Základy pravděpodobnosti a matematické statistiky

11. Analýza rozptylu

FAKULTA
STROJNÍ
ČVUT V PRAZE

prof. RNDr. Gejza Dohnal, CSc.

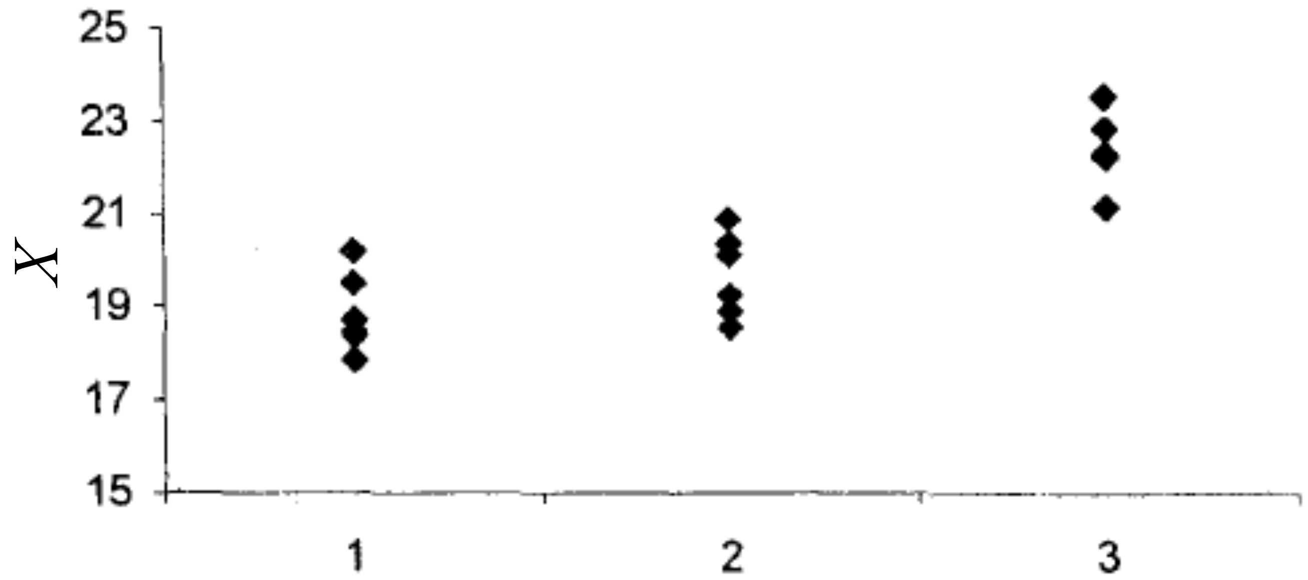
ak. rok 2023/2024

11. Analýza rozptylu

Srovnání více náhodných veličin

Příklad: Chceme porovnat kvalitu vláken dodávaných třemi různými výrobci.
Jako kritérium kvality je zvolena pevnost vláken v tahu.

$$X_1^1, X_2^1, \ldots, X_{n_1}^1, \quad X_1^2, X_2^2, \ldots, X_{n_2}^2, \quad X_1^3, X_2^3, \ldots, X_{n_3}^3,$$



nulová hypotéza H$_0$: $\mu_1 = \mu_2 = \mu_3$

alternativní hypotéza H$_A$: $\mu_i \neq \mu_j$ pro některou dvojici i,j (oboustranná)

hladina významnosti $\alpha = 5\%$

Jedna z možností je provést k(k-1)/2 porovnání pomocí dvouvýběrových testů.

ALE: tím se výrazně zvýší hladina významnosti.

Srovnání více náhodných veličin

Příklad: Chceme porovnat kvalitu vláken dodávaných třemi různými výrobci.
Jako kritérium kvality je zvolena pevnost vláken v tahu.

$$X_1^1, X_2^1, \ldots, X_{n_1}^1, \quad X_1^2, X_2^2, \ldots, X_{n_2}^2, \quad X_1^3, X_2^3, \ldots, X_{n_3}^3,$$

	hladina významnosti v *t*-testu					
počet porovnávání k	0,20	0,10	0,05	0,02	0,01	0,001
2	0,20	0,10	0,05	0,02	0,01	0,001
3	0,41	0,23	0,13	0,05	0,03	0,003
4	0,58	0,36	0,21	0,09	0,05	0,006
5	0,71	0,47	0,23	0,13	0,07	0,009
10	0,96	0,83	0,63	0,37	0,23	0,034
20	1,00	0,98	0,92	0,71	0,52	0,109
∞	1,00	1,00	1,00	1,00	1,00	1,000

nulov

alter

hladi

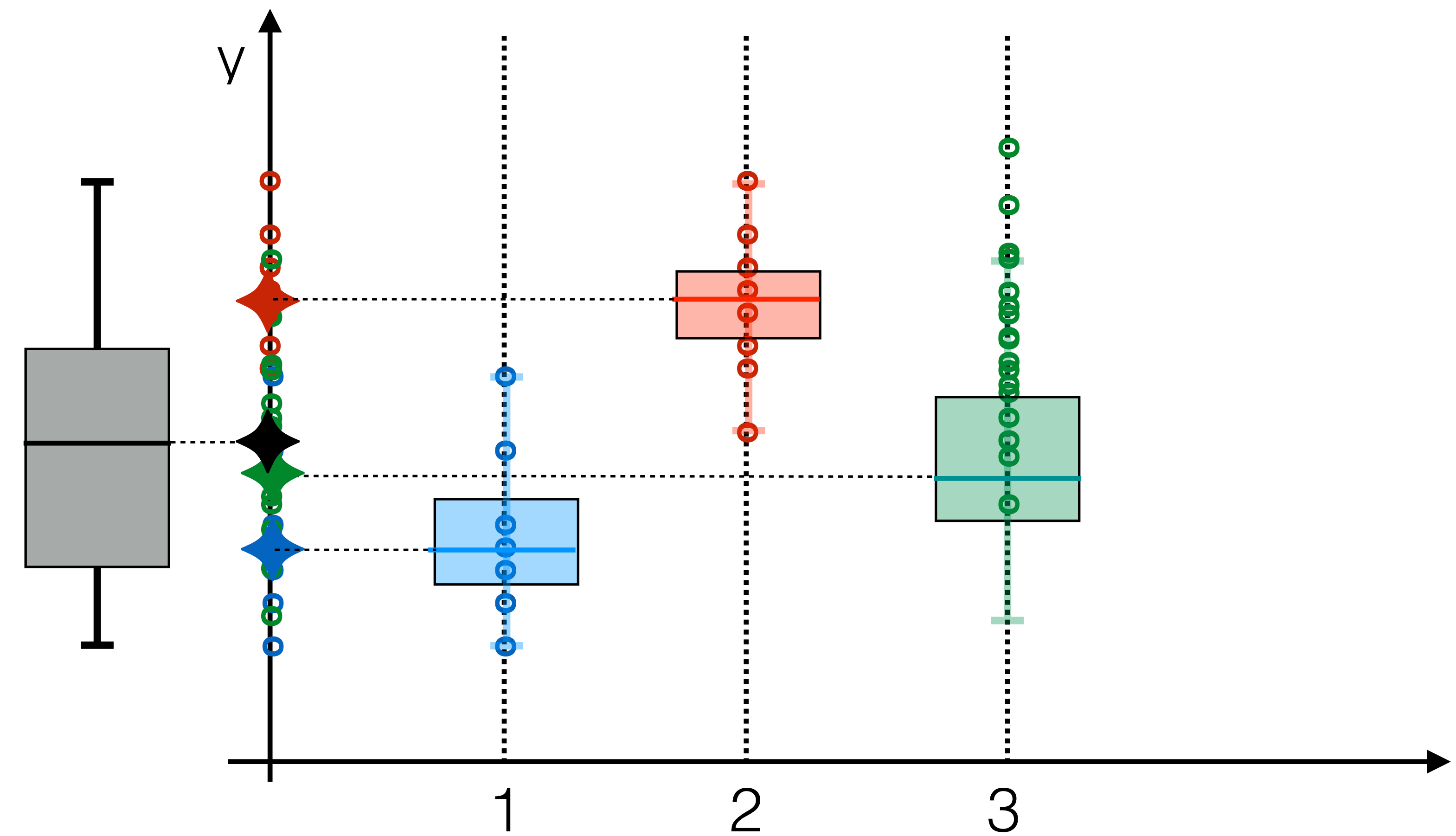
Jedna z možností je provést k(k-1)/2 porovnání pomocí dvouvýběrových testů.

ALE: tím se výrazně zvýší hladina významnosti.

