Kalužová Anna; Nohelová Lucie; Siegl František; Štolbová Daniela; Vališová Martina

Základní pojmy

- Statistika = teoretická a praktická činnost zkoumající hromadné jevy
- Soubor = homogenní množství všech možných opakování definovaných přesně co do kvality a rozsahu
- Výběr = skupina opakování/variant vybraných ze souboru

Základní popisná statistika

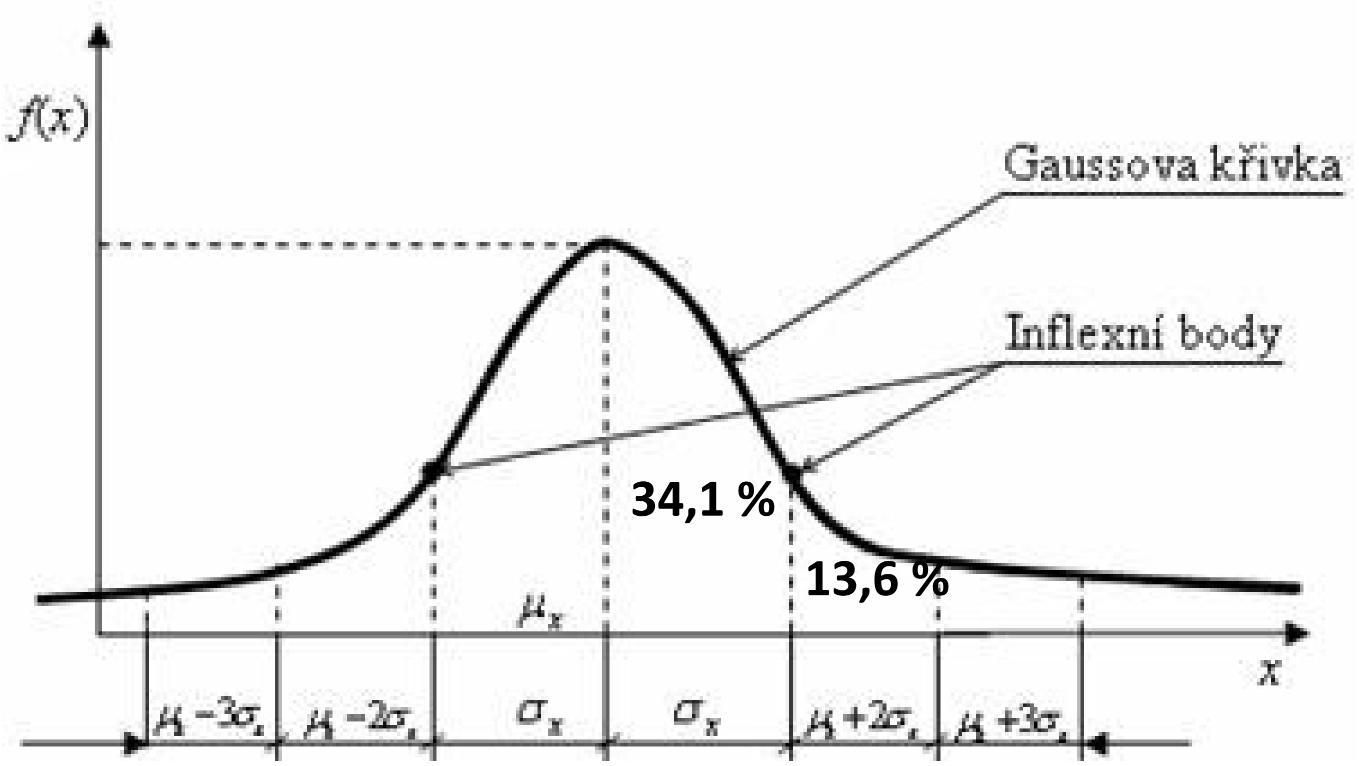
- Aritmetický průměr → $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
- Medián
- Směrodatná odchylka, rozptyl → $$s = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$$
- Relativní směrodatná odchylka → $$s_r = \frac{s}{\bar{x}}.100$$

NORMÁLNÍ ROZDĚLENÍ - Gaussova křivka

- Rozdělení pravděpodobnosti spojité náhodné veličiny (IQ)

μ … střední hodnota

σ … směrodatná odchylka



EXCEL:
=NORM.DIST(x; střední hodnota; směrodatná odchylka, 1)

Poissonovo rozdělení

- Popisuje pravděpodobnost výskytu sledovaného znaku za danou (časovou) jednotku
- Platí: jev je <u>náhodný</u>

λ .... Intenzita

x .... Počet opakování (n -> ∞)

$$\boxed{p(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \text{ kde } x = 0,1,2,3,\ldots}$$

Př. Za 1 hodinu spadne na Zemi 20 meteorů. Jaká je pravděpodobnost (p), že v následujících 10 min spadnou 3 meteory?

p(x) =?

λ = 20/60

x = 3

Parametrické vs. neparametrické testy

- Parametrický test = test, pro jehož odvození je nutné specifikovat typ rozdělení, případně jeho parametry.

- Neparametrický test = test, pro jehož odvození není nutné specifikovat typ rozdělení.

Postup při hledání vhodného tes

- Zjistit, zda naměřené hodnoty splňují kritéria p použití parametrických či neparametrických tes
  - ✓ Pro soubory s n > 25 s normálním rozděle a stejnými rozptyly použít <u>parametrické tes</u>
  - ✓ Pro ostatní soubory <u>neparametrické testy</u>.

t-test

- Jednovýběrový: testujeme výběr od konstanty (střední hodnota základního souboru)
- Dvouvýběrový: porovnání dvou výběrových souborů
  a) Párový t-test: 2 měření u jednoho výběrového souboru (1. měření před aplikací pokusného zásahu, 2. po aplikaci pokusného zásahu)
     - Testujeme hypotézu, že střední hodnota měření před pokusem a po pokusu se rovnají

