STATISTIKA

451102 Veronika Hlaváčková
446928 Radka Vaňušaníková
460533 Tamara Juriňáková
451410 Mária Bugajová
461123 Michaela Pešková

Priemer vs medián

- Priemer - súčet všetkých hodnôt vydelený ich počtom

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

- Medián - prostredná pozorovaná hodnota, ktorá delí pozorované hodnoty na dve polovice (na hodnoty menšie a hodnoty väčšie ako medián)

$$\tilde{x} = x_{((n+1)/2)} \qquad \text{pro } n \text{ liché}$$

$$\tilde{x} = \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right) \qquad \text{pro } n \text{ sudé}$$

Rozptyl dát

- vyjadruje variabilitu
- aritmetický priemer štvorcov (druhých mocnín) odchýlok od aritmetického priemeru
- veľmi ovplyvniteľný odľahlými hodnotami

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

- odmocnina z rozptylu= **smerodajná odchýlka (SMODCH)**
- uvádza sa v rovnakých jednotkách ako pozorované dáta
  - *pre n<10* $s_R = k_n . R$
  - *pre n>10* $s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$

Stredná (štandardná) chyba priemeru

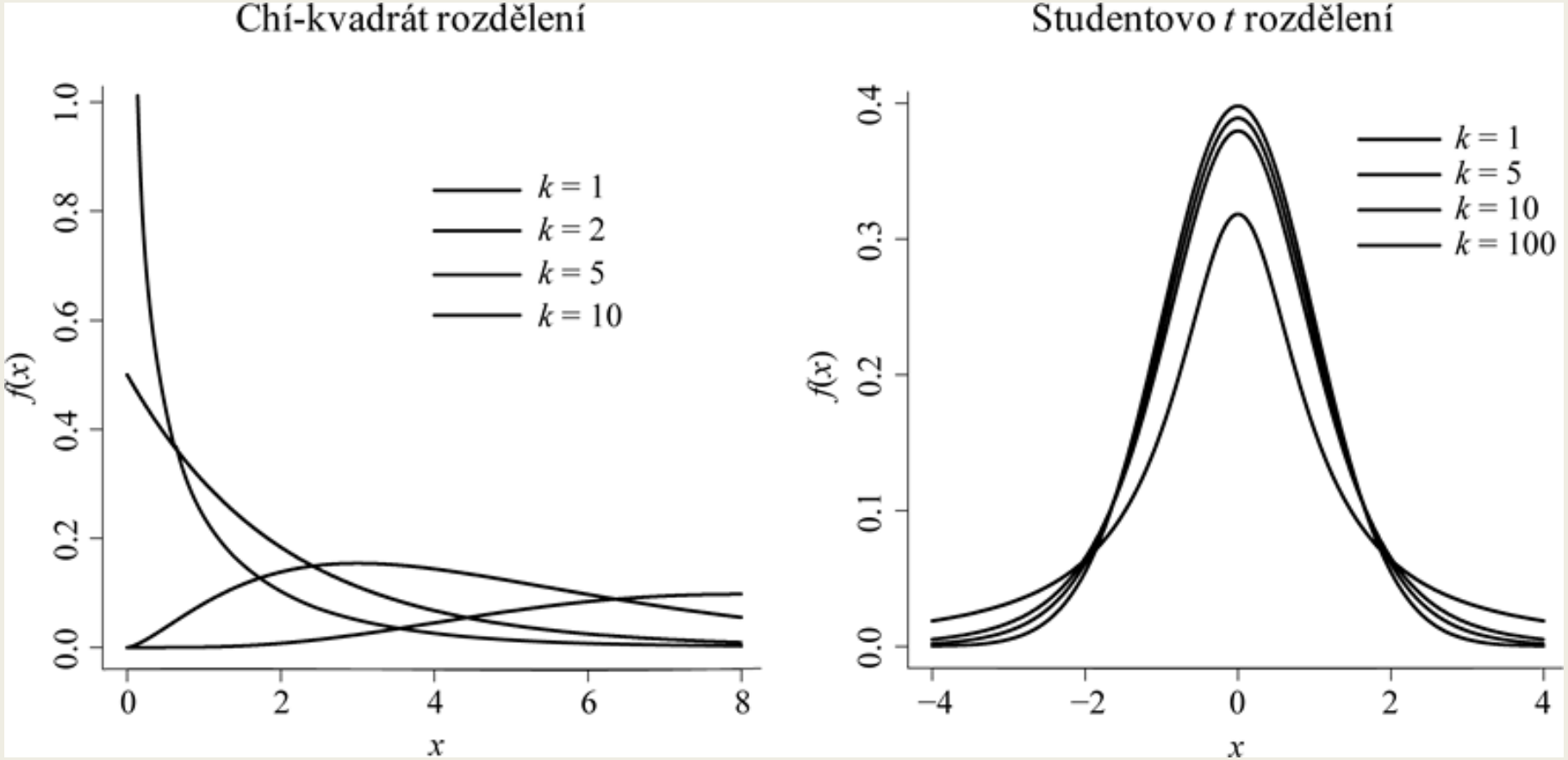
- meria rozptýlenosť vypočítaného aritmetického priemeru v rôznych výberových súboroch z jedného veľkého základného súboru
- je mierou presnosti s akou aritmetický priemer odhaduje skutočnú strednú hodnotu základného súboru