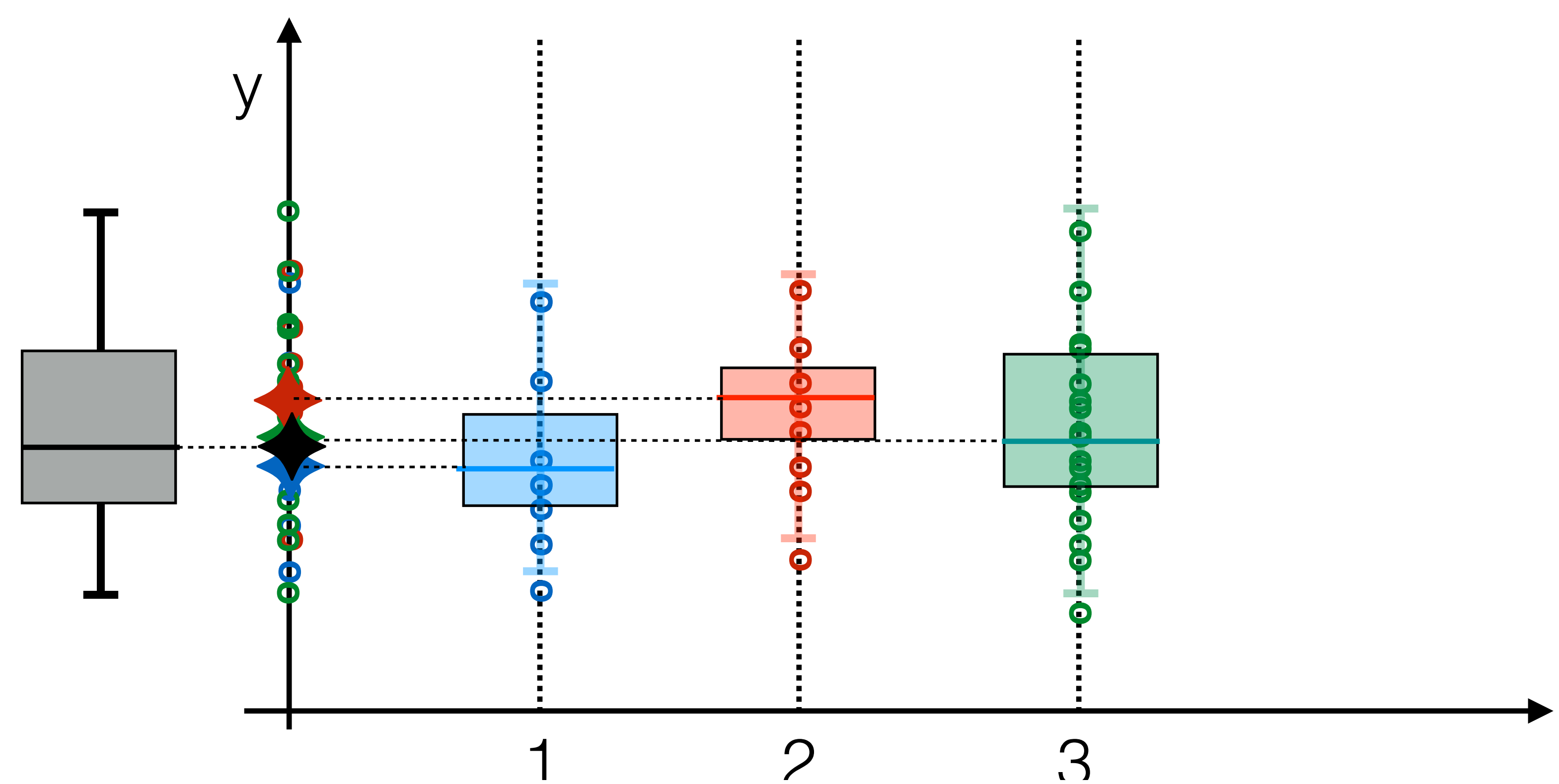
Tedy je třeba provádět tzv. simultánní test.

Metoda ANOVA pro srovnání více náhodných veličin

Rozptyl mezi skupinami je **velký** vzhledem k součtu rozptylů uvnitř skupin => střední hodnoty **jsou různé** => nulovou hypotézu **zamítáme**



Rozptyl mezi skupinami je **malý** vzhledem k součtu rozptylů uvnitř skupin => střední hodnoty lze považovat **za nerozlišitelné** => nulovou hypotézu **nelze zamítnout**



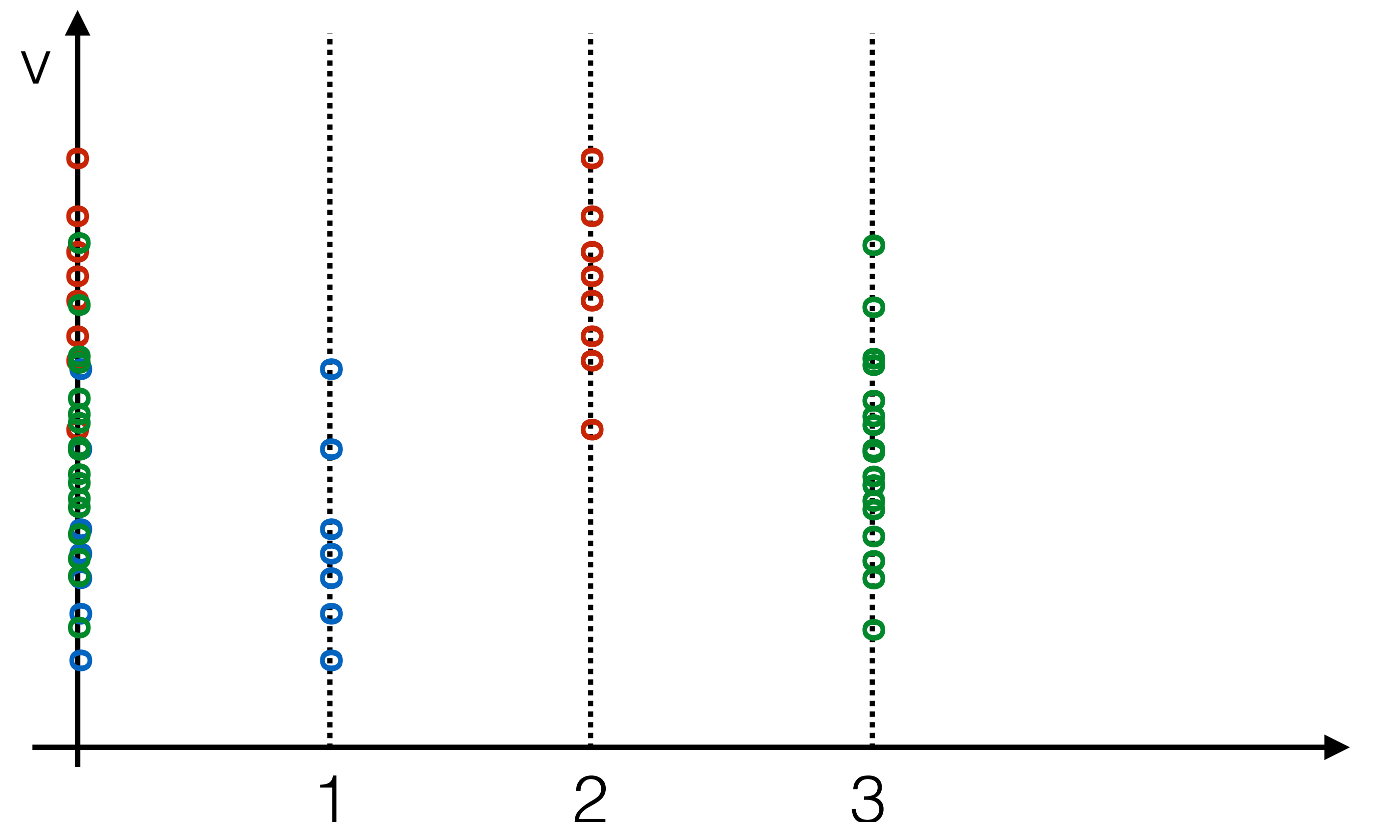
Metoda ANOVA pro srovnání více náhodných veličin

ANalysis Of VAriance = ANOVA

pro 1 faktor = všechna data rozdělíme do skupin podle jednoho hlediska (úrovní jednoho faktoru)

k úrovní = k skupin

v každé skupině n měření



1) celkový průměr všech naměřených hodnot: $\bar{y} = \frac{1}{n.k}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}$

2) průměry v jednotlivých skupinách: $\bar{y}_i = \frac{1}{n}\sum_{j=1}^{n} y_{ij}$

3) celkový součet čtvercových odchylek: $SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y})^2$

4) součet čtvercových odchylek uvnitř skupin $SS_E = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2$

5) součet čtvercových odchylek mezi skupinami $SS_F = n\sum_{i=1}^{k}(\bar{y}_i - \bar{y})^2$