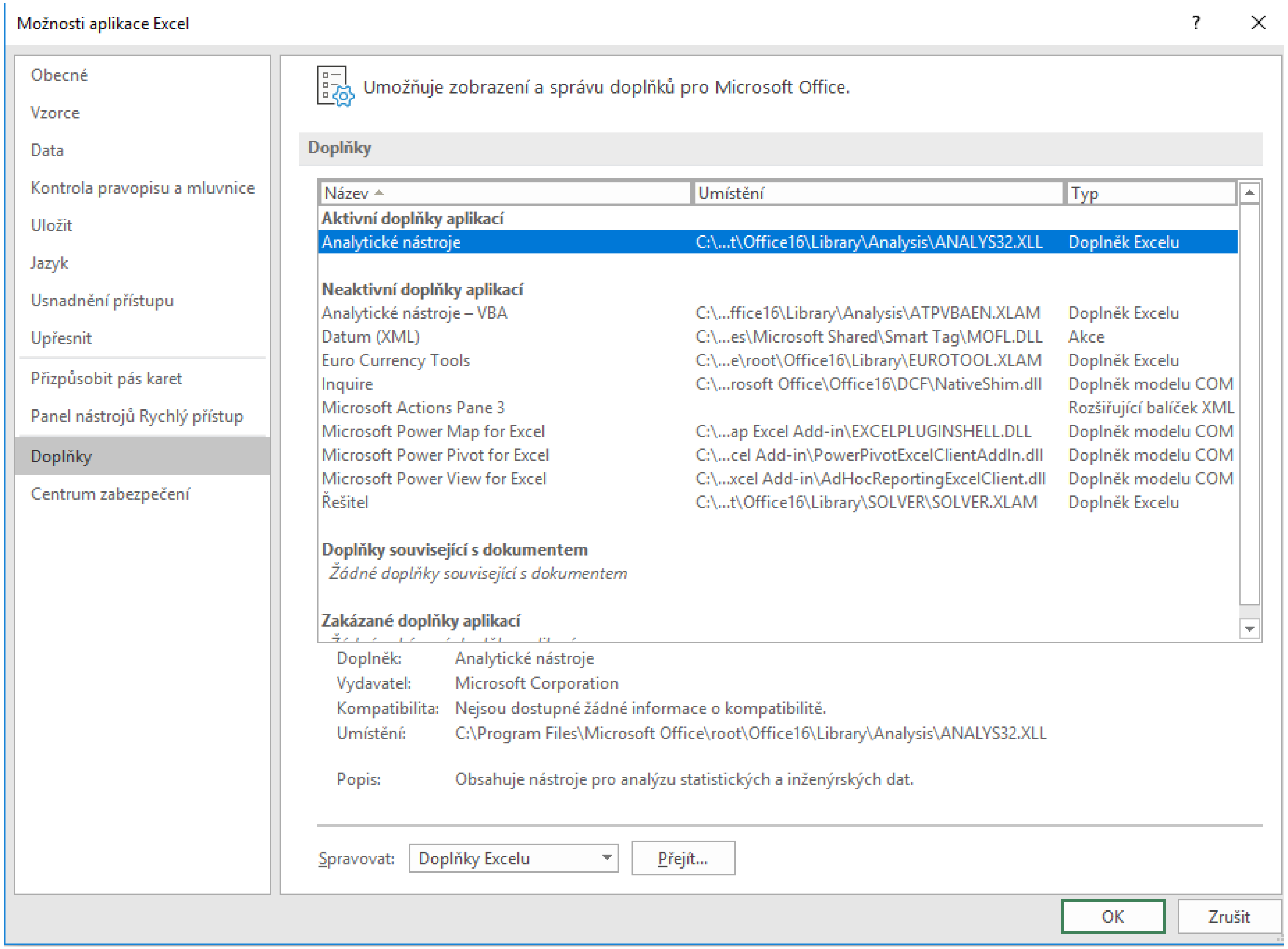
  b) Nepárový t-test: porovnávaná data pocházející ze dvou různých skupin (např. porovnání hodnot pokusné a kontrolní skupiny)

F-test

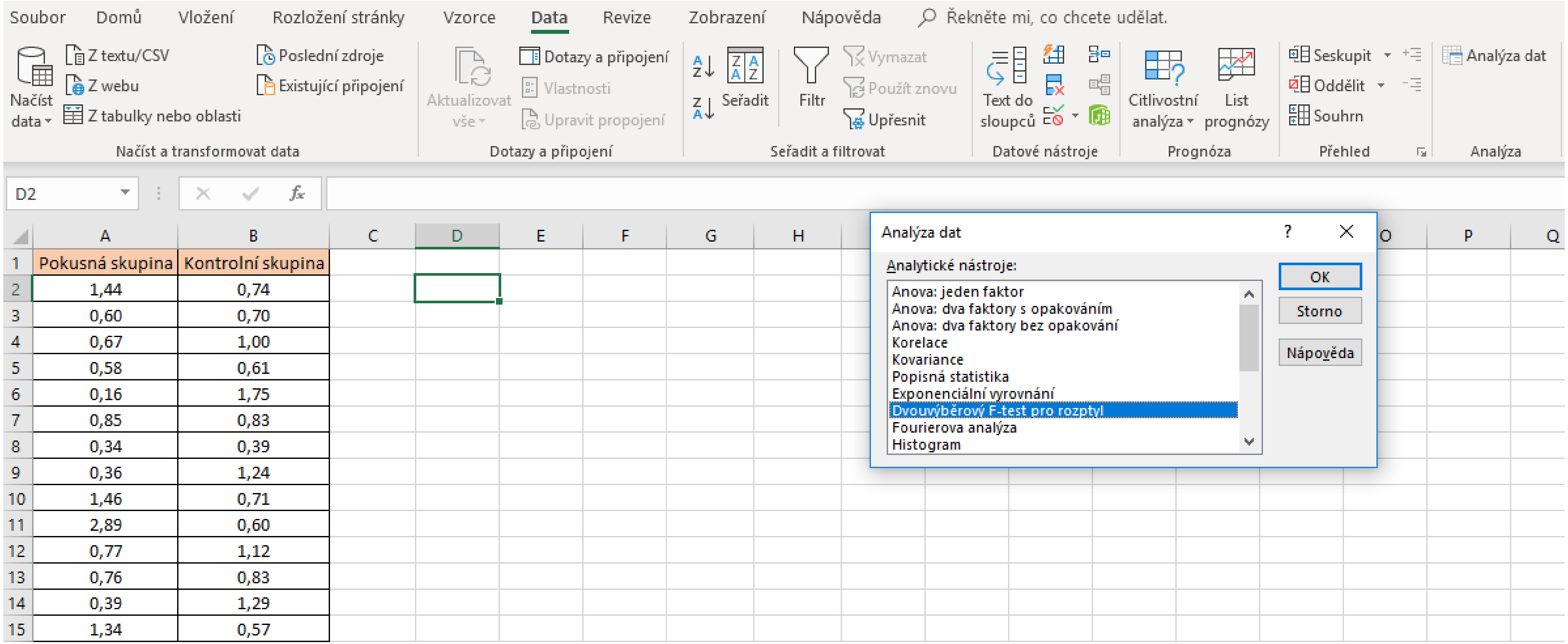
- U dvouvýběrového nepárového t-testu musíme nejdříve provést F-test
- A to kvůli otestování rozdílu rozptylů obou souborů
- Podle výsledku F-testu zvolíme dvouvýběrový t-test s rovností/nerovností rozptylů