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

Modelové rozloženia dát

Rozloženie	Parametre	Stručný popis
Studentovo	-stupne voľnosti- uvažuje veľkosť vzorky -priemer -rozptyl	-simuluje normálne rozloženie pre menšie vzorky čísel -pre väčšie súbory (n>100) sa limitne blíži k normálnemu rozloženiu
Pearsonovo (Chi-kvadrát)	-stupne voľnosti- uvažuje veľkosť vzorky	-slúži predovšetkým k porovnaniu početnosti javov v dvoch a viac kategóriách -používa sa k modelovaniu rozloženia odhadu rozptylu normálne rozložených dát
Fisher-Snedecorovo	-dvojité stupne voľnosti- uvažuje veľkosť dvoch vzoriek	-používa sa k testovaniu hodnôt priemerov-F test pre porovnanie dvoch výberových rozptylov; F test, ANOVA atď.

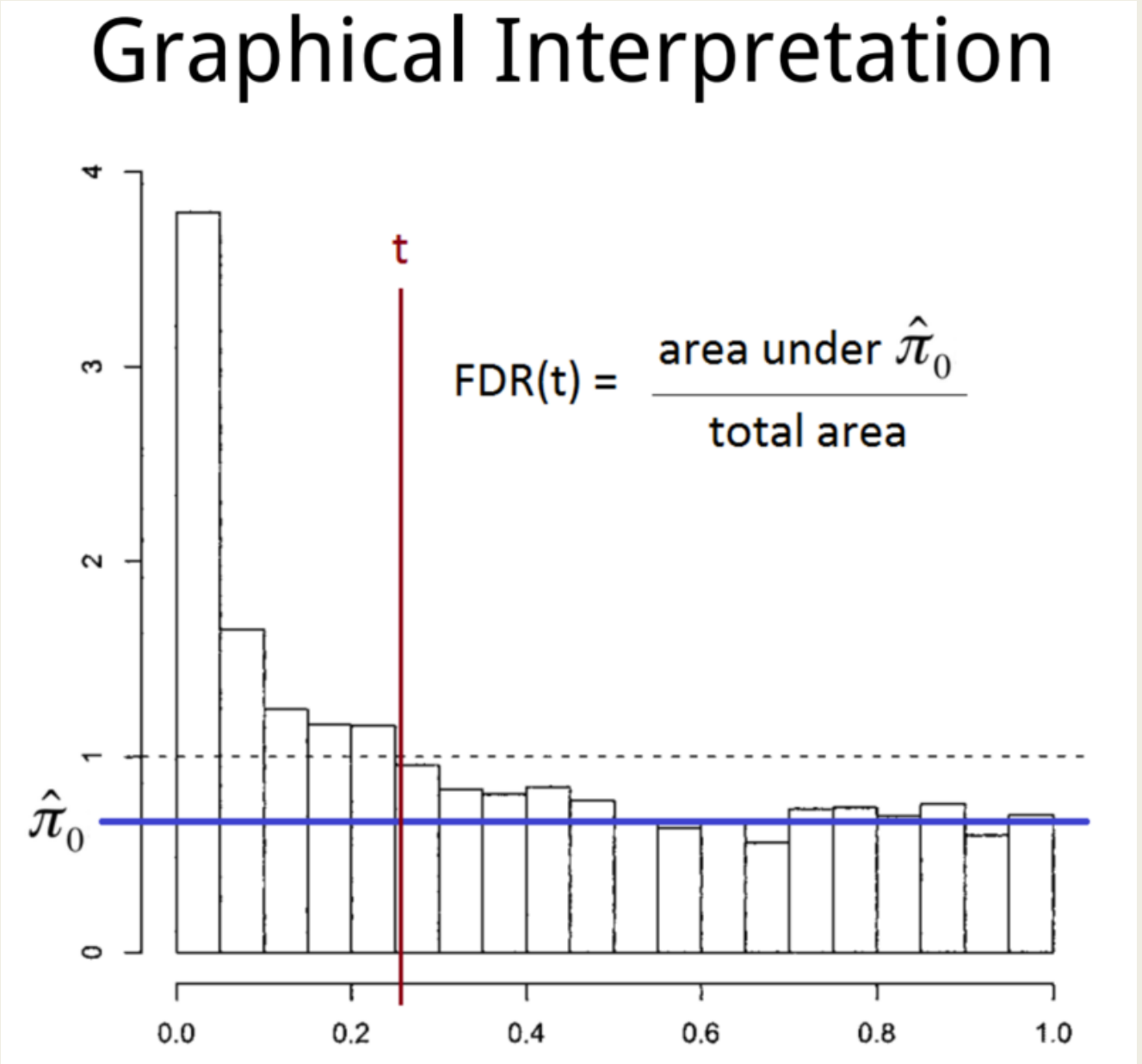
Modelové rozloženie dát



Ukážky hustôt náhodných veličín s chi-kvadrát rozdelením a studentovým t rozdelením.