Metoda ANOVA pro srovnání více náhodných veličin

Tabulka analýzy rozptylu:

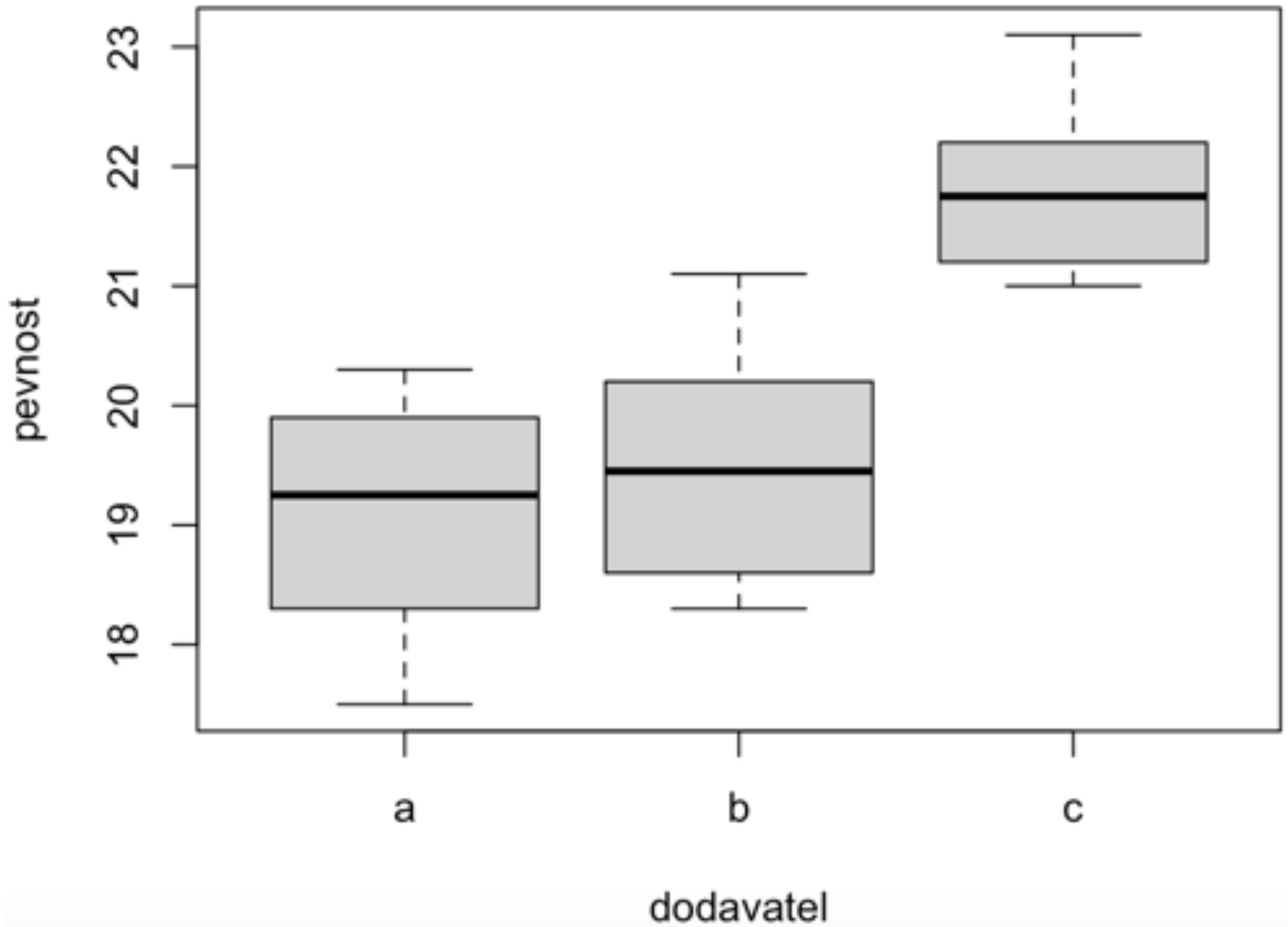
Zdroj variability	Součet čtverců	Stupně volnosti	Průměrný čtverec	Testová statistika	p-hodnota
faktor (mezi skupinami)	SS_F	k-1	$MS_F = \frac{SS_F}{k-1}$	$F = \frac{MS_F}{MS_E}$	
rezidua (uvnitř skupin)	SS_E	$k(n$-1)	$MS_E = \frac{SS_E}{k(n-1)}$		
celkový	SS_T	kn-1			

Testová statistika $F = \frac{\frac{SS_F}{k-1}}{\frac{SS_E}{k(n-1)}} = \frac{MS_F}{MS_E}$ má Fisherovo-Snedecorovo rozdělení o (k-1) a k(n-1) stupních volnosti.

Metoda ANOVA pro srovnání více náhodných veličin

Příklad: Chceme porovnat kvalitu vláken dodávaných třemi různými výrobci. Jako kritérium kvality je zvolena pevnost vláken v tahu.

$$X_1^1, X_2^1, \dots, X_{n_1}^1, \quad X_1^2, X_2^2, \dots, X_{n_2}^2, \quad X_1^3, X_2^3, \dots, X_{n_3}^3,$$




```
> pvl <- data.frame(read.table("pevnost_vlaken.txt", header=T))
> boxplot(pevnost~dodavatel, data=pvl)
```

Metoda ANOVA pro srovnání více náhodných veličin

Příklad: Chceme porovnat kvalitu vláken dodávaných třemi různými výrobci.
Jako kritérium kvality je zvolena pevnost vláken v tahu.