Excel: stáhnutí doplňku analýza dat

- Soubor → Možnosti → Doplňky → Analytické nástroje



Excel: F-test

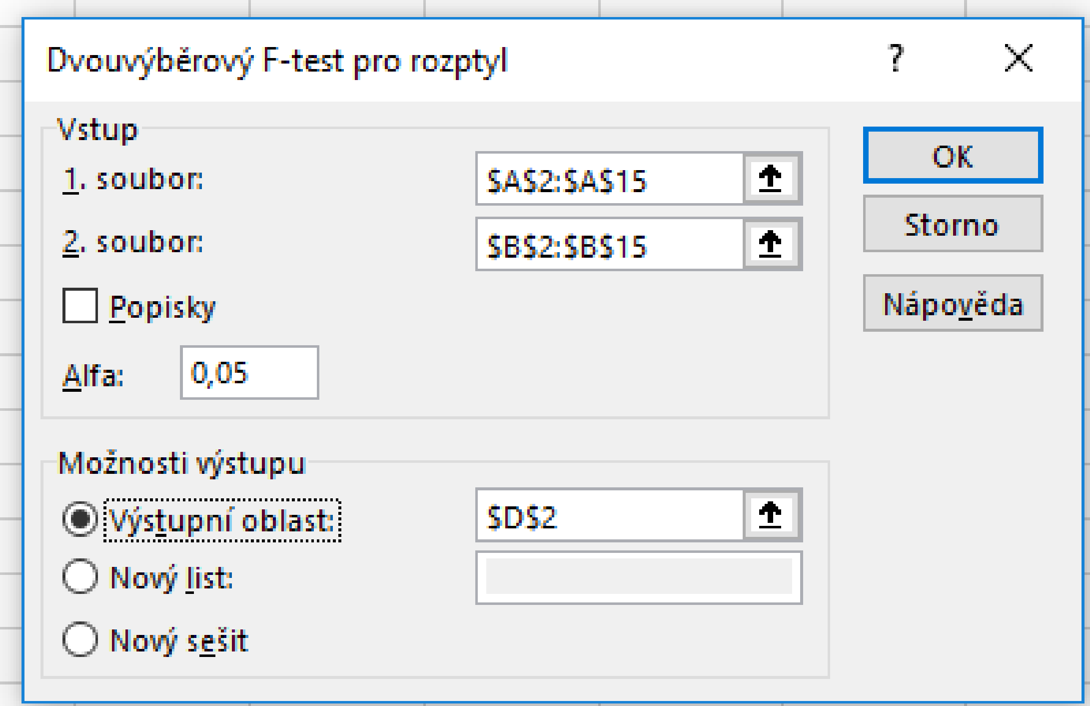


	A	B
1	Pokusná skupina	Kontrolní skupina
2	1,44	0,74
3	0,60	0,70
4	0,67	1,00
5	0,58	0,61
6	0,16	1,75
7	0,85	0,83
8	0,34	0,39
9	0,36	1,24
10	1,46	0,71
11	2,89	0,60
12	0,77	1,12
13	0,76	0,83
14	0,39	1,29
15	1,34	0,57

Excel: F-test

Soubor s větším rozptylem musí být jako první

	A	B
1	Pokusná skupina	Kontrolní skupina
2	1,44	0,74
3	0,60	0,70
4	0,67	1,00
5	0,58	0,61
6	0,16	1,75
7	0,85	0,83
8	0,34	0,39
9	0,36	1,24
10	1,46	0,71
11	2,89	0,60
12	0,77	1,12
13	0,76	0,83
14	0,39	1,29
15	1,34	0,57



Excel: F-test

	A	B	C	D	E	F
1	Pokusná skupina	Kontrolní skupina				
2	1,44	0,74		Dvouvýběrový F-test pro rozptyl		
3	0,60	0,70				
4	0,67	1,00			*Soubor 1*	*Soubor 2*
5	0,58	0,61		Stř. hodnota	0,900714	0,884286
6	0,16	1,75		Rozptyl	0,496623	0,130749
7	0,85	0,83		Pozorování	14	14
8	0,34	0,39		Rozdíl	13	13
9	0,36	1,24		F	3,798276	
10	1,46	0,71		P(F<=f) (1)	0,011247	
11	2,89	0,60		F krit (1)	2,576927	
12	0,77	1,12				
13	0,76	0,83		**P < 0,05 → dvouvýběrový t-test s nerovností rozptylů**		
14	0,39	1,29		**P > 0,05 → dvouvýběrový t-test s rovností rozptylů**		
15	1,34	0,57				

Excel: t-test

	A	B	C	D	E	F	G
1	Pokusná skupina	Kontrolní skupina					
2	1,44	0,74		Dvouvýběrový t-test s nerovností rozptylů			
3	0,60	0,70					
4	0,67	1,00			*Soubor 1*	*Soubor 2*	
5	0,58	0,61		Stř. hodnota	0,900714	0,884286	
6	0,16	1,75		Rozptyl	0,496623	0,130749	
7	0,85	0,83		Pozorování	14	14	
8	0,34	0,39		Hyp. rozdíl stř. hodnot	0		
9	0,36	1,24		Rozdíl	19		
10	1,46	0,71		t Stat	0,077607		
11	2,89	0,60		P(T<=t) (1)	0,469476		
12	0,77	1,12		t krit (1)	1,729133		
13	0,76	0,83		P(T<=t) (2)	0,938952		
14	0,39	1,29		t krit (2)	2,093024		
15	1,34	0,57					
16				**P < 0,05 → změna je statisticky významná na hladině významnosti 95 %**			
17				**P > 0,05 → změna není statisticky významná na hladině významnosti 95 %**			

Neparametrické testy

- Nemají předpoklady o rozložení vstupujících dat, lze je tedy použít i při asymetrickém rozložení, odlehlých hodnotách, či nedetekovatelném rozložení
- Nelze u nich předpokládat normální rozdělení pravděpodobností sledovaného znaku
- Snížená síla těchto testů je způsobena redukcí informační hodnoty původních dat, kdy neparametrické testy nevyužívají původní hodnoty, ale nejčastěji pouze jejich **pořadí**
- **<u>Neparametrické testy testují nulovou hypotézu</u>**
- **<u>Mann-Whitneyho nepárový test, Wilcoxonův párový test, chí-kvadrát</u>**

Mann-Whitney test

- Používá se pro hodnocení **nepárových** pokusů, kdy porovnáváme 2 různé výběrové soubory (*pokusný zásah A*, *B*). Testujeme hypotézu, že veličina *X* odpovídající pokusnému zásahu „*A*“ a veličina *Y* odpovídající pokusnému zásahu „*B*“ mají totéž rozdělení pravděpodobností
- Pracuje se s pořadím, ne s původními hodnotami
- Menší z obou součtu pořadí je porovnán s kritickou hodnotou testu, pokud je tato hodnota menší než kritická hodnota testu, zamítáme nulovou hypotézu shody distribučních funkcí obou skupin

Mann-Whitney-ův test pro neparametrické testování oboustranné hypotézy, že není rozdíl ve zdravotním stavu stromků lišících se svým kultivarem

Zdravotní stav stromků byl odhadnut pomocí přibližné stupnice s hodnotami od 1 (zcela zdravý strom) do 5 (zcela mrtvý strom).
H_0: Zdravotní stav jedinců se neliší mezi dvěma kultivary A a B.
H_A: Zdravotní stav jedinců se liší mezi kultivary A a B.