Q-hodnota

- štatistická metóda na vylúčenie falošne pozitívnych výsledkov (FDR= False Discovery Rate)
- FDR
  - *stanovuje upravené P - hodnoty pre jednotlivé testy*
  - *kontroluje počet falošne pozitívnych výsledkov v týchto testoch*



E - očakávaná hodnota, t - prahová hodnota, π_0 - podiel prvkov, ktoré sú skutočne nulové, π_0 nezávisí na t

z-skóre

- Převod hrubého skóre na standardizovanou stupnici
- z-transformace: lineární transformace, která posunuje a rovnoměrně mění měřítko, zatímco nedeformuje vzdálenost mezi hodnotami
- z-skóre je rovno 0, SMODCH je rovna 1

$$z = \frac{(x - \mu)}{\sigma}$$

x – naměřená hodnota, μ – průměr,
σ - SMODCH

Správný výběr statistického testu na základě:

- typu dat
- homogenity rozptylu porovnávaných skupin
- normality dat
- typu hypotézy

- parametrické testy (symetrická data: tělesná výška, krevní tlak)
- neparametrické testy (asymetrická data: relativní exprese biomolekul)
- jednovýběrové testy (porovnání množiny dat s referenční hodnotou, která platí pro celou populaci - počet leukocytů u pacientu a průměrný počet leukocytů)
- dvouvýběrové testy (srovnání dvou nezávislých skupin - pacienti a kontrola, sledování určitého znaku u mužů, žen)
- párové testy (srovnání dvou skupin na sobě závislých dat)

T-test

- Testovanie štatistických hypotéz
- Sú rozdiely medzi skupinami len náhodné?
- Inferenčná štatistika
  - *generalizácia*
  - *oddelenie náhody od zákonitostí*

Predpoklady

- Normálna distribúcia dát
- Reprezentačná vzorka z populácie
- Primeraný počet vzoriek (20-30 vzoriek)
- Homogenita rozptylu

T-hodnota

	A	B	C	D
1		**Skupina 1**	**Skupina 2**	
2		15,2	15,9	
3		15,3	15,9	
4		16	15,2	
5		15,8	16,6	
6		15,6	15,2	
7		14,9	15,8	
8		15	15,8	
9		15,4	16,2	
10		15,6	15,6	
11		15,7	15,6	
12		15,5	15,8	
13		15,2	15,5	
14		15,5	15,5	
15		15,1	15,5	
16		15,3	14,9	
17		15	15,9	
18	**priemer**	15,38125	15,68125	
19	**sm.odchýlka**	0,31245	0,406971	
20	**odchýlka**	0,097625	0,165625	
21	**počet**	16	16	
22				
23				

$$t = \frac{signál}{šum} = \text{odchýlka medzi skupinami/odchýlka v rámci skupín}$$

priemer

$$t = \frac{|X1 - X2|}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

odchýlka

počet vzoriek

t=2,3

T-test = (B2:B17;C2:C17;2;2) = 0,026198

T - test

Degrees of freedom	Significance level					
	20% (0.20)	10% (0.10)	5% (0.05)	2% (0.02)	1% (0.01)	0.1% (0.001)
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.043	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.158	2.617	3.373
∝	1.282	1.645	1.960	2.326	2.576	3.291

