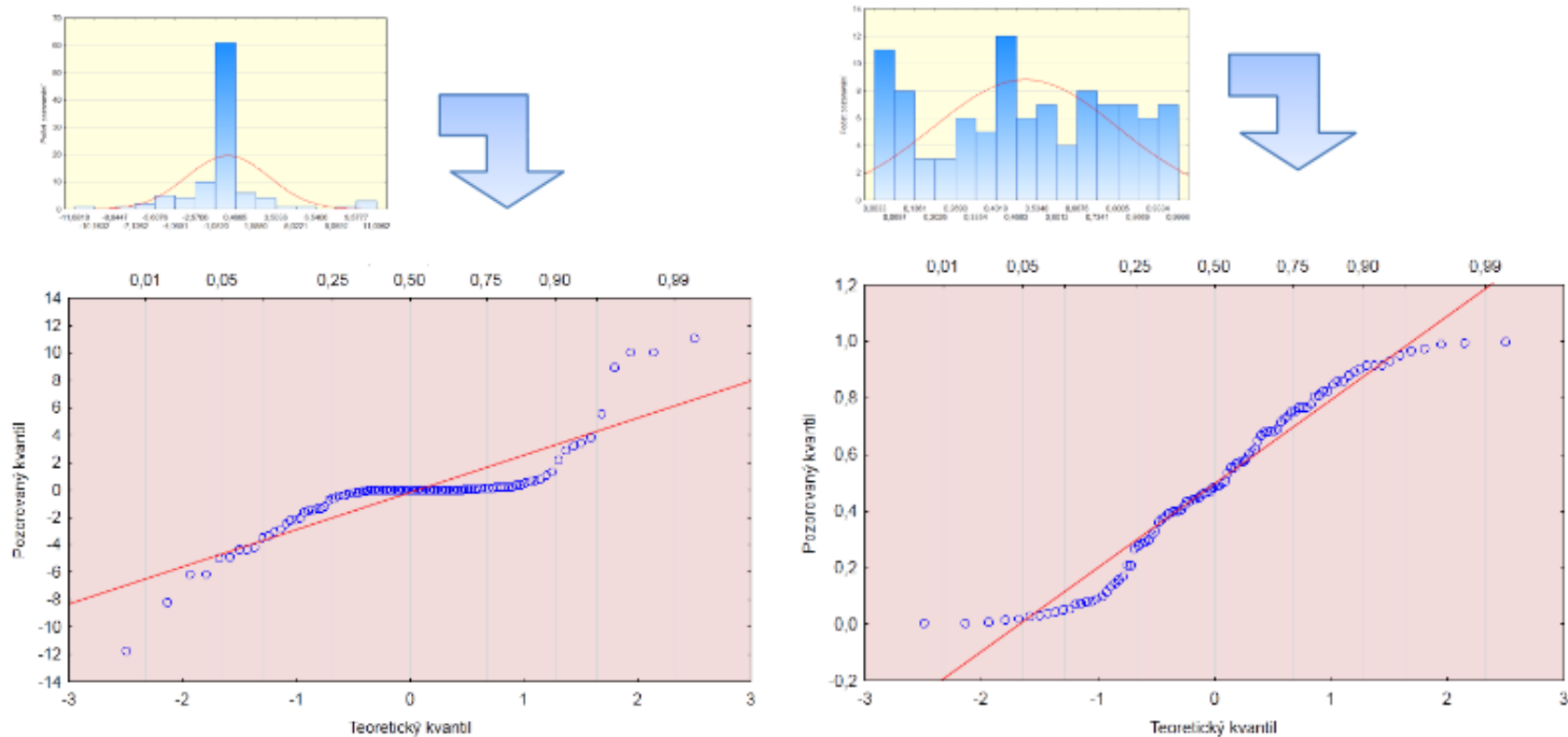
- Histogram
- Box plot
- P-P graf
- Q-Q graf



„parabola“ indukuje nesymetriu (zošikmenie) dát



Tvar „S“ indukuje nenulovú špicatost'



http://www.statsoft.cz/file1/PDF/newsletter/2013_10_09_StatSoft_Jak_se_pozna_normalita_pomoci_grafu.pdf