Zdravotní stav jedinců kultivaru A: 2, 2, 1, 2, 3, 4, 2, 3, 1, 5 $n_1 = 10$
Zdravotní stav jedinců kultivaru B: 4, 5, 3, 1, 4, 3, 5, 2, 1, 2 $n_1 = 10$

Po převodu na pořadí (s průměrným pořadím pro jedince se shodnou hodnotou zdravotního stavu) dostáváme:
Kultivar A: 13,5; 13,5; 18,5; 13,5; 8,5; 5,0; 13,5; 8,5; 18,5; 2,0 $R_1=115{,}0$
Kultivar B: 5,0; 2,0; 8,5; 18,5; 5,0; 8,5; 2,0; 13,5; 18,5; 13,5 $R_2= 95{,}0$

$U=n_1n_2 + n_1(n_1+1)/2 – R_1 = (10)(10) + (10)(11)/2 – 115 = 100 + 55 – 115 = 40$
$U' = n_1n_2 – U = 60$
$p=0{,}481$ (exaktní odhad), H_0 zdravotní stav stromků nezávisí na kultivaru proto nezamítáme

Wilcoxonův test

- Používá se pro hodnocení párových testů (obdoba párového t-testu)
- 1. spočítat rozdíly mezi hodnotami pozorování v párech
- 2. zbylé rozdíly seřadíme podle velikosti jejich absolutní hodnoty (vzestupně)
- 3. poté se spočte součet pořadí kladných a součet pořadí záporných rozdílů (označujeme je T_+ a T_-)
- 4. menší z hodnot lze porovnat se známým rozdělením této statistiky nebo užít aproximaci normálním rozdělením
- Používá se opět nulová hypotéza

Příklad Wilcoxonova testu pro jeden výběr

Chceme srovnat průměrný energetický příjem skupiny 11 žen ve věku 22 – 30 let s doporučenou hodnotou (7725 kJ).

Nulovou a alternativní hypotézu vyjádříme jako: $H_0 : \tilde{x} = x_0 \quad H_1 : \tilde{x} \neq x_0$

Žena	Denní energetický příjem v kJ	Diference od hodnoty 7725 kJ	Pořadí absolutní hodnoty diference
1	5260	-2465	11
2	5470	-2255	10
3	5640	-2085	9
4	6180	-1545	8
5	6390	-1335	7
6	6515	-1210	6
7	6805	-920	4
8	7515	-210	1,5
9	7515	-210	1,5
10	8230	505	3
11	8770	1045	5

Výpočet testové statistiky: $S^+ = \sum_{y_i > 0} R_i = 8$ a $S^- = \sum_{y_i < 0} R_i = 58$

$$\min(S^+, S^-) = 8$$

Kritická hodnota z tabulek pro *n* = 11: $w_n(\alpha) = w_{11}(0{,}05) = 10$

Výsledná hodnota statistiky min(S^+,S^-) je menší než 10: ⟹ Zamítáme H_0

Kontingenční tabulky

- Kvalitativní znaky (je/není; ano/ne;…);
- Testování hypotéz nebo k posouzení, zda mezi znaky existuje nějaký vztah
- Přehledné sledování závislosti mezi dvěma nebo více proměnnými
- Dvourozměrné tabulky – závislost dvou proměnných
  - Čtyřpolní tabulky (2×2) – nejjednodušší příklad

	ano	ne	součet
Kontrolní	a	b	a+b
Pokusná	c	d	c+d
Součet	a+c	b+d	n

Byla vakcinována telata. V kontrolní i pokusné skupině bylo 12 pacientů (celkem 24). V pokusné (vakcinované) skupině onemocněla 2 telata, neonemocnělo 10 telat. V kontrole (nevakcinované) onemocnělo 8 telat, neonemocněla 4 telata.

2 × 2 kontingenční tabulka

	onemocnělo	**neonemocnělo**	**součet**
Kontrola	**8**	**4**	**12**
Vakcinace	**2**	**10**	**12**
Součet	**10**	**14**	**24**

Vylučování odlehlých výsledků paralelních hodnot

- Odlehlá hodnota = <u>hrubá chyba</u> (nepatří do souboru) a <u>ne náhodná!</u>
- Posouzení na základě typu rozdělení náhodné veličiny:
  - NORMÁLNÍ ROZDĚLENÍ
  - NEZNÁMÉ ROZDĚLENÍ

NORMÁLNÍ ROZDĚLENÍ

- Grubbsův test extrémních odchylek (pro n ≥ 3)
  1. Seřadit hodnoty výběrového souboru do vzestupné variační řady
  2. Výpočet $\bar{x}$ a s ze všech hodnot souboru
  3. Výpočet testovacího kritéria pro poslední/první hodnotu řady

$$T_n = \frac{x_n - \bar{x}}{s_n} \qquad \text{nebo} \qquad T_1 = \frac{\bar{x} - x_1}{s_n}$$

  4. Porovnání s tabelovanou kritickou hodnotou

$$T_1\,(T_n) \geq T_{p,n} \longrightarrow \boxed{\textbf{ODLEHLÁ}}$$

NEZNÁMÉ ROZDĚLENÍ

- Dixonův test extrémních odchylek

1. Seřadit hodnoty výběrového souboru do vzestupné variační řady
   Variační rozpětí: $R = x_{max} - x_{min}$
2. Výpočet testovacího kritéria pro poslední/první hodnotu řady

$$Q_n = \frac{x_n - x_{n-1}}{R} \qquad Q_1 = \frac{x_2 - x_1}{R}$$

3. Porovnání s tabelovanou kritickou hodnotou

$$Q_n \geq Q_p \longrightarrow \boxed{\textbf{ODLEHLÁ}}$$

Při rozboru křemičitanu byl nalezen tento obsah SiO_2: 52,44 %, 53,82 %, 52,91 %, 50,10 %, 54,03 %, 53,89 %. Je některý z výsledků odlehlý na hladině významnosti α = 0,05 ?

Obsah SiO_2 v křemičitanu (%)	Var. řada
52,44	50,10
53,82	52,44
52,91	52,91
50,10	53,82
54,03	53,89
53,89	54,03

- Grubbs:

$$\bar{x} = 52,865$$

$$\sum(x_i - \bar{x})^2 = 11,14775$$

$$S = \sqrt{\frac{1}{6} \cdot 11,1477} = 1,363$$

$$T_1 = \frac{52,865 - 50,10}{1,363} = \boxed{2,029 > T_{p,n} = 1,996} \rightarrow \boxed{\text{ODLEHLÁ}}$$

- Dixon:

$$R = 54,03 - 50,10 = 3,93\ \%$$

$$Q1 = \frac{52,44 - 50,10}{3,93} = \boxed{0,595 > Q_{p,n} = 0,560} \rightarrow \boxed{\text{ODLEHLÁ}}$$

