# Přednáška 8

## Vícevýběrové testy parametrických hypotéz

— testy shody rozptylů (Bartletův test, Hartleyův test, Cochranův test, Leveneův test)

— analýza rozptylu (test shody středních hodnot)

  —post hoc analýza pro analýzu rozptylu

— Kruskal-Wallisův test (test shody mediánů)

  —post hoc analýza pro Kruskal-Wallisův test

# Používané značení

| Číslo skupiny | 1 | 2 | … | $k$ |
|---|---|---|---|---|
| Náhodný výběr | $X_{11}$<br>$\vdots$<br>$X_{1n_1}$ | $X_{21}$<br>$\vdots$<br>$X_{2n_2}$ | $\vdots$ | $X_{k1}$<br>$\vdots$<br>$X_{kn_k}$ |
| Rozsah skupiny | $n_1$ | $n_2$ | | $n_k$ |
| Průměr skupiny | $\bar{X}_1$ | $\bar{X}_2$ | | $\bar{X}_k$ |
| Rozptyl skupiny | $s_1^2$ | $s_2^2$ | | $s_k^2$ |

celkový rozsah všech *k* výběrů: $n = \sum_{i=1}^{k} n_i$,

průměr *i*-tého výběru (angl. population means): $\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$,

celkový průměr všech *k* výběrů (angl. grand mean): $\bar{\bar{X}} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}$,

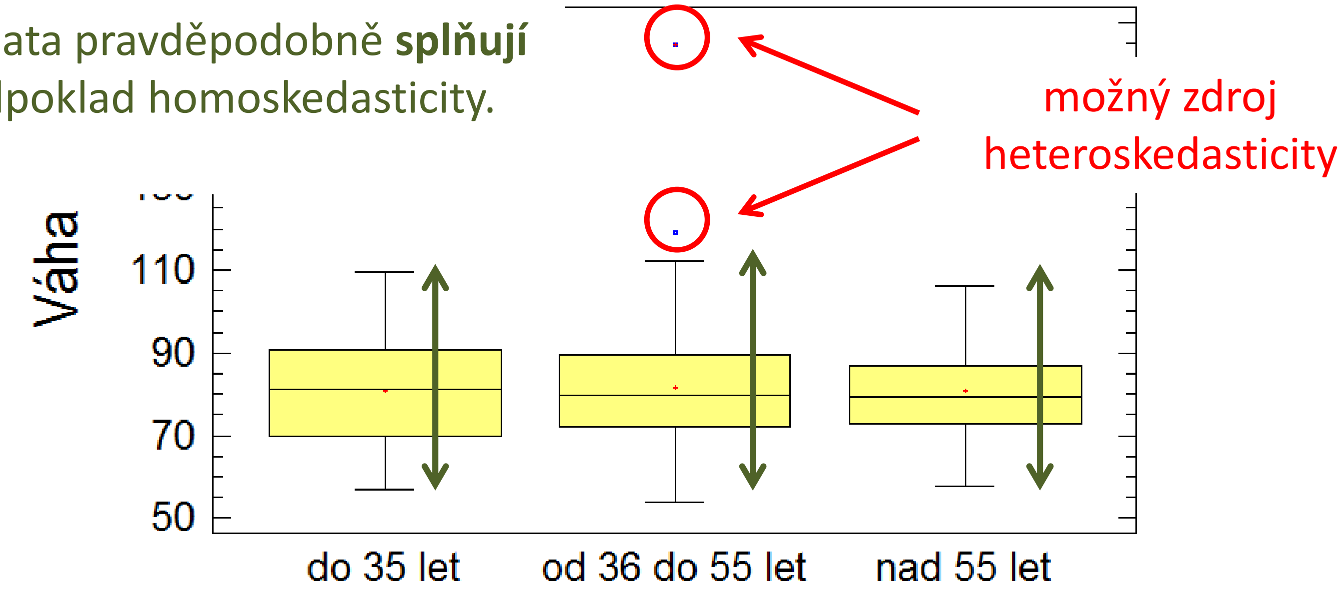
výběrový rozptyl *i*-tého výběru: $s_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}_i\right)^2$.

# Testy shody rozptylů

# Testy shody rozptylů

- Homoskedasticita (shoda rozptylů) je častým předpokladem testů o shodě středních hodnot.
- Jak posoudit homoskedasticitu pomocí grafů exploratorní analýzy?