df (stupne voľnosti)
$= (n_1\text{-}1) + (n_2\text{-}1) = 30$

t-hodnota > kritická hodnota

↓

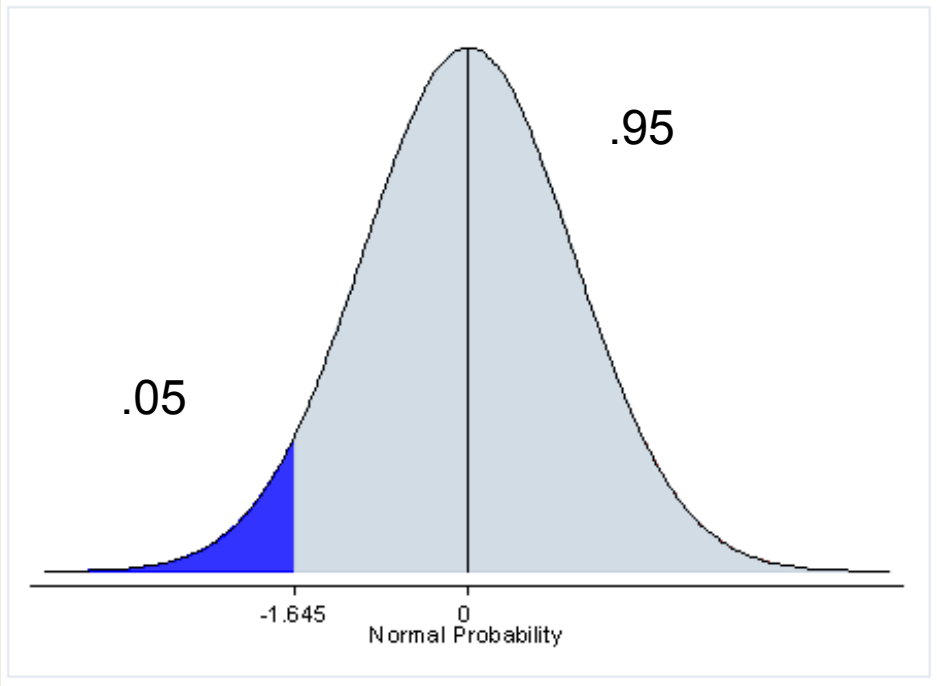
p-hodnota

P - hodnota

- Každá t-hodnota má p-hodnotu
- Pravdepodobnosť, že vzor údajov by mohol byť vytvorený náhodnými údajmi

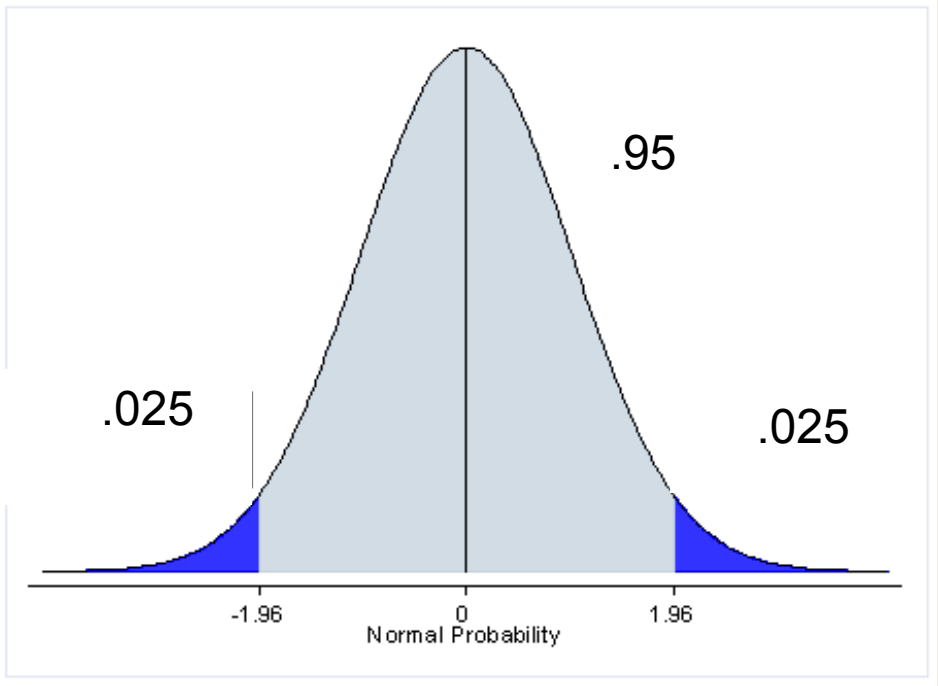
$p < 0,05$	< 5% pravdepodobnosť, že dáta sú len náhodné
$p > 0,05$	> 5% pravdepodobnosť, že dáta sú len náhodné

Typy t-testov



Jednovýberový

sú chlapci výrazne vyšší ako dievčatá?



Dvojvýberový

je výrazný rozdiel medzi výškou dievčat a výškov chlapcov?