Lineární regrese

Regrese = závislost mezi veličinami **x** a **y**

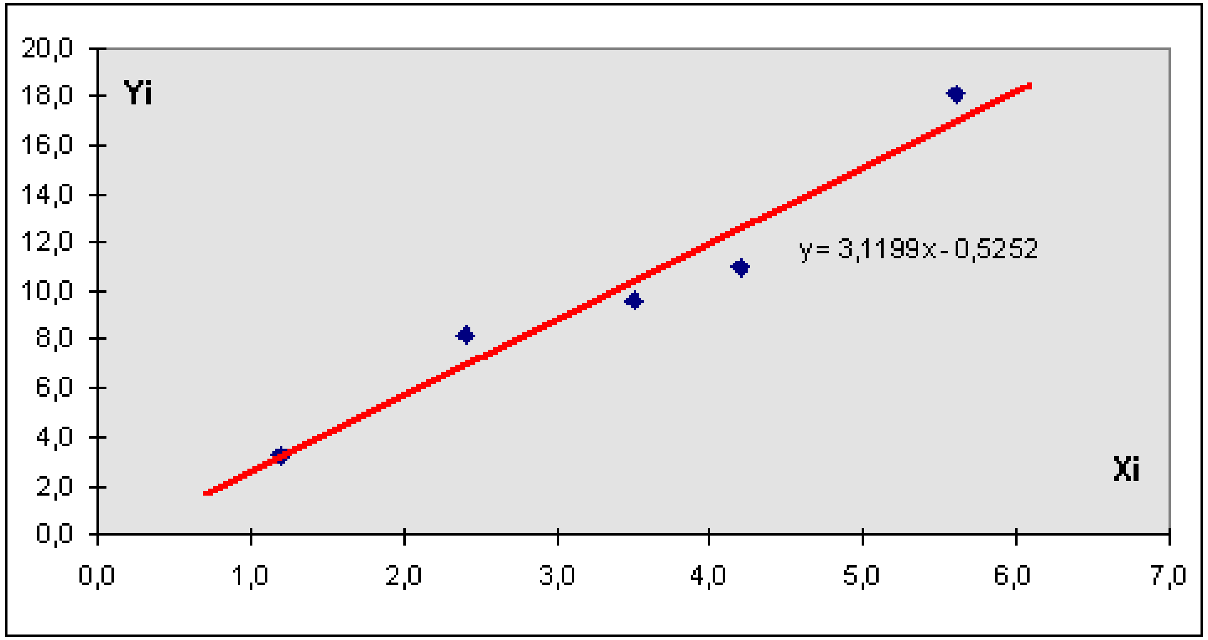
x … nezávislá proměnná (vysvětlující)

y … závislá proměnná (vysvětlovaná)

Závislost x a y : **Regresní přímka**

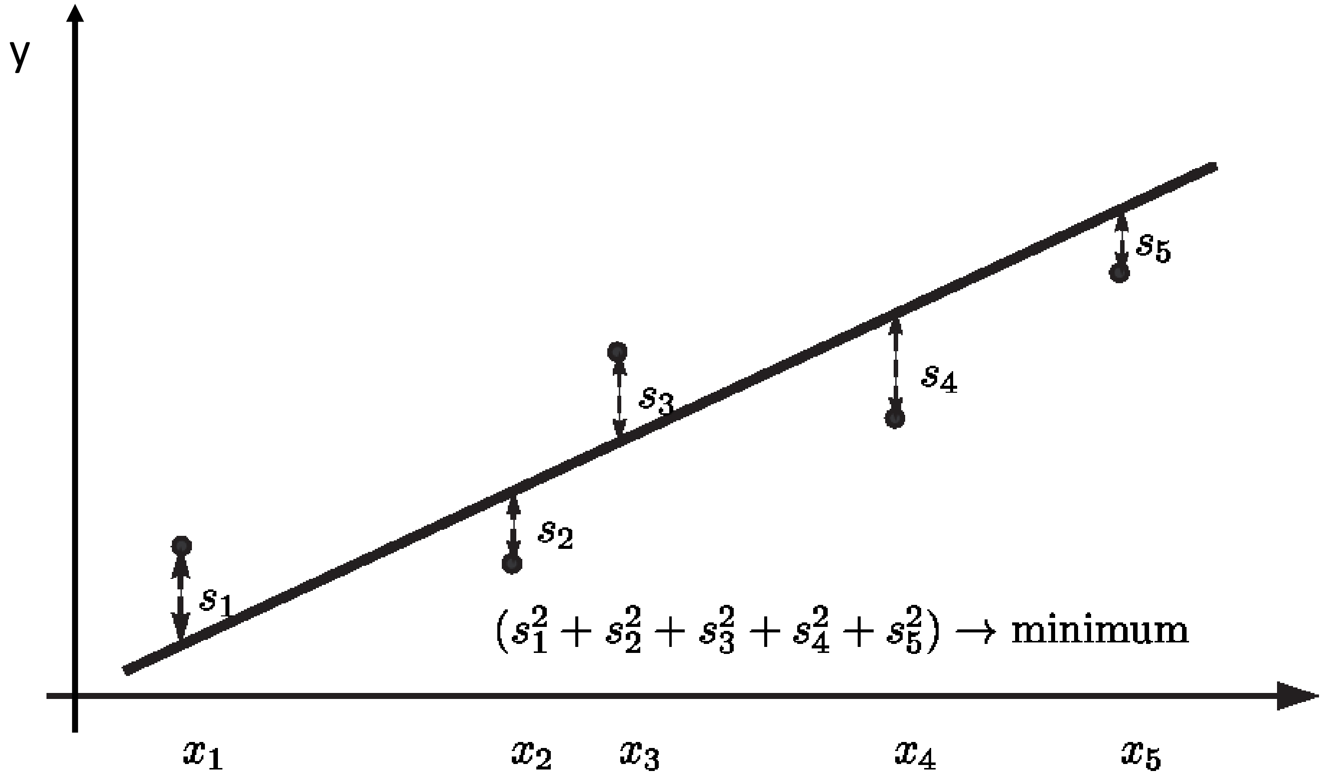
$$\boxed{y = f(x) = kx + q}$$

k, q … regresní koeficienty



Metoda nejmenších čtverců

- Aproximace závislosti mezi naměřenými veličinami
- Založena na minimalizaci tzv. reziduálního součtu čtverců



Korelační koeficient (R, r_{xy})

= vhodnost použití lineární regrese pro proložení závislosti mezi 2 veličinami

r_{xy} **… <-1 ; +1>** $\boxed{r_{xy} > 0{,}99}$ … „**Pravidlo dvou devítek**“

+1 … pozitivní korelace

-1 … negativní korelace

R^2 … koeficient determinace, <0;1> Vyšší hodnoty = vyšší úspěšnost regrese

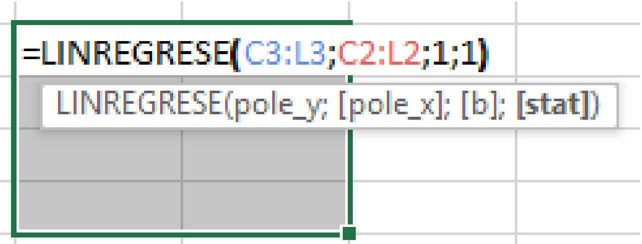
Provedení

1. **Hodnoty x (nezávislá) a y (závislá)**

t[°C]	18,3	24,5	29,9	37,3	42,7	47,7	51	58,7	62,7	66,1
R [Ω]	1073	1087	1106	1137	1164	1187	1196	1229	1236	1256