$$X_1^1, X_2^1, \ldots, X_{n_1}^1, \quad X_1^2, X_2^2, \ldots, X_{n_2}^2, \quad X_1^3, X_2^3, \ldots, X_{n_3}^3,$$

```
> oneway <- aov(pevnost~dodavatel, data=pvl)
> summary(oneway)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
dodavatel	2	26.23	13.117	13.62	0.000425 ***
Residuals	15	14.45	0.963		
Total	17	40.68			

—

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Závěr: nulovou hypotézu zamítáme. Pevnost vláken různých dodavatelů se statisticky významně liší.

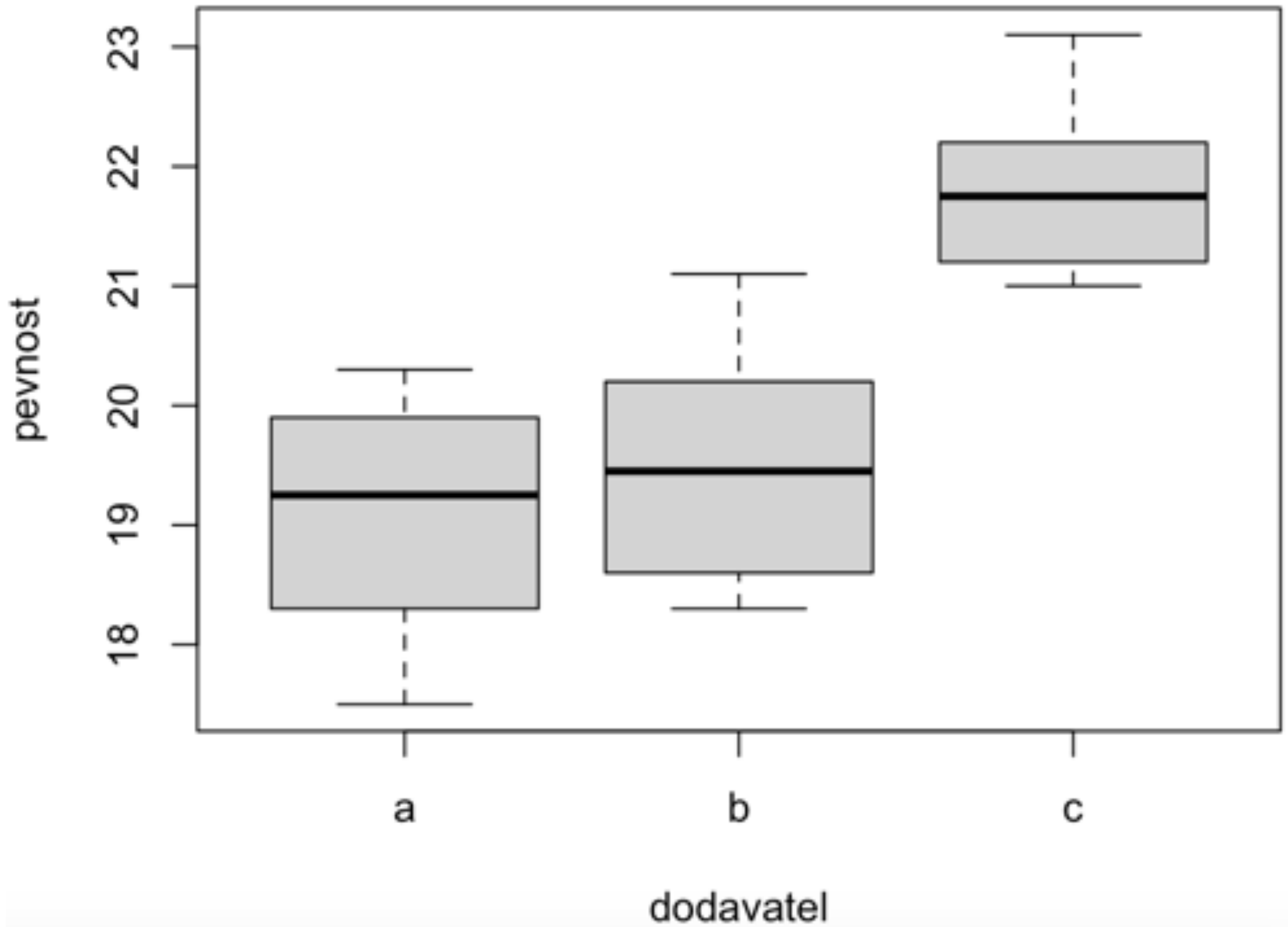
Jak se liší dodavatelé mezi sebou?

Pokud zamítneme nulovou hypotézu v ANOVA, měli bychom provést tzv. Mnohonásobné porovnání

Metoda ANOVA pro srovnání více náhodných veličin

Příklad: Chceme porovnat kvalitu vláken dodávaných třemi různými výrobci.
Jako kritérium kvality je zvolena pevnost vláken v tahu.

$$X_1^1, X_2^1, \ldots, X_{n_1}^1, \quad X_1^2, X_2^2, \ldots, X_{n_2}^2, \quad X_1^3, X_2^3, \ldots, X_{n_3}^3,$$



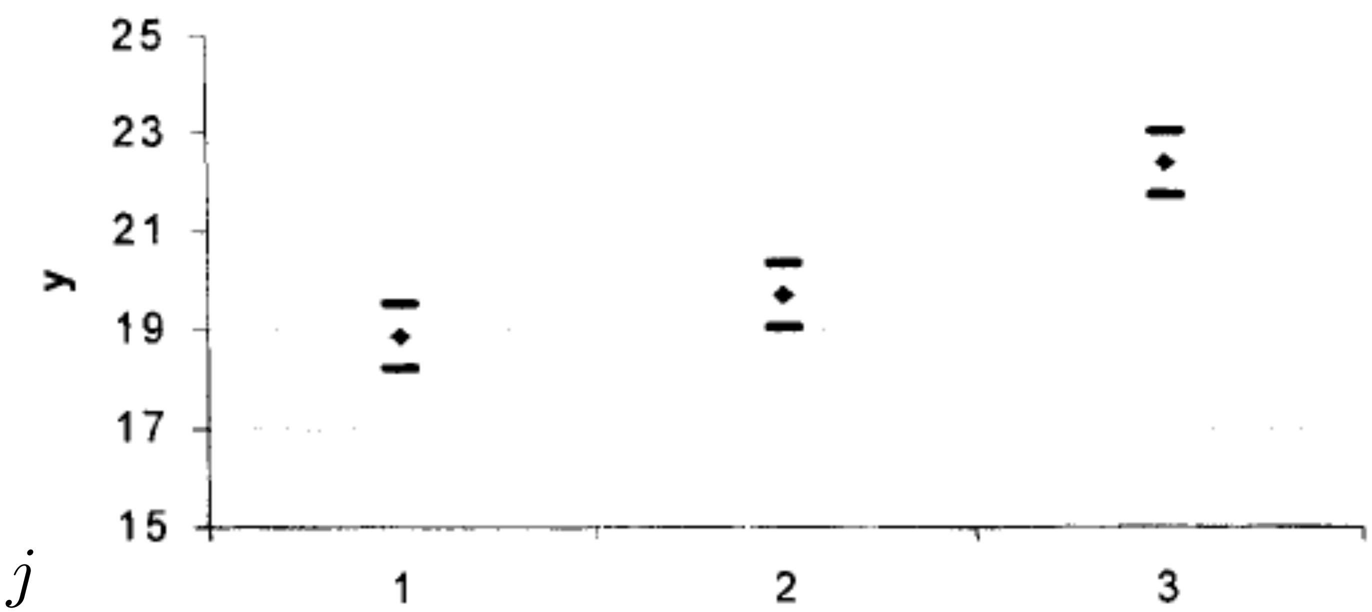
Jak se liší dodavatelé mezi sebou?

Pokud zamítneme nulovou hypotézu v ANOVA, měli bychom provést tzv. Mnohonásobné porovnání

Metoda ANOVA pro srovnání více náhodných veličin

Metody mnohonásobného srovnávání:

- Tukeyova metoda
- Schefféo metoda
- Bonferroniho metoda
- Duncanova metoda
- …



$$T = \frac{|\hat{\mu}_i - \hat{\mu}_j|}{S*}, \quad i, j = 1, 2, \ldots, k, \; i \neq j$$

> TukeyHSD(oneway, conf.level = .95)

Dodavatele 1 a 2 nelze statisticky odlišit, dodavatel 3 je významně lepší (jeho vlákna dosahují statisticky významně vyšší pevnosti než od dodavatelů 1 a 2).

Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = pevnost ~ dodavatel, data = pvl)

$dodavatel

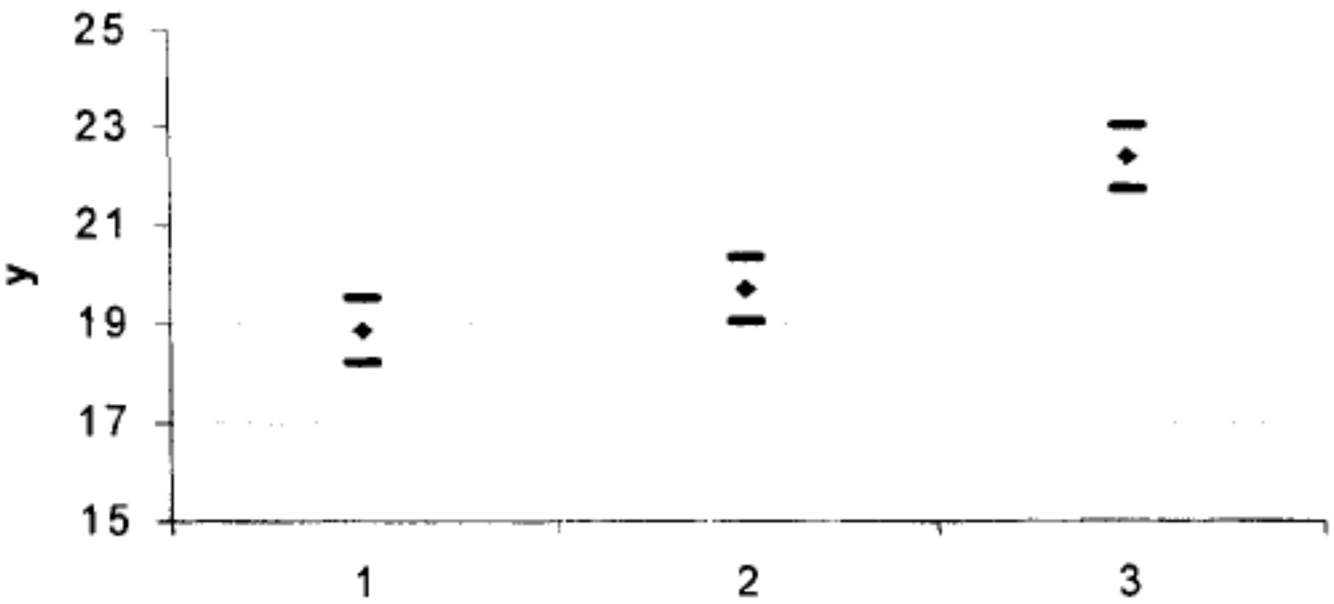
=> z tabulky *p*-hodnot při porovnání po dvojicích vyplývá, že

	diff	lwr	upr	p adj	
b-a	0.4333333	-1.0385665	1.905233	0.7296030	- a ~ b jsou nerozlišitelné
c-a	2.7500000	1.2781002	4.221900	0.0005817	- c ~ a se liší
c-b	2.3166667	0.8447669	3.788566	0.0026211	- c ~ b se liší

Metoda ANOVA pro srovnání více náhodných veličin

Metody mnohonásobného srovnávání:

- Tukeyova metoda
- Schefféo metoda
- Bonferroniho metoda
- Duncanova metoda
- …



$$T = \frac{|\hat{\mu}_i - \hat{\mu}_j|}{S*}, \quad i, j = 1, 2, \ldots, k, \; i \neq j$$