Typy t-testov

	Výška	
	Chlapci	Dievčatá
	72	71
	68	70
	74	68
	69	68
	63	67
	68	65
	71	65
	69	64
	62	62
	75	62
priemer	69,1	66,2
sm.odchýlka	4,011234	2,95973

Nepárový

žiadne spojenie

	Výška chlapcov	
	Vek 5	Vek 15
	54	73
	55	65
	56	71
	48	65
	39	63
	45	70
	48	65
	43	68
	51	63
	54	71
priemer	49,3	67,4
sm.odchýlka	5,404628	3,46987

Párový

rovnaká skupina je opakovane meraná

TESTOVÁNÍ ODLEHLÝCH HODNOT

Dean Dixonův Q-test

$$Q_1 = \frac{(x_2 - x_1)}{R}$$

$$Q_n = \frac{(x_n - x_{n-1})}{R}$$

- Vyloučení hrubých chyb
- Vhodný pro soubor s malým počtem paralelních stanovení (pro n < 7)
- Místo směrodatné odchylky používá rozpětí
- $R = x_{max} - x_{min}$
- když Qn < Q k pak výsledek není odlehlý a zůstane součástí souboru dat
- když Qn > Q k pak výsledek je odlehlý a výsledek se vyloučí ze souboru dat

Příklad Q-test

$$Q_n = \frac{(x_n - x_{n-1})}{R}$$

Obsah křemičitanu ve vzorku [%]	
1	50,10
2	52,44
3	52,91
4	53,82
5	53,89
6	54,03

Výpočty	
rozptyl	3,93
Q_1	0,59542
Q_6	0,03562

Grubsův test T-test

$$T_1 = \frac{\left(\bar{X} - x_1\right)}{s}$$

$$T_n = \frac{\left(x_n - \bar{X}\right)}{s}$$

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

- pro n > 7
- Parametrický test
- Používáme parametry souboru: průměr a směrodatnou odchylku
- o kolik směrodatných odchylek se liší extrémní hodnota od průměru
- když Tn < Tk pak výsledek není odlehlý a zůstane součástí souboru dat
- když Tn > Tk pak výsledek je odlehlý a výsledek se vyloučí ze souboru dat

Příklad t-test

Titrační stanovení manganu [%]	
1	37,41
2	37,76
3	37,84
4	37,88
5	37,90
6	37,92
7	38,01
8	38,23
9	38,33
10	38,42
11	38,59
12	39,90

Výpočty	
průměr	38,18
sm.odch	0,60

t-test	
T1	1,282
T12	2,850

test odlehlosti výsledků – tabelované hodnoty Q_k a T_k

	T_k		Q_k	
n	α=0,05	α=0,01	α=0,05	α=0,01
3	1,412	1,416	0,941	0,988
4	1,689	1,723	0,765	0,889
5	1,869	1,955	0,642	0,760
6	1,996	2,130	0,560	0,698
7	2,093	2,265	0,507	0,637
8	2,172	2,374	0,468	0,590
9	2,237	2,464	0,437	0,555
10	2,294	2,540	0,412	0,527
11	2,343	2,606		
12	2,387	2,663		

ANOVA

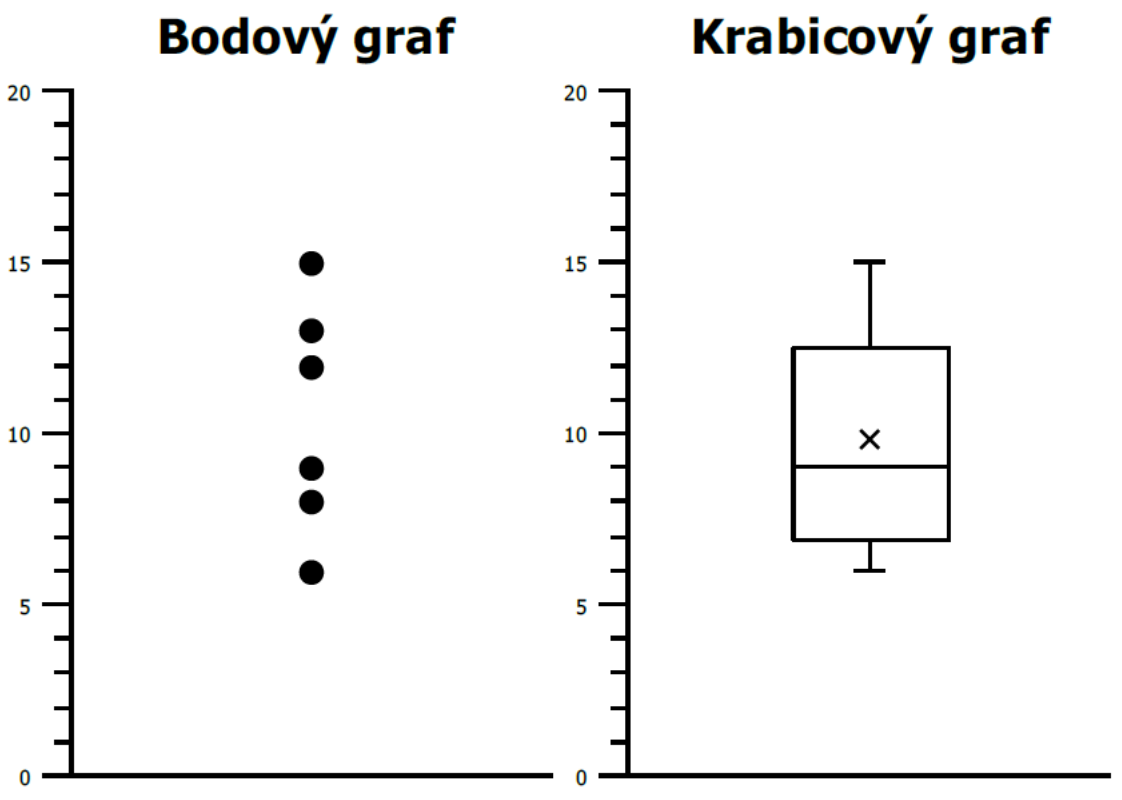
- Porovnávání průměrů v k>2 diskrétních skupinách