Tato data pravděpodobně **splňují** předpoklad homoskedasticity.



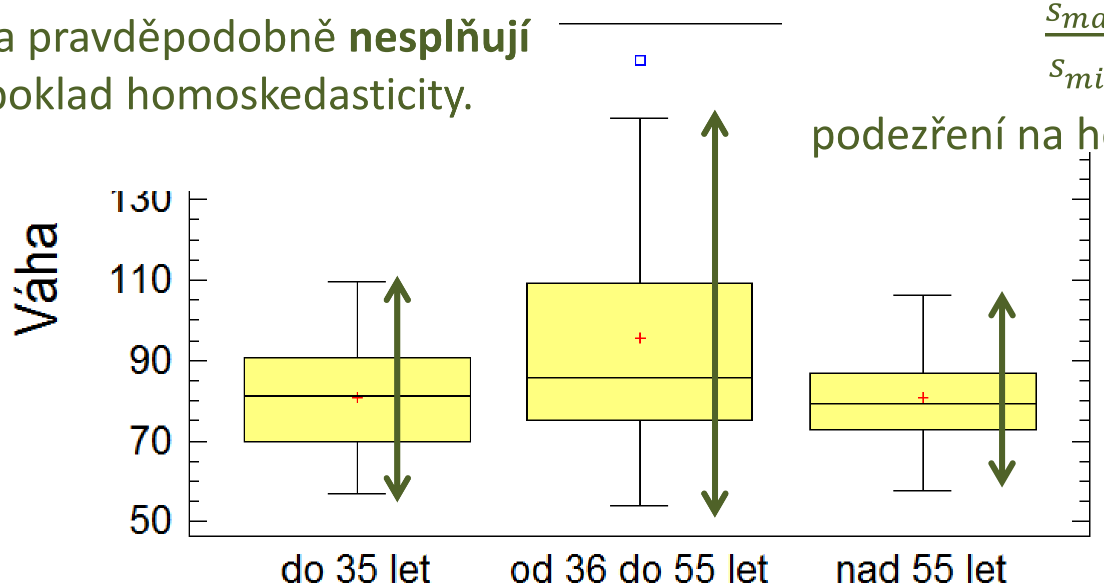
# Testy shody rozptylů

- Homoskedasticita (shoda rozptylů) je častým předpokladem testů o shodě středních hodnot.
- Jak posoudit homoskedasticitu pomocí grafů exploratorní analýzy?

Tato data pravděpodobně **nesplňují** předpoklad homoskedasticity.

$$\frac{s_{max}}{s_{min}} > 2$$

podezření na heteroskedasticitu



# Bartletův test

Předpokládejme, že máme $k > 2$ **nezávislých** výběrů z **normálního rozdělení**.

Nulová a alternativní hypotéza:

$H_0$: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$,   $H_A$: $\overline{H_0}$ (alespoň jedna dvojice rozptylů se liší)

Testová statistika:

$$B = \frac{1}{C}\left[(n-k)\ln MS_e - \sum_{i=1}^{k}(n_i - 1)\ln s_i^2\right],$$

kde $MS_e = \frac{1}{n-k}\sum_{i=1}^{k}(n_i - 1)s_i^2$ (MSE… reziduální rozptyl), $C = 1 - \frac{1}{3(k-1)}\left(\frac{1}{n-k} - \sum_{i=1}^{k}\frac{1}{n_i-1}\right)$.

p-hodnota: $p\text{–}hodnota = 1 - F_0(x_{OBS})$, kde $F_0(x)$ je distr. f-ce rozdělení $\chi^2$ s $n-k$ stupni volnosti.

# Hartleyův test

Předpokládejme, že máme $k > 2$ **nezávislých** výběrů **stejného rozsahu** z **normálního rozdělení**.

Nulová a alternativní hypotéza:

$H_0$: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$,   $H_A$: $\neg H_0$ (alespoň jedna dvojice rozptylů se liší)

Testová statistika:

$$F_{max} = \frac{\max s_i^2}{\min s_i^2}$$

Nulová hypotéza se zamítá, je-li pozorovaná hodnota $F_{max}$ větší nebo rovna kritické hodnotě $h_\alpha(k, n_1 - 1)$, která je tabelována ve speciálních tabulkách (tabulka T8).