```
> pairwise.t.test(pvl$pevnost, pvl$dodavatel, p.adj = "bonf")

    Pairwise comparisons using t tests with pooled SD

data:  pvl$pevnost and pvl$dodavatel

  a       b
b 1.00000 -
c 0.00063 0.00291

P value adjustment method: bonferroni
```

=> z tabulky *p*-hodnot při porovnání po dvojicích vyplývá, že

- a s b jsou nerozlišitelné
- c se liší jak od a, tak i od b

Metoda ANOVA pro srovnání více náhodných veličin

Metody mnohonásobného srovnávání:

označme k počet tříd a $N = \sum_{i=1}^{k} n_i$

- Tukeyova metoda

- musí být $n_1 = n_2 = \ldots = n_k$

$$Q = \frac{|\hat{\mu}_i - \hat{\mu}_j|}{S*}, \quad S* = \sqrt{\frac{S_E}{n(N-k)}}$$

$$Q < q_{1-\alpha}(k,\ N-k)$$

- Scheffého metoda

$$S = \frac{|\hat{\mu}_i - \hat{\mu}_j|}{S*}, \quad S* = \sqrt{\frac{S_E}{N-k}(k-1)\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

$$S < \sqrt{F_{1-\alpha}(k-1,\ N-k)}$$

- Bonferroniho metoda

$$T = \frac{|\hat{\mu}_i - \hat{\mu}_j|}{S*}, \quad S* = \sqrt{\frac{S_E}{N-k}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

$$T < t_{1-\alpha/2}(N-k)$$

Metoda ANOVA pro srovnání více náhodných veličin

Ověření podmínek pro použití ANOVA

- Reprezentativnost výběru, náhodnost
- Nezávislost pozorování
- Normalita dat
- Stejné rozptyly (homoskedasticita)

https://is.muni.cz/th/151390/prif_m/diplomova_prace_ed.pdf

Vícenásobná ANOVA

One-way ANOVA

- Zatím jsme měli datový soubor, který bylo možné roztřídit do několika skupin podle hodnoty jednoho faktoru jednoduché třídění

Příklad: pevnost vláken ~ dodavatel (*a, b, c)*

výkon ~ stroj (A1, A2)

	pevnost	dodavatel
1	18.3	a
2	17.5	a
3	20.3	a
4	19.9	a
5	19.8	a
6	18.7	a
7	18.3	b
8	18.6	b
9	19.2	b
10	19.7	b
11	21.1	b
12	20.2	b
13	21.0	c
14	21.4	c
15	21.2	c
16	22.2	c
17	23.1	c
18	22.1	c

- V mnoha případech se data dají roztřídit podle více faktorů
Multi-way ANOVA vícenásobné třídění

Příklad: výnos plodiny ~ druh hnojiva + typ půdy

	Výnos kukuřice na parcele číslo							
varianta	1	2	3	4	5	6	7	8
V1	1,29	1,19	1,23	1,33	1,27	1,29	1,31	1,2
V2	1,3	1,33	1,29	1,37	1,35	1,25	1,38	1,29
V3	1,2	1,24	1,25	1,24	1,2	1,21	1,28	1,17
V4	1,03	1,14	1,09	1,2	1,07	1,19	1,01	1,05

Vícenásobná ANOVA

Příklad: výnos plodiny ~ druh hnojiva + typ půdy

	Výnos kukuřice na parcele číslo								
varianta	1	2	3	4	5	6	7	8	průměry
V1	1,29	1,19	1,23	1,33	1,27	1,29	1,31	1,2	1,26375
V2	1,3	1,33	1,29	1,37	1,35	1,25	1,38	1,29	1,32
V3	1,2	1,24	1,25	1,24	1,2	1,21	1,28	1,17	1,22375
V4	1,03	1,14	1,09	1,2	1,07	1,19	1,01	1,05	1,0975
průměry	1,205	1,225	1,215	1,285	1,2225	1,235	1,245	1,1775	1,22625

Je výnos kukuřice závislý na variantě, typu půdy nebo na obojím?

32 naměřených hodnot můžeme rozdělit
- do 4 skupin po 8 podle varianty
- do 8 skupin po 4 podle parcely

dvojné třídění

$$X = \mu_0 + \alpha_1 V_1 + \alpha_2 V_2 + \alpha_3 V_3 + \alpha_4 V_4 + \epsilon$$

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$$
$$H_A : \exists i : \alpha_i \neq 0$$

$$H_0 : \mu_{V_1} = \mu_{V_2} = \mu_{V_3} = \mu_{V_4} = \mu_0$$
$$H_A : \exists i, j : \mu_{V_i} \neq \mu_{V_j}$$

model1 = vynos ~ varianta

Vícenásobná ANOVA

Příklad: výnos plodiny ~ druh hnojiva + typ půdy

	Výnos kukuřice na parcele číslo								
varianta	1	2	3	4	5	6	7	8	průměry
V1	1,29	1,19	1,23	1,33	1,27	1,29	1,31	1,2	1,26375
V2	1,3	1,33	1,29	1,37	1,35	1,25	1,38	1,29	1,32
V3	1,2	1,24	1,25	1,24	1,2	1,21	1,28	1,17	1,22375
V4	1,03	1,14	1,09	1,2	1,07	1,19	1,01	1,05	1,0975
průměry	1,205	1,225	1,215	1,285	1,2225	1,235	1,245	1,1775	1,22625

Je výnos kukuřice závislý na variantě, typu půdy nebo na obojím?

32 naměřených hodnot můžeme rozdělit
- do 4 skupin po 8 podle varianty
- do 8 skupin po 4 podle parcely

dvojné třídění

$$X = \mu_0 + \beta_1 P_1 + \beta_2 P_2 + \beta_3 P_3 + \beta_4 P_4 + \beta_5 P_5 + \beta_6 P_6 + \beta_7 P_7 + \beta_8 P_8 + \epsilon$$

$$H_0 : \beta_1 = \cdots = \beta_8 = 0 \qquad H_0 : \mu_{P_1} = \cdots = \mu_{P_8} = \mu_0$$

$$H_A : \exists j : \beta_j \neq 0 \qquad H_A : \exists i, j : \mu_{P_i} \neq \mu_{P_j}$$

model2 = vynos ~ parcela

Vícenásobná ANOVA

Příklad: výnos plodiny ~ druh hnojiva + typ půdy

	Výnos kukuřice na parcele číslo								
varianta	1	2	3	4	5	6	7	8	průměry
V1	1,29	1,19	1,23	1,33	1,27	1,29	1,31	1,2	1,26375
V2	1,3	1,33	1,29	1,37	1,35	1,25	1,38	1,29	1,32
V3	1,2	1,24	1,25	1,24	1,2	1,21	1,28	1,17	1,22375
V4	1,03	1,14	1,09	1,2	1,07	1,19	1,01	1,05	1,0975
průměry	1,205	1,225	1,215	1,285	1,2225	1,235	1,245	1,1775	1,22625

Je výnos kukuřice závislý na variantě, typu půdy nebo na obojím?

32 naměřených hodnot můžeme rozdělit
- do 4 skupin po 8 podle varianty
- do 8 skupin po 4 podle parcely

dvojné třídění

$$X = \mu_0 + \sum_{i=1}^{4} \alpha_i V_i + \sum_{j=1}^{8} \beta_j P_j$$

model3 = vynos ~ varianta + parcela

model4 = vynos ~ varianta * parcela

Vícenásobná ANOVA

Příklad: výnos plodiny ~ druh hnojiva + typ půdy