Kontrola | Vzorek 1 | Vzorek 2

- Proměnná, která kategorizuje jednotlivé pozorované skupiny se nazývá **KATEGORIÁLNÍ FAKTOR**

ANOVA

Vizualizace dat



Odlehlé hodnoty

- Hrubé chyby, překlepy, prokazatelné selhání lidí či techniky
- Důsledky poruch, chybného měření, technologických chyb

- Vyloučení odlehlých hodnot
  - *ANOVA*
- Zachování odlehlých hodnot
  - *Kruskalův-Wallisův test*

ANOVA

Předpoklady analýzy

- Normalita rozložení dat
- Nezávislost výběru
- Shodnost rozptylu (homoskedosticita)

Nulová hypotéza ANOVy

- H_0 – všechny střední hodnoty jsou si rovny
- H_A – alespoň jedna dvojice středních hodnot se liší
  - *Platí-li H_A je pro potřeba využít Post hoc analýzu*

ANOVA

Rozklad celkové variability dat

- Systematická část
  - *Kategorizace skupin,*
    tzv. vysvětlitelná variabilita
- Náhodná část
  - *Chyby neovlivnitelné, přítomné ve všech měřeních*

Post hoc analýza

- Porovnání středních hodnot všech dvojic populací
  - *Fischerovo LSD*
  - *Bonferroniho metoda*
  - *Scheffého metoda*
  - *Tukeyho metoda*
  - *Tukey HSD*

Kruskalův-Wallisův test

- Neparametrická ANOVA
- Porovnávání mediánů v k>2 nezávislých skupinách
- Výběry nesplňují požadavky pro použití parametrické ANOVy

Kruskalův-Wallisův test

Nulová hypotéza

- H_0 – mediány všech porovnávaných skupin jsou si rovny
- H_A – alespoň jedna dvojice mediánů porovnávaných skupin se liší

Kruskalovo-Wallisova Post hoc analýza

- Porovnání mediánů všech dvojic populací
  - *Dunnova metoda*
  - *Neméniova metoda*

Neparametrické testy

- srovnání souboru statistických dat, u kterých není předpoklad normálního rozdělení pravděpodobností sledovaného znaku
- použitelné i pro symetrická data a při výskytu odlehlých hodnot
- menší síla testu (místo původních hodnot využívají jejich pořadí)
- nižší citlivost a přesnost
- testují nulovou hypotézu