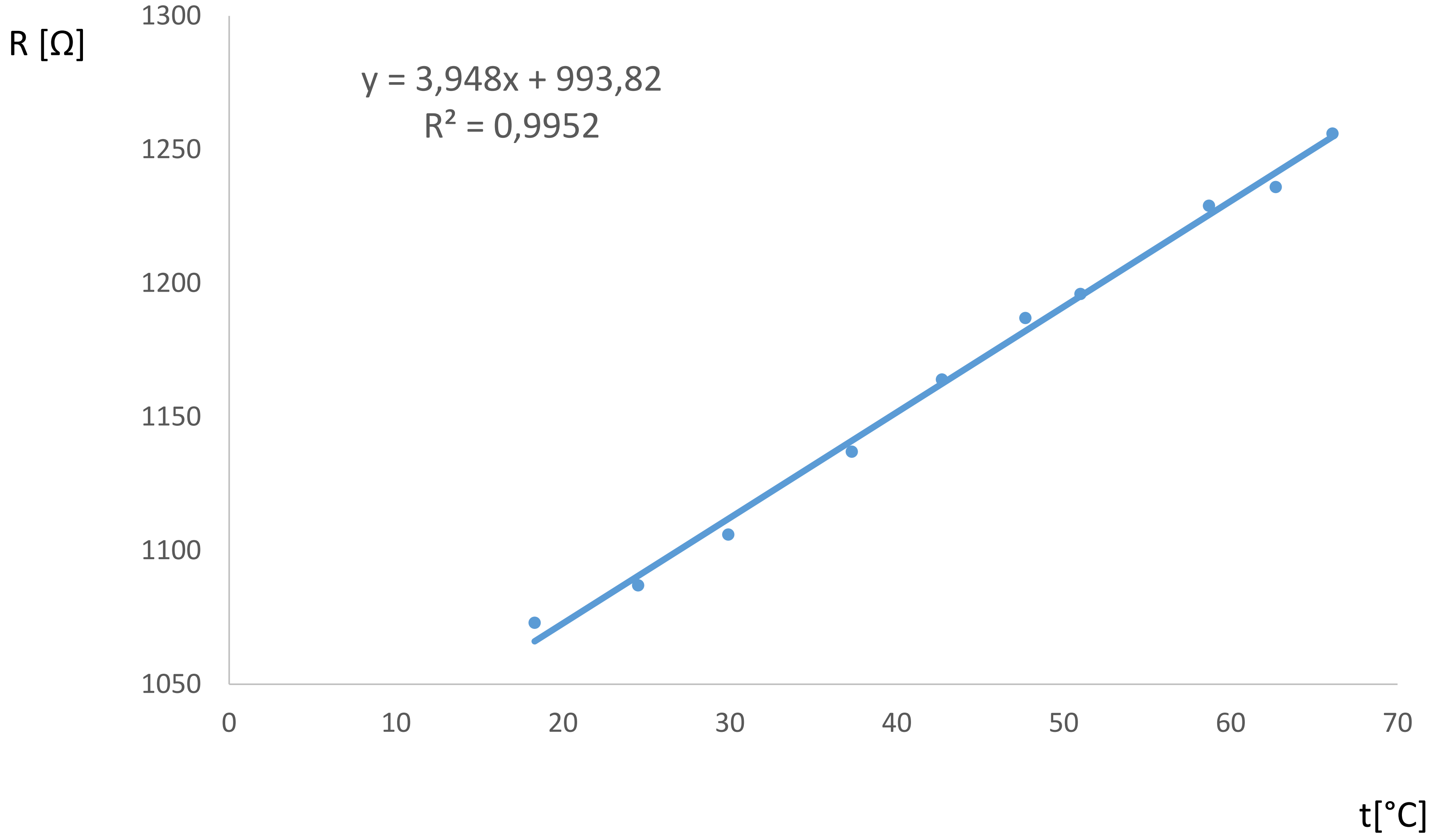
2. **Korelační koeficient** (r_{xy}) – vhodnost použití lineární regrese pro proložení závislosti mezi 2 veličinami
3. **Regresní koeficienty** (**k, q**), odchylky regresních koeficientů (σ_k, σ_q)
   - Excel: **LINREGRESE()**

     1. Vyznačit **3 řádky x 2 sloupce**
     2. Vybereme oblasti pro dané parametry
     3. **CTRL + SHIFT + ENTER**



k	3,94796	993,82403	q
σ_k	0,09691	4,5108905	σ_q
R^2	0,995203	4,7504819	

4. **Graf** – bodový, spojnice trendu (regresní přímka)
   (zobrazit rovnici grafu, hodnotu spolehlivost)


R [Ω]
y = 3,948x + 993,82
R² = 0,9952
1300
1250
1200
1150
1100
1050
0
10
20
30
40
50
60
70
t[°C]