```
> vynosy<- data.frame(read.table("vynosy.txt", header=T))
```

$$X = \mu_0 + \alpha_1 V_1 + \alpha_2 V_2 + \alpha_3 V_3 + \alpha_4 V_4 + \epsilon$$

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$$

$$H_A : \exists i : \alpha_i \neq 0$$

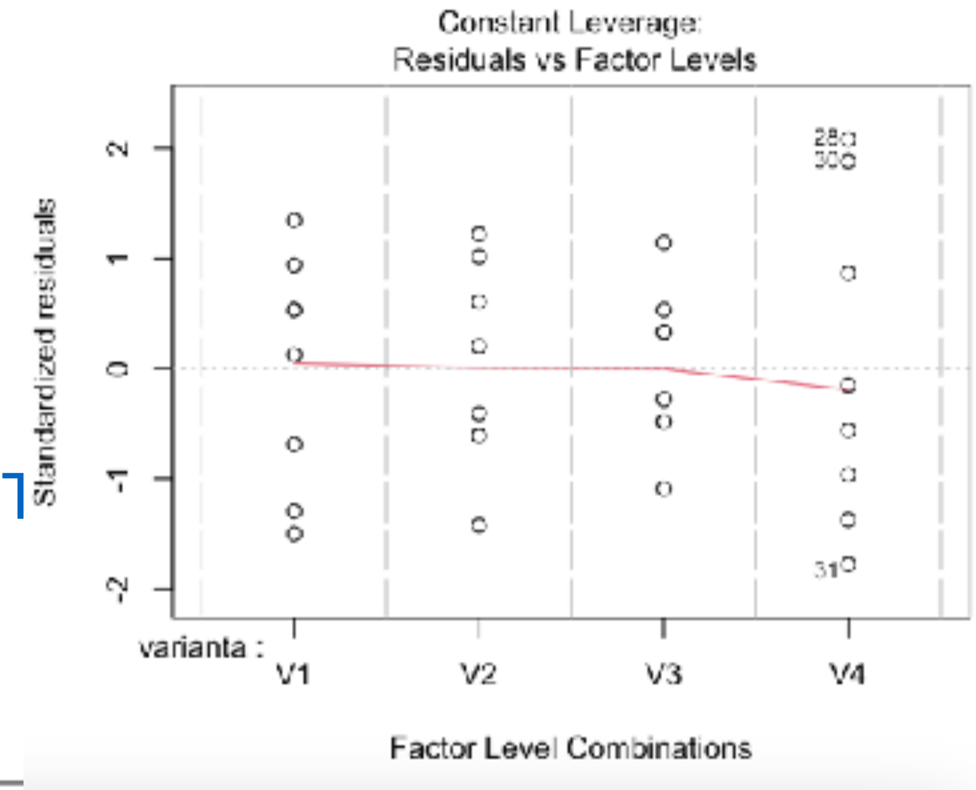
```
> oneway1<- aov(vynos~varianta, data=vynosy)
> summary(oneway1)
            Df  Sum Sq   Mean Sq  F value   Pr(>F)
varianta     3  0.21423  0.07141    25.79   3.29e-08 ***
Residuals   28  0.07752  0.00277
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> plot(oneway1)
```

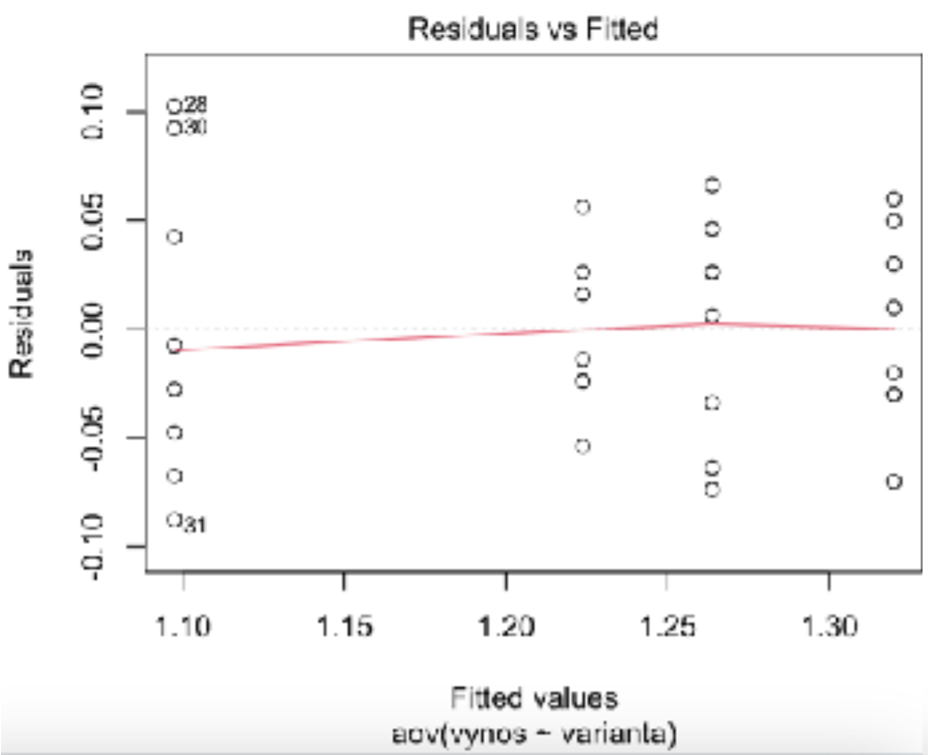
Vícenásobná ANOVA

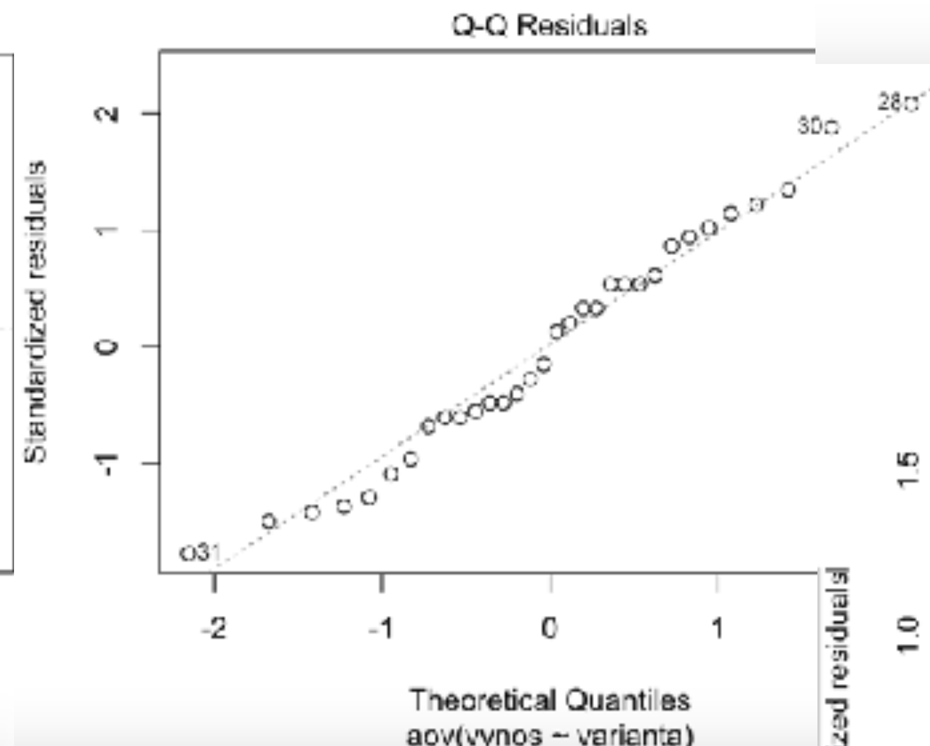
Příklad: výnos plodiny ~ druh hnojiva + typ půdy

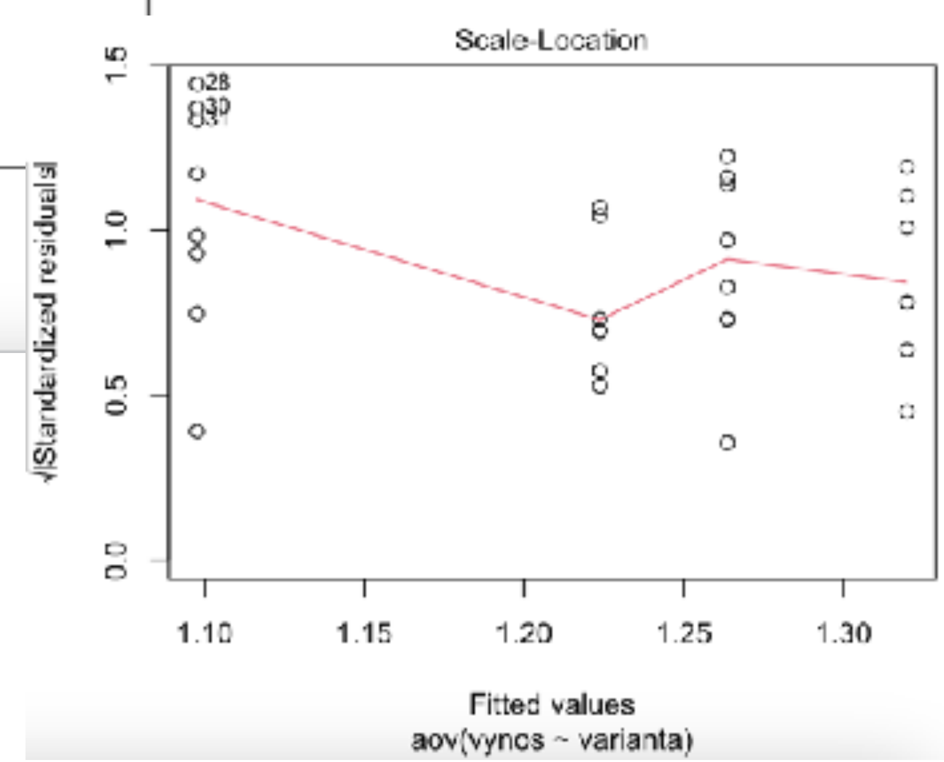
> vynosy<- data.frame(read.table("vynosy.txt", header=T

$$X = \mu_0 + \alpha_1 V_1 + \alpha_2 V_2 + \alpha_3 V_3 + \alpha_4 V_4 + \epsilon$$









> plot(oneway1)

Vícenásobná ANOVA

Příklad: výnos plodiny ~ druh hnojiva + typ půdy

$$X = \mu_0 + \beta_1 P_1 + \beta_2 P_2 + \beta_3 P_3 + \beta_4 P_4 + \beta_5 P_5 + \beta_6 P_6 + \beta_7 P_7 + \beta_8 P_8 + \epsilon$$

$$H_0 : \beta_1 = \cdots = \beta_8 = 0$$

$$H_A : \exists j : \beta_j \neq 0$$