# Cochranův test

Předpokládejme, že máme $k > 2$ **nezávislých** výběrů **stejného rozsahu** z **normálního rozdělení**.

Nulová a alternativní hypotéza:

H$_0$: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$,   H$_A$: $\neg H_0$ (alespoň jedna dvojice rozptylů se liší)

Testová statistika:

$$G_{max} = \frac{\max s_i^2}{s_1^2 + \cdots + s_k^2}.$$

K zamítnutí nulové hypotézy vedou vysoké pozorované hodnoty $G_{max}$. Kritické hodnoty $c_\alpha(k, n_1 - 1)$ jsou uvedeny v tabulce T9.

# Leveneův test

Předpokládejme, že máme $k > 2$ **nezávislých** výběrů ze spojitého rozdělení.

Nulová a alternativní hypotéza:

$\mathrm{H}_0$: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$,   $\mathrm{H}_A$: $\neg H_0$ (alespoň jedna dvojice rozptylů se liší)

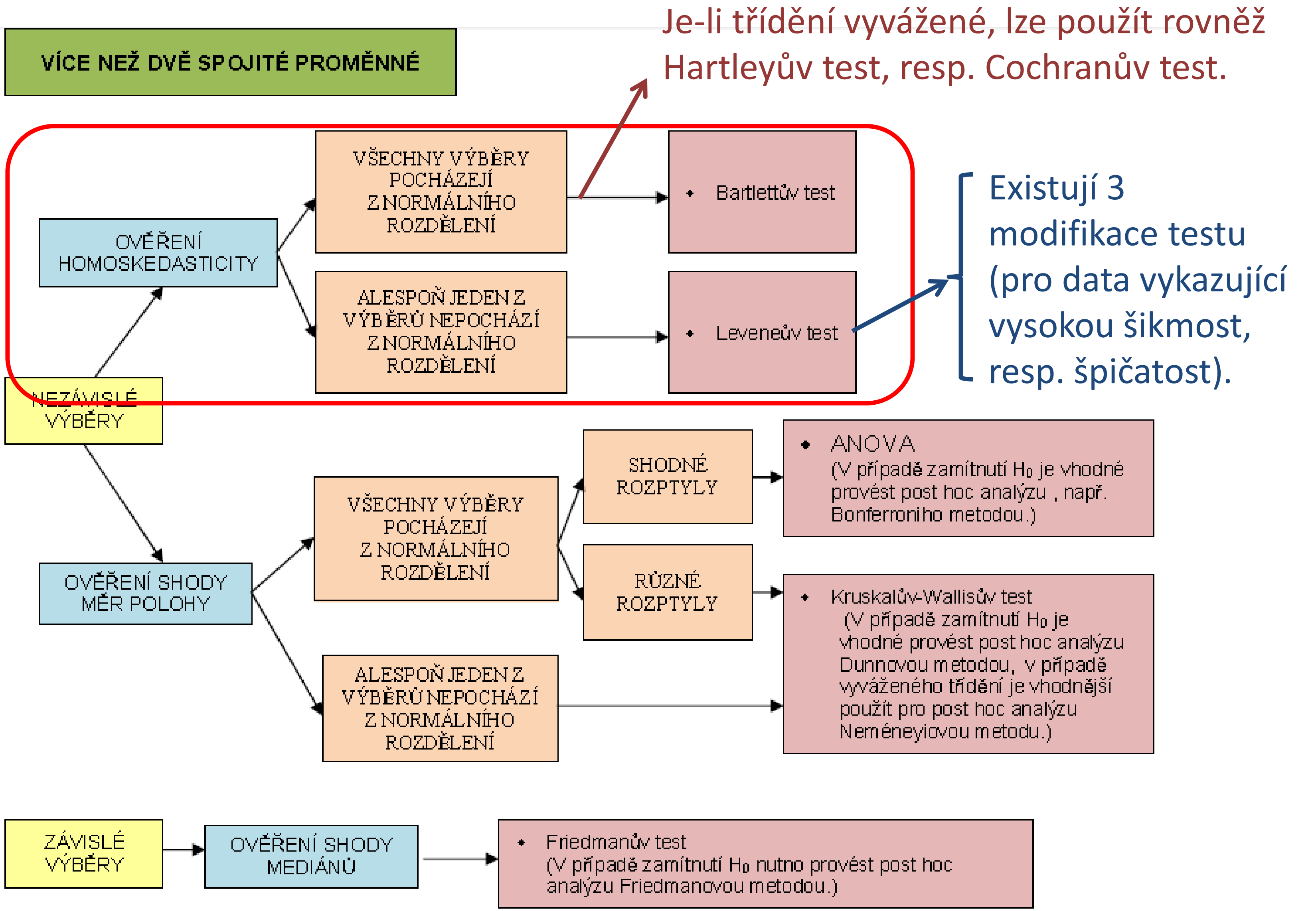
Testová statistika:

$$F_Z = \frac{\frac{SS_{ZB}}{k-1}}{\frac{SS_{Ze}}{n-k}},$$

kde $Z_{ij} = |X_{ij} - \bar{X}_i|$, $\bar{Z}_i = \frac{\sum_{j=1}^{n_i} Z_{ij}}{n_i}$, $\bar{\bar{Z}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{Z_{ij}}{n}$, $SS_{ZB} = \sum_{i=1}^{k} n_i \left(\bar{Z}_i - \bar{\bar{Z}}\right)^2$, $SS_{Ze} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Z_{ij} - \bar{Z}_i\right)^2$.

p-hodnota: $p\text{–}hodnota = 1 - F_0(x_{OBS})$, kde $F_0(x)$ je distribuční funkce Fisher-Snedecorova rozdělení s $k - 1$ stupni volnosti v čitateli a $n - k$ stupni volnosti ve jmenovateli.

# Testy shody rozptylů



Při sledování kvality pěnového polystyrénu (EPS) byla sledována hustota EPS [kg/m$^3$] čtyř různých výrobců A, B, C, D. Hustota byla stanovena pro 7 produktů každého z výrobců. Výsledky byly vepsány do níže uvedené tabulky. Ověřte homoskedasticitu objemové hmotnosti EPS jednotlivých výrobců.

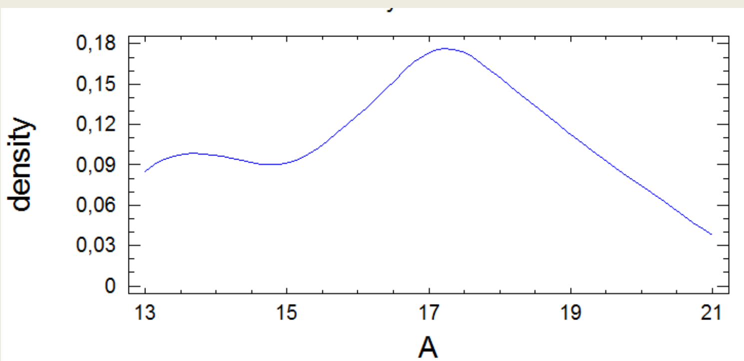
| Výrobce | Objemová hmotnost EPS [kg/m$^3$] | | | | | | |
|---|---|---|---|---|---|---|---|
| A | 14,3 | 13,0 | 17,6 | 16,9 | 16,1 | 20,0 | 18,4 |
| B | 19,1 | 22,5 | 21,2 | 21,0 | 20,3 | 17,4 | 22,7 |
| C | 19,7 | 16,8 | 15,8 | 20,1 | 18,2 | 18,6 | 18,9 |
| D | 13,2 | 12,6 | 12,9 | 13,7 | 17,3 | 11,2 | 15,0 |

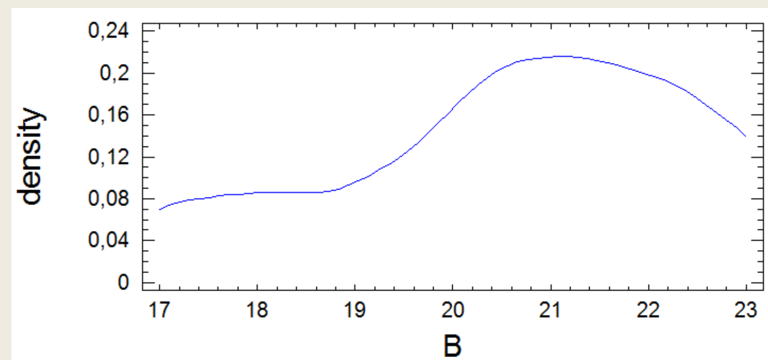
H$_0$: $\sigma_A^2 = \sigma_B^2 = \sigma_C^2 = \sigma_D^2$,
H$_A$: $\neg H_0$ (alespoň jedna dvojice rozptylů se liší)