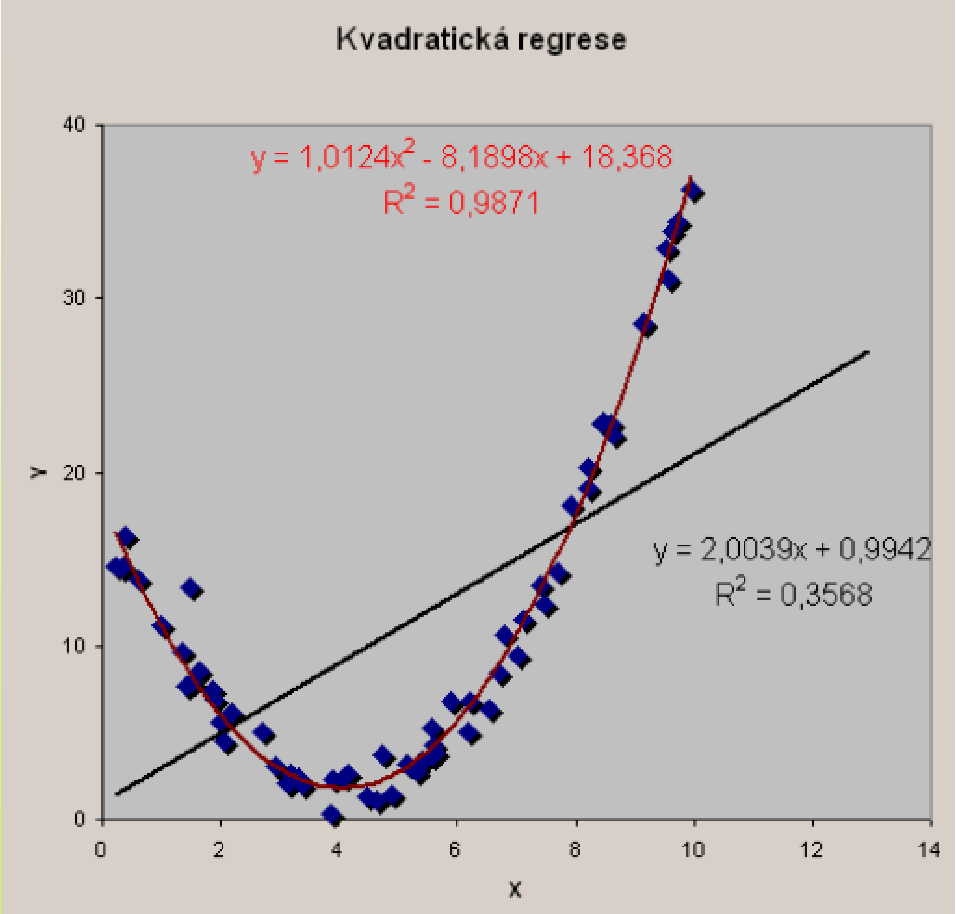
Nelineární regrese

- pro popis závislosti veličin využívá funkce nelineární v parametrech (tyto funkce nelze na lineární v parametrech převést pomocí žádné transformace)
- Funkci hledám v předepsaném tvaru (exponenciální, polynomiální,…) parametry nalezneme metodou nejmenších čtverců
- Lze provést linearizaci vztahu pomocí transformace proměnných nebo použít zobecněné lineární modely, ale také polynomiální regresi nebo nelineární regresi

Nelineární regrese

- Koeficient determinace R^2 – popisná míra vhodnosti použití regresní rovnice pro predikování
- Hodnoty blízké nule naznačují, že zvolená funkce není vhodná
- Naopak, hodnoty blízké 1 naznačují, že rovnice je velmi vhodná pro extrapolaci
- Malá hodnota ale nemusí znamenat nízký stupeň závislosti mezi proměnnými, ale může signalizovat špatně zvolenou regresní funkci

$$R^2 = \frac{S_{reg}}{S_{yy}}$$



Korelační analýza

- Cílem je určit sílu závislosti mezi dvěma veličinami = síla statistické závislosti
- Vyjádřena korelačními koeficienty r (Pearsonův, Spearmanův, Kendallův ...)
- r nabývá hodnot od -1,1
- 1 – přímá korelace
- -1 – nepřímá korelace

Korelační koeficienty

- Pearsonův - pro lineární závislost dvou náhodných veličin s normálním dvourozměrným rozdělením

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

- Spearmanův – neparametrický, robustní vůči odlehlým hodnotám, pracuje pouze s pořadním hodnot. Popisuje jak vztah funk odpovídá monotónní fci, která může být i nelineární

$$r_s = 1 - \frac{6\sum D^2}{n(n^2-1)}$$

- Mnohonásobý – vyjadřuje sílu zavislosti jedné proměné na dvou a více jiných proměných

Korelační koeficienty

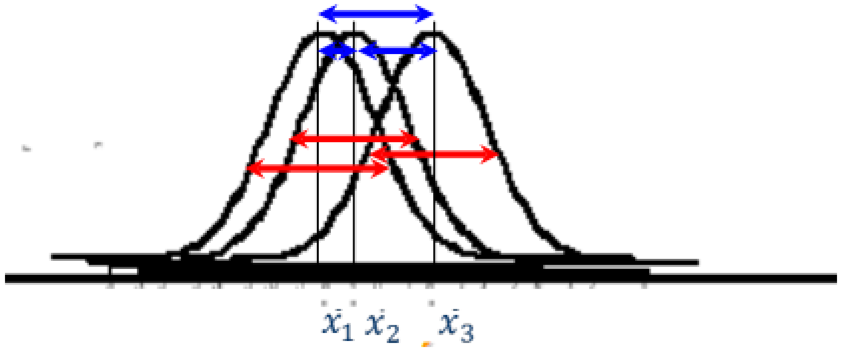
- Parciální – výpočet síly závislosti dvou proměnných v souboru více proměnných za současného zanedbání ostatních proměnných
- Kendallův – neparametrický koeficient nezávisloti, citlivější na některé nelineární vztahy, tento test nečiní žádný předpoklad povahy pravděpodobnostního rozdělení

ANOVA

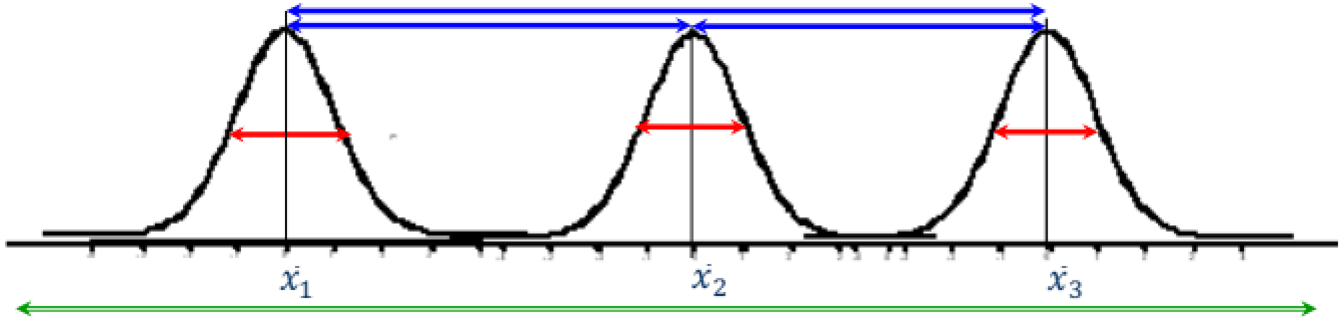
- Analysis of variance (analýza rozptylu)
- Test shody středních hodnoty pro více výběrů (pro 2 výběry – T-test)
- Anova analyzuje zdroje variability u lineárních statistických modelů – vnitrovýběrová variabilita X mezivýběrová variabilita
- Základ při výpočtu variability při analýze rozptylu jednofaktorové ANOVy je F-test