Neparametrické testy - rozdělení

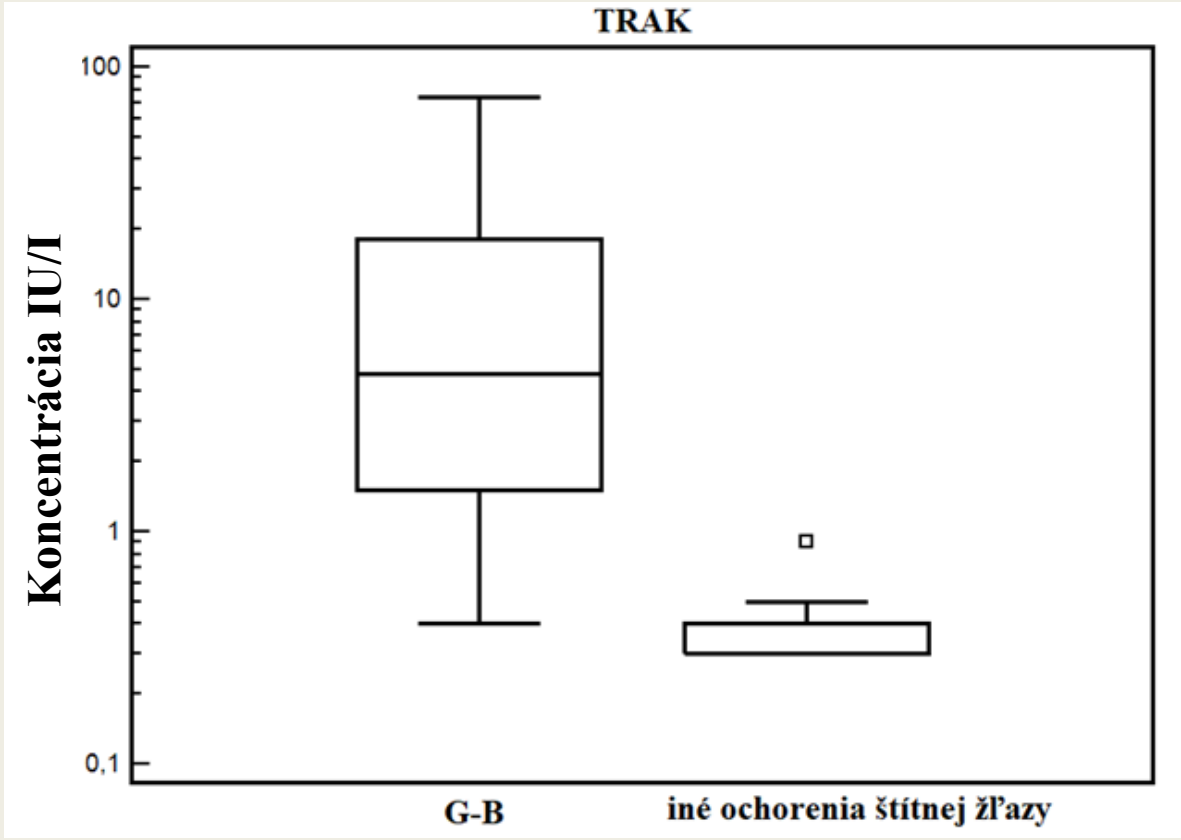
- jednovýběrové statistické testy
  - *jednovýběrový t-test*
  - *jednovýběrový test rozptylu*
- dvouvýběrové statistické testy
  - *nepárový Mann-Whitneyho test*
  - *párový Wilcoxon a znaménkový test*

Mann-Whitney test

- Neparametrický ekvivalent t-testu nezávislých vzorků
- Ověření významnosti rozdílu mezi dvěma nezávislými vzorky
- Distribuce (rozložení) vzorku není normální, není možná jejich transformace na normální distribuci pomocí logaritmické transformace
- Počítá s pořadím dat v souborech, ne s originálními daty

Mann-Whitney test

- měřená koncentrace protilátek u 42 pacientů s G-B a 21 pacientů s jiným onemocněním ŠŽ
- nulová hypotéza je, že není rozdíl mezi skupinami pacientů s G-B a pacienty s jiným onemocněním ŠŽ



Mann-Whitney test

- pořadí sloučených hodnot
- pořadí hodnot v jednotlivých skupinách dat je sečteno a menší ze součtů je použit pro srovnání s kritickou hodnotou testu
- výsledkem testu je P <0,05, nulovou hypotézu tedy zamítáme a výsledek testu potvrzuje statisticky signifikatní rozdíl mezi porovnávanými skupinami pacientů

	G-B	iné ochorenia štítnej žľazy bez AITx
Počet vzoriek	42	21
Najnižšia hodnota TRAK	0,4000	0,3000
Najvyššia hodnota TRAK	74,0000	0,9000
Medián	4,8000	0,3000
95% interval spoľahlivosti pre medián	2,5728-9,2540	0,3000-0,4000
Hladina významnosti	P<0,0001	

Wilcoxonův test

- neparametrická obdoba párového t-testu a jednovýběrového t-testu
- neparametrický test pro párové hodnoty dvou závislých souborů, u kterých není jistota normálního rozložení dat
- porovnání stejné skupiny respondentů ve dvou podmínkách (před experimentální manipulací a po ní, hladina analytu před a po podání léku…)