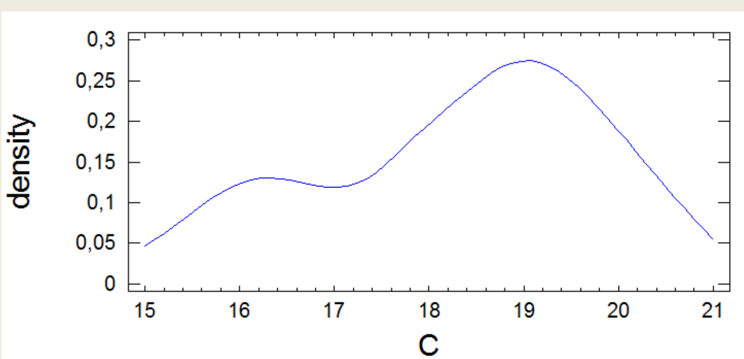
Předpoklady: Nezávislé výběry **OK**, normální rozdělení **?** každého z výběrů, vyváženost tříd **OK**

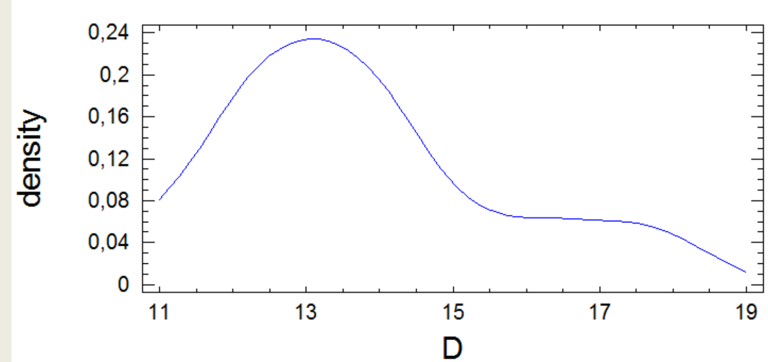
Při sledování kvality pěnového polystyrénu (EPS) byla sledována hustota EPS [kg/m$^3$] čtyř různých výrobců A, B, C, D.  Hustota byla stanovena pro 7 produktů každého z výrobců.  Výsledky byly vepsány do níže uvedené tabulky. Ověřte homoskedasticitu objemové hmotnosti EPS jednotlivých výrobců.

| Výrobce | Objemová hmotnost EPS [kg/m$^3$] | | | | | | |
|---|---|---|---|---|---|---|---|
| A | 14,3 | 13,0 | 17,6 | 16,9 | 16,1 | 20,0 | 18,4 |
| B | 19,1 | 22,5 | 21,2 | 21,0 | 20,3 | 17,4 | 22,7 |
| C | 19,7 | 16,8 | 15,8 | 20,1 | 18,2 | 18,6 | 18,9 |
| D | 13,2 | 12,6 | 12,9 | 13,7 | 17,3 | 11,2 | 15,0 |