```
> oneway2<- aov(vynos~parcela, data=vynosy)
> summary(oneway2)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
parcela	7	0.0274	0.003914	0.355	0.919
Residuals	24	0.2643	0.011015		

```
> plot(oneway2)
```

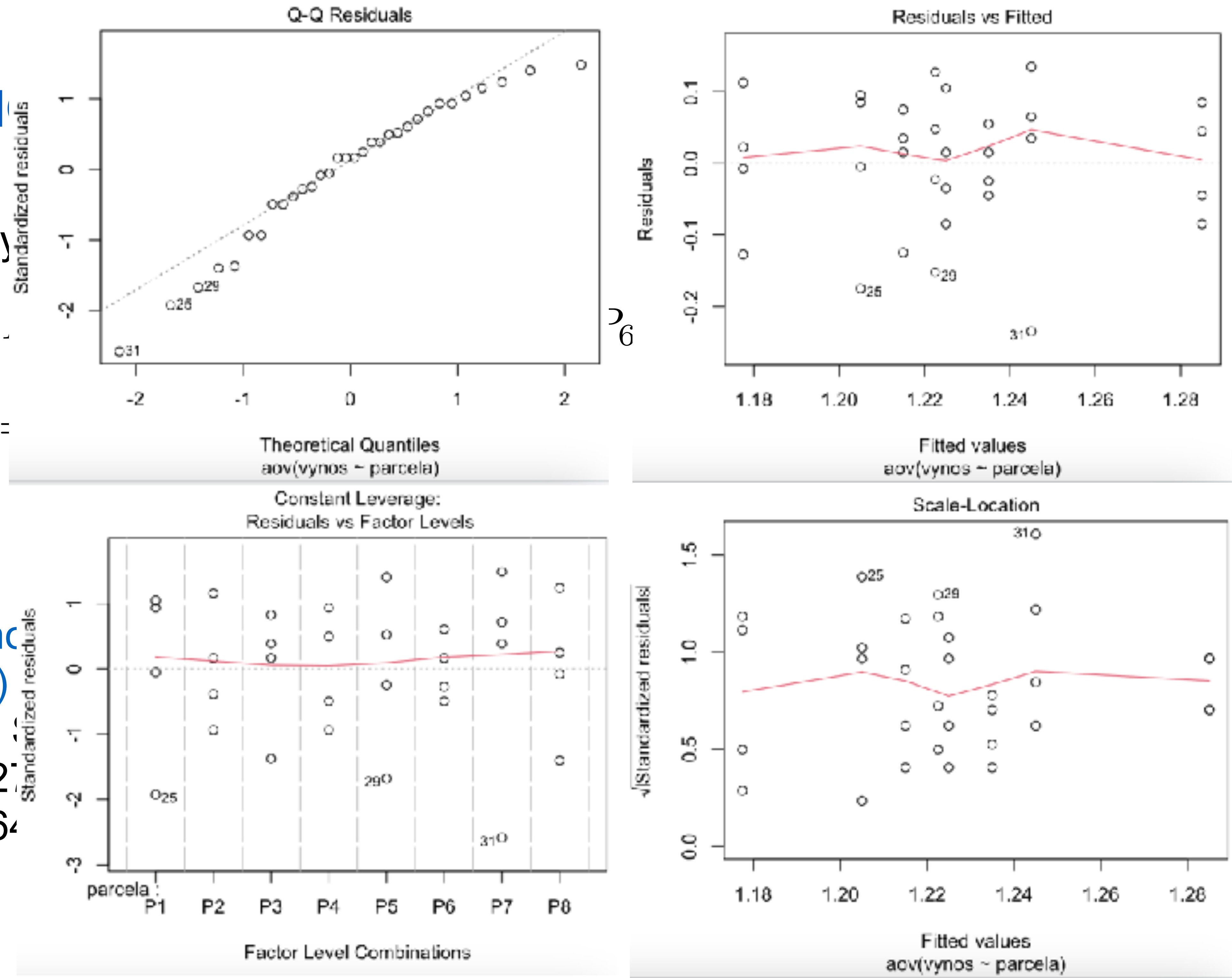
Vícenásobná AN

Příklad: výnos plodiny

$$X = \mu_0 + \beta_1 P_1 + \beta_2 \ldots P_6$$

$$H_0 : \beta_1 = \cdots = \beta_8 =$$

$$H_A : \exists j : \beta_j \neq 0$$




```
> oneway2<- aov(vyno
> summary(oneway2)
              Df   Sum 
parcela        7   0.027
Residuals     24   0.264
```

```
> plot(oneway2)
```

```
> shapiro.test(oneway2$residuals)
```

Shapiro-Wilk normality test

data: oneway2$residuals
W = 0.95112, p-value = 0.155

Vícenásobná ANOVA

Příklad: výnos plodiny ~ druh hnojiva + typ půdy

$$X = \mu_0 + \sum_{i=1}^{4} \alpha_i V_i + \sum_{j=1}^{8} \beta_j P_j$$

```
> twoway1 <- aov(vynos~vynos + parcela, data=vynosy)
> summary(twoway1)
             Df   Sum Sq   Mean Sq   F value    Pr(>F)
varianta      3   0.21423   0.07141    29.92    9e-08 ***
parcela       7   0.02740   0.00391     1.64    0.179
Residuals    21   0.05012   0.00239
```

```
> twoway2 <- aov(vynos~vynos*parcela, data=vynosy)
> summary(twoway2)
                   Df   Sum Sq   Mean Sq
varianta            3   0.21423  0.07141
parcela             7   0.02740  0.00391
varianta:parcela   21   0.05012  0.00239
```

Neparametrická analýza rozptylu

Odezva $X_1, X_2, \ldots, X_n$ se neřídí normálním rozdělením => nelze použít ANOVA.

Kruskal – Wallisův test

- Je neparametrickou obdobou analýzy rozptylu jednoduchého třídění.
- Slouží k ověření nulové hypotézy H_0, že $k > 2$ nezávislých náhodných výběrů o rozsazích $n_1, n_2, \ldots, n_k$ pochází z jednoho základního souboru.
- Předpokládáme, že tyto náhodné výběry byly pořízeny ze základních souborů se spojitými distribučními funkcemi $F_1(x), F_2(x), \ldots, F_k(x)$.
- Nulovou hypotézu H_0 můžeme zapsat takto

$$H_0: F_1(x) = F_2(x) = \ldots = F_k(x) \text{ pro všechna } x.$$

Postup při stanovení testového kritéria:

1) Máme k dispozici $\boldsymbol{k}$ výběrových souborů o četnostech $n_1, n_2, \ldots, n_k$.

2) Všechny výběrové soubory sloučíme do jediného souboru.

3) Každé hodnotě souboru přiřadíme vzestupně pořadové číslo, stejným hodnotám pak pořadí průměrné.

4) Následně sečteme pořadová čísla jednotlivých pozorování pro každý původní výběrový soubor zvlášť a získáme součty $T_1, T_2, \ldots, T_k$ (T_i ; $i = 1, \ldots, k$, je tedy součet pořadových čísel pro i-tý výběr)

Neparametrická analýza rozptylu

Odezva $X_1, X_2, \ldots, X_n$ se neřídí normálním rozdělením => nelze použít ANOVA.

Kruskal – Wallisův test

Testová statistika má tvar:
kde $n = n_1 + n_2 + \ldots + n_k$.

$$KW = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n+1)$$

Statistika KW má za platnosti H_0 při $n_i \to \infty$ asymptoticky χ^2 – rozdělení o k-1 stupních volnosti.

Pokud $KW > \chi^2_\alpha(n-1)$, přijímáme hypotézu alternativní, podle které se hodnoty nejméně dvou porovnávaných výběrových souborů od sebe statisticky významně liší.