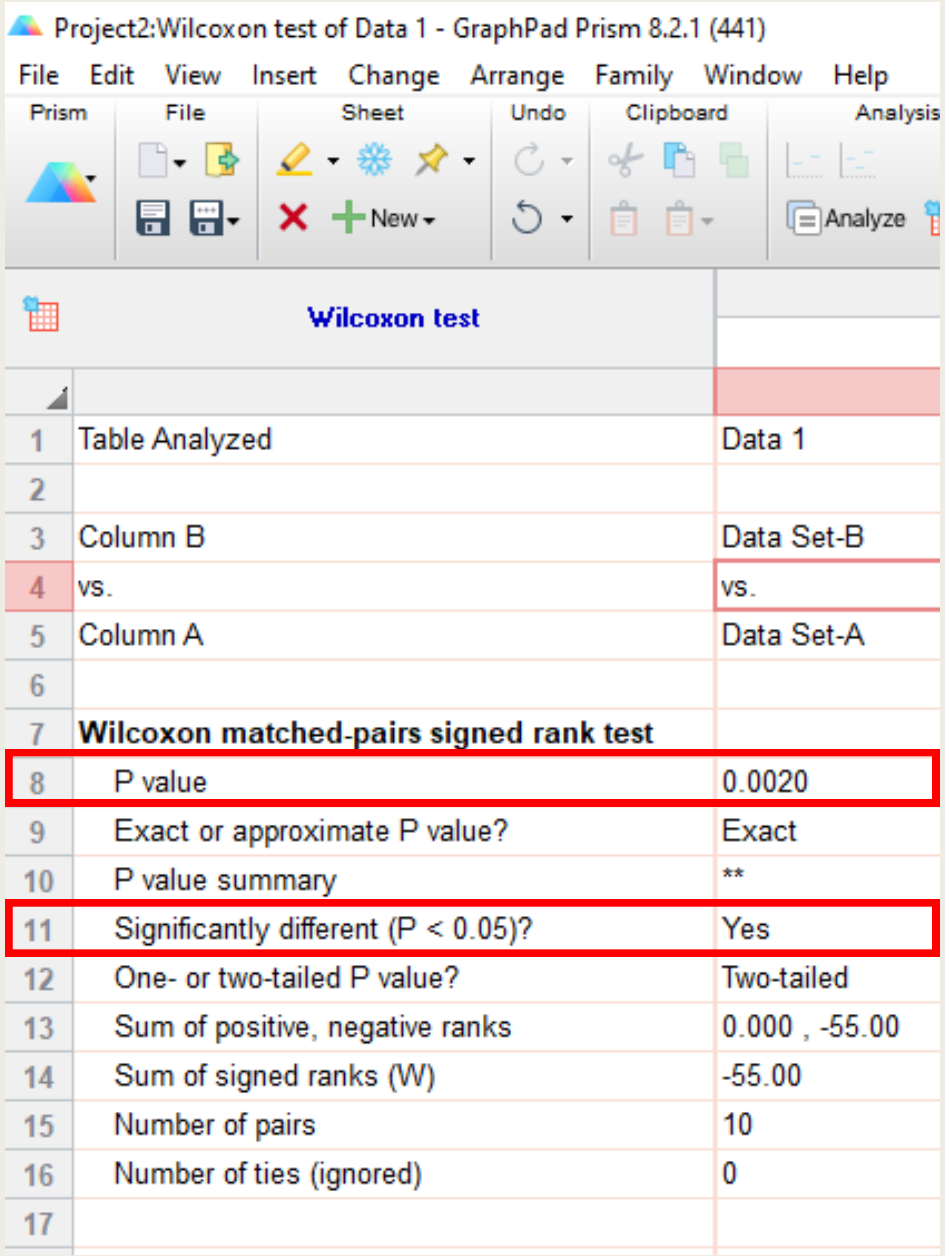
Wilcoxonův test

- vytvoření rozdílu párových dat v rámci každého vzorku a následné seřazení rozdílu podle pořadí nezávisle na znaménku
- pořadí záporných a kladných rozdílu je poté zvlášť sečteno a menší z těchto hodnot je srovnána s kritickou hodnotou testu

Pacient	Glc před lekmi (mmol/l)	Glc po lecich (mmol/l)
1	9,7	3,6
2	10,3	3,9
3	11,3	4,2
4	9,7	4,8
5	8,8	3,5
6	9,9	3,7
7	8,7	4
8	7,9	4,5
9	8,8	3,7
10	9	4,1

Wilcoxonův test

- statistický software GraphPad Prism 8
- nulová hypotéza: není předpoklad statisticky signifikantního rozdílu mezi hladinou glukózy před a po podání léku



Wilcoxon test

1	Table Analyzed	Data 1
2		
3	Column B	Data Set-B
4	vs.	vs.
5	Column A	Data Set-A
6		
7	**Wilcoxon matched-pairs signed rank test**	
8	P value	0.0020
9	Exact or approximate P value?	Exact
10	P value summary	**
11	Significantly different (P < 0.05)?	Yes
12	One- or two-tailed P value?	Two-tailed
13	Sum of positive, negative ranks	0.000 , -55.00
14	Sum of signed ranks (W)	-55.00
15	Number of pairs	10
16	Number of ties (ignored)	0
17		

Použitá literatura

- ŠÁNA, Jiří a Ondřej SLABÝ. Úvod do molekulární medicíny: cvičení (biomarkerové studie). Brno: Masarykova univerzita, 2017. ISBN 978-80-210-8538-1.
- Neparametrické testy, Institut biostatistiky a analýz, Masarykova univerzita, J. Jarkovský, L. Dušek
- Encyklopedie laboratorní medicíny pro klinickou praxi 2015 [online]. Dostupné z: http://www.demo4.smitka.eu/encyklopedie/A/PIABE.htm
- https://statistikapspp.sk/wilcoxonov-test/
- https://portal.matematickabiologie.cz

DĚKUJEME ZA POZORNOST!!!