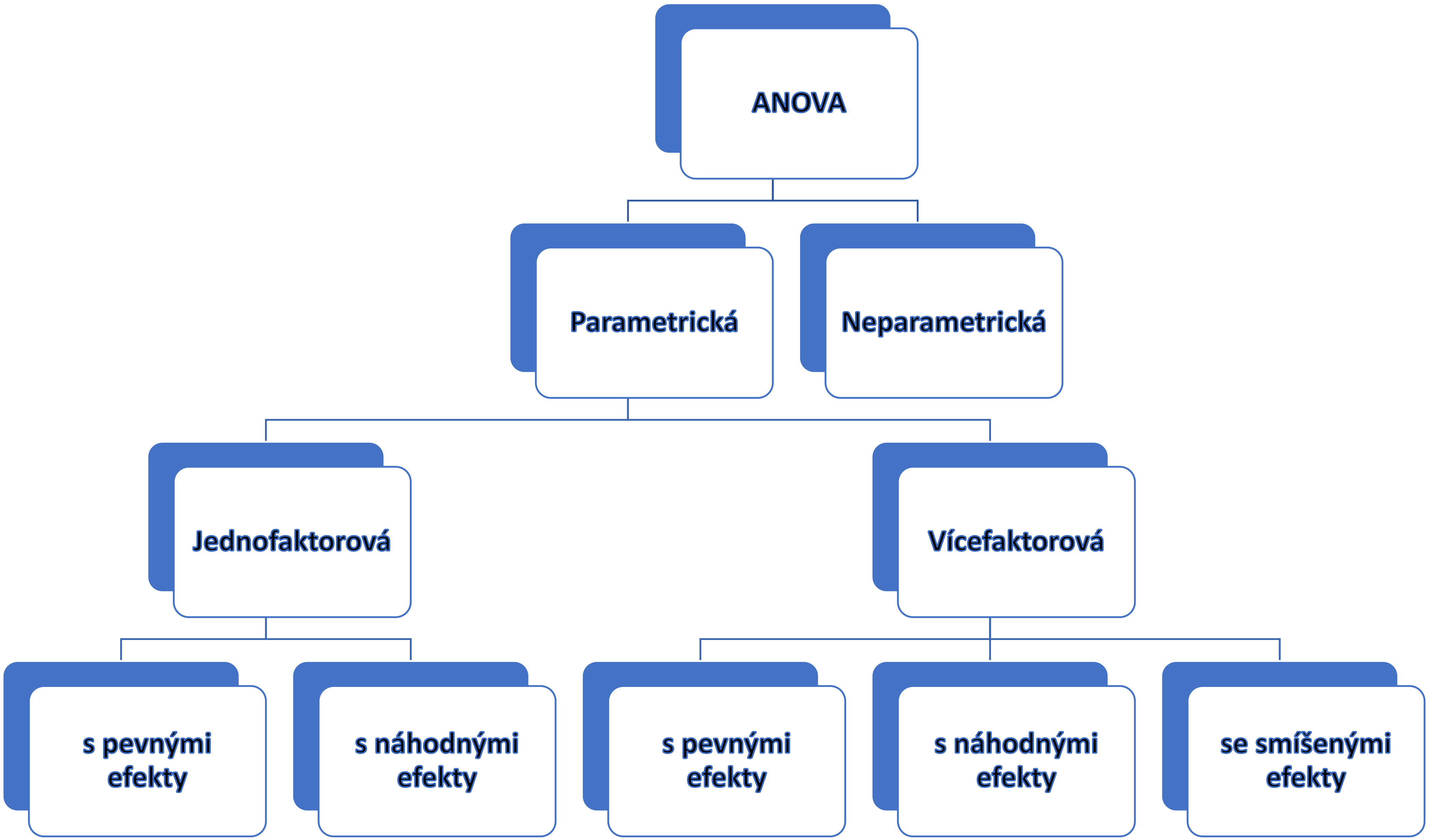
ANOVA - princip

- Variabilita mezi v … odnotami výběrů
- Variabilita uvnitř … ɔlo středních hodnot v rámci výběru



- **Anova testuje pc … s vnitrovýběrovou variabilitou**
- Mezivýk … ; vysokou pravděp … utí nulové hypotéz




ANOVA
Parametrická
Neparametrická
Jednofaktorová
Vícefaktorová
s pevnými efekty
s náhodnými efekty
s pevnými efekty
s náhodnými efekty
se smíšenými efekty

Vícefaktorová ANOVA

- Vliv dvou a více faktorů (např. vliv živné půdy a způsob kultivace u mikroorganismů)
- Nejčastěji 2-F Anova, 3-F a více řešitelná, ale obtížně interpretovatelná
- Zkoumáme efekty, které způsobují jednotlivé faktory (hlavní efekt) a efekty, které vznikají interakcí těchto faktorů (interakční efekt)
- Mezivýběrovou variabilitu lze v tomto případě rozložit na variabilitu způsobenou faktorem A, B, .. a variabilitu způsobenou interakčním efektem.
- Interakce se vyskytuje tehdy, pokud není účinek jednoho faktoru stejný při změně úrovni druhého faktoru.

Neparametrická ANOVA

- Kruskal-Wallisův test (zobecnění Mann-Whitneyho testu pro více než dva výběry)
- V případě, že nejsou splněny podmínky pro parametrickou Anovu (normalita výběru, homogenita rozptylu) + v případě velmi malých výběrů
- Není testována shoda konkrétních parametrů, ale shoda výběrových distribučních funkcí srovnávaných souborů

Neparametrická ANOVA

M1	M2	M3
6	2	9
4	4	6
9	4	11
12		

Sdružený soubor	Neupravené pořadí	Upravené pořadí
2	1	1
4	2	3
4	3	3
4	4	3
6	5	5,5
6	6	5,5
9	7	7,5
9	8	7,5
11	9	9
12	10	10

M1	M2	M3
5,5	1	7,5
3	3	5,5
7,5	3	9
10		
26	7	22

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

Testy normality

- Slouží k určení, zda lze rozdělení dat považovat za normální

- <u>Grafické metody</u>: histogram, Q-Q graf (kvantil-kvantil), nebo P-P graf (pravděpodobnost-pravděpodobnost)
- <u>Ověření normality výpočtem</u>: Shapirův-Wilkův test, Andersonův-Darlingův test, Kolmogorovův-Smirnovův test, Lillieforsův test, Chí kvadrát atd.
- <u>Hrubý odhad</u>: porovnání aritmetického průměru s mediánem (neměly by se lišit o více než 10 %)

Shapirův-Wilkův test

- Zjištění, zda se body sestrojeného kvantil-kvantilového grafu (Q-Q plotu) významně liší od regresní přímky proložené těmito body
- Především pro výběry menších rozsahů n < 50
- Testová statistika W – čím blíže 1, tím více svědčí pro normalitu (pokud hodnota testové statistiky nepřekročí tabelovanou kritickou hodnotu Shapiro-Wilkova testu, nulovou hypotézu zamítáme na dané hladině významnosti)

$$W = \frac{b^2}{S^2} = \frac{\left(\sum_{i=1}^{k} a_{n-i+1}(y_{n-i+1} - y_i)\right)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

Postup

1. Hodnoty jednotlivých pozorování seřadíme vzestupně
2. Vypočteme: $S^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i^2) - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$
3. Vypočteme: $b = a_n(y_n - y_1) + \cdots + a_{k+2}(y_{k+2} - y_k)$
4. Vypočteme testovou statistiku: $W = \frac{b^2}{S^2}$
5. Hodnotu testové statistiky porovnáme s tabelovanou kritickou hodnotu Shapiro-Wilkova testu a učiníme závěr o zamítnutí, resp. Nezamítnutí nulové hypotézy na hladině významnosti α

Děkujeme Vám za pozornost!