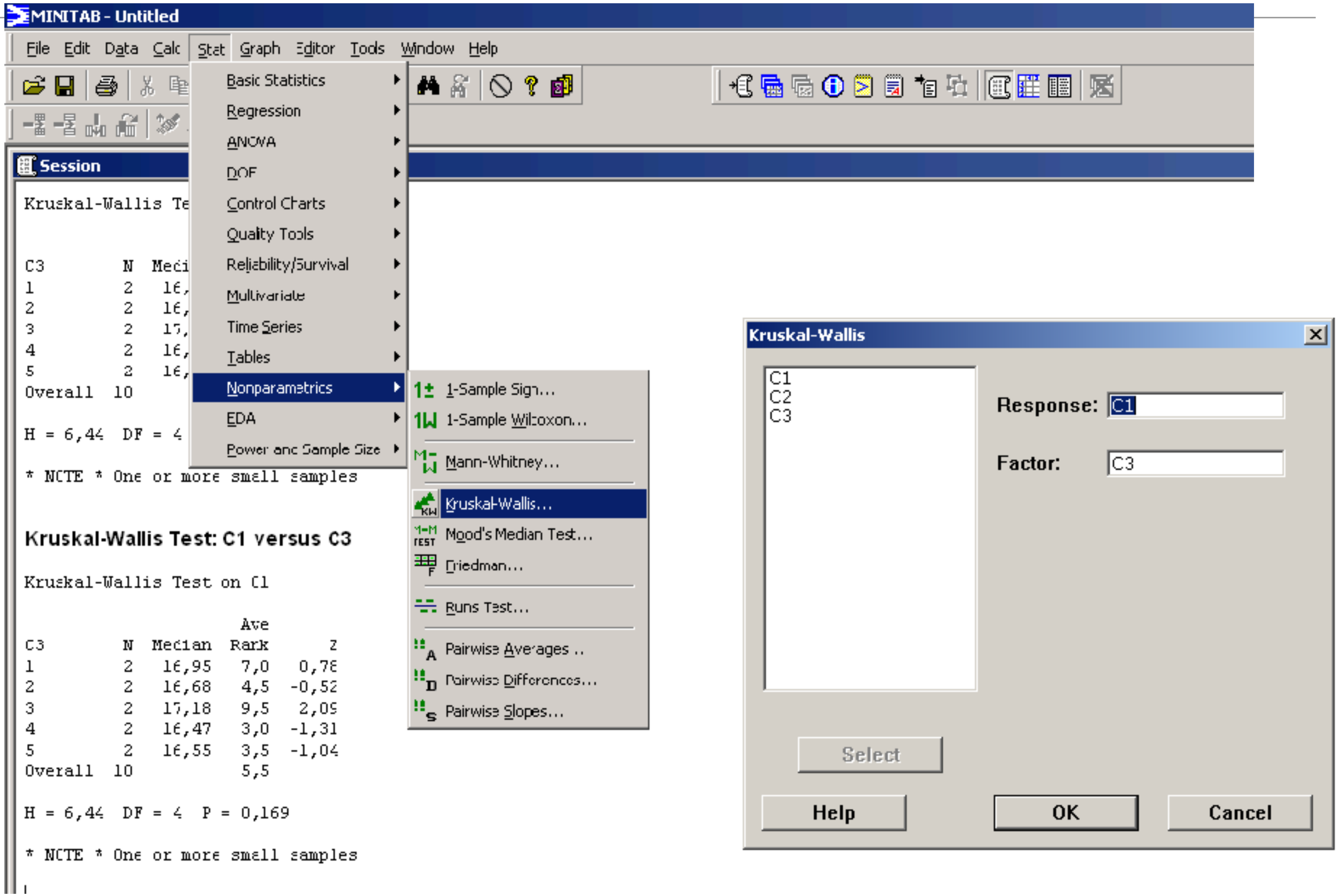
Jestliže se v posloupnosti zjištěných údajů vyskytnou shodné hodnoty, kterým se přiřazuje průměrné pořadí, je nutno hodnotu KW dělit korekčním faktorem

$$K = 1 - \frac{1}{n^3 - n} \sum_{j=1}^{p} (t_j^3 - t_j)$$

kde p je počet tříd se stejným pořadím a t_i počet pořadí v i-té třídě.

Opravené testové kritérium se stanoví jako $KW_{opr} = \dfrac{KW}{K}$

Neparametrická analýza rozptylu



Neparametrická analýza rozptylu

Neparametrické metody mnohonásobného porovnávání.

Neményiho metoda

- Předpokládáme tzv. „*vyvážený pokusný plán*“, tzn., že všech $\boldsymbol{k}$ výběrů má stejný rozsah (tedy $n_1 = n_2 = \ldots = n_k = N$).
- Spočítáme $\frac{k(k-1)}{2}$ diferencí $|T_i - T_j|$ které porovnáváme s kritickými hodnotami.
- Kritické hodnoty D_α se hledají pro hladinu významnosti α, pro k porovnávaných tříd a N opakování v každé třídě ($n_1 = n_2 = \ldots = n_k = N$).
- Je-li nějaká diference větší nebo rovna kritické hodnotě D_α pro Neményiho metodu, zamítá se hypotéza o neprůkaznosti diference, tzn. že i-tý a j-tý výběr pocházejí z téhož rozdělení.

Neparametrická analýza rozptylu

Neparametrické metody mnohonásobného porovnávání.

Dunnova metoda

- Použijeme ji v případě „ne*vyváženého pokusného plánu*“, tzn., že výběry nemají stejný rozsah (tedy nemusí být $n_1 = n_2 = \ldots = n_k$).
- Spočítáme $\frac{k(k-1)}{2}$ diferencí $|T_i - T_j|$ které porovnáváme s kritickými hodnotami.
- Pokud pro nějakou diferenci platí $|T_i - T_j| > u_\beta \sqrt{\frac{n(n-1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$

  kde u_β je β- kvantil standardního normálního rozdělení, $\beta = \frac{\alpha}{k(k-1)}$,

  zamítáme na hladině významnosti α hypotézu, že i-tý výběr (s rozsahem n_i) a j-tý

  výběr (s rozsahem n_j) pocházejí z téhož rozdělení.
-

Neparametrická analýza rozptylu

Příklad: pevnost vláken ~ dodavatel (*a, b, c)*

	pevnost	dodavatel
1	18.3	a
2	17.5	a
3	20.3	a
4	19.9	a
5	19.8	a
6	18.7	a
7	18.3	b
8	18.6	b
9	19.2	b
10	19.7	b
11	21.1	b
12	20.2	b
13	21.0	c
14	21.4	c
15	21.2	c
16	22.2	c
17	23.1	c
18	22.1	c

```
> kruskal.test(pevnost ~ dodavatel, data = pevnost)

	Kruskal-Wallis rank sum test

data:  pevnost by dodavatel
Kruskal-Wallis chi-squared = 10.862, df = 2, p-value = 0.004379

> library(DescTools)
> NemenyiTest(pevnost ~ dodavatel, data = pevnost)

 Nemenyi's test of multiple comparisons for independent samples (tukey)

    mean.rank.diff   pval
b-a              0 1.0000
c-a              9 0.0098 **
c-b              9 0.0098 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Neparametrická analýza rozptylu

Odezva $X_1, X_2, \ldots, X_n$ se neřídí normálním rozdělením => nelze použít ANOVA.

Friedmanův test

- Je neparametrickou obdobou analýzy rozptylu dvojného třídění (pro dva faktory - například pro jeden hlavní a druhý blokový).
- Předpokládejme, že máme měření y_{ij}, kde i je označení bloku, j je označení úrovně faktoru (ošetření). Pro každou dvojici (i,j) máme jedno měření.
- Označme

  k - počet úrovní faktoru

  n - počet bloků

  r_{ij} - pořadí naměřené hodnoty odezvy y_{ij} v bloku i

  $\bar{r}_{.j}$ - průměrné pořadí odezvy při úrovni j přes všechny bloky $\quad \bar{r}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} r_{ij}$

  $\bar{r}$ - průměrné pořadí přes všechny úrovně a přes všechny bloky $\quad \bar{r} = \frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij}$

Položme $SS_t = n\sum_{j=1}^{k}(\bar{r}_{\cdot j} - \bar{r})^2$ a $SS_e = \frac{1}{n(k-1)}\sum_{i=1}^{n}\sum_{j=1}^{k}(r_{ij} - \bar{r})^2$

Testová statistika má tvar $Q = \frac{SS_t}{SS_e}$

(Pro velká n a k ($n > 15$, $k > 4$) má přibližně chí-kvadrát rozdělení o k-1 stupních volnosti)

Neparametrická analýza rozptylu