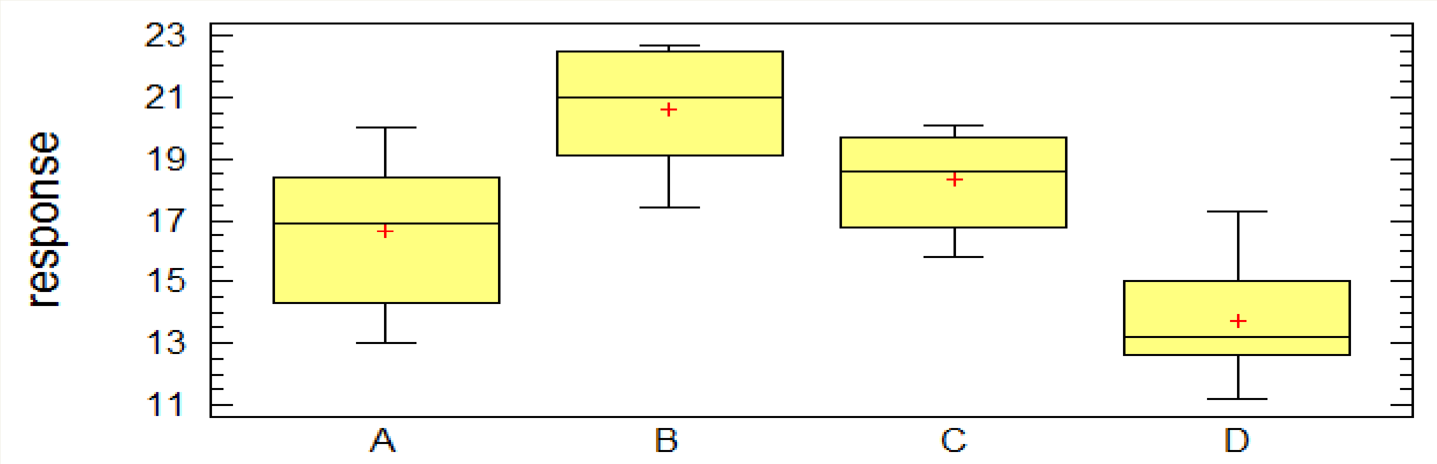


Při sledování kvality pěnového polystyrénu (EPS) byla sledována hustota EPS [kg/m$^3$] čtyř různých výrobců A, B, C, D. Hustota byla stanovena pro 7 produktů každého z výrobců. Výsledky byly vepsány do níže uvedené tabulky. Ověřte homoskedasticitu objemové hmotnosti EPS jednotlivých výrobců.



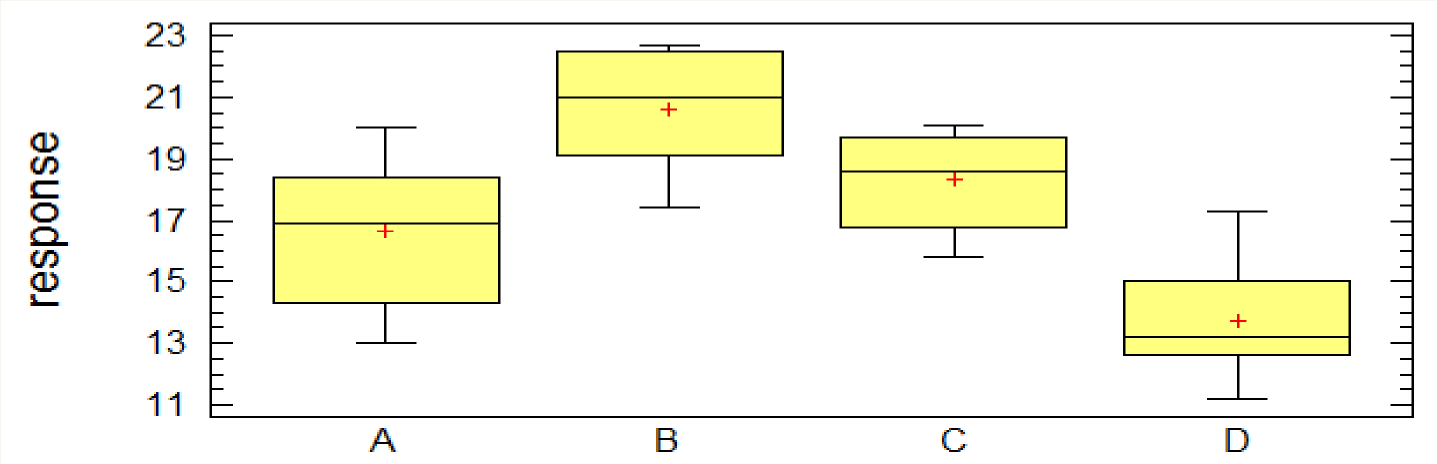

H$_0$: $\sigma_A^2 = \sigma_B^2 = \sigma_C^2 = \sigma_D^2$,
H$_A$: $\neg H_0$ (alespoň jedna dvojice rozptylů se liší)

Předpoklady: Nezávislé výběry, normální rozdělení každého z výběrů, vyváženost tříd

OK ??? OK

Cochran's C test: 0,371092 P-Value = 0,64871
Bartlett's test: 1,0505 P-Value = 0,775697
~~Hartley's test: 2,42857~~
Levene's test: 0,338333 P-Value = 0,797764

Při sledování kvality pěnového polystyrénu (EPS) byla sledována hustota EPS [kg/m$^3$] čtyř různých výrobců A, B, C, D. Hustota byla stanovena pro 7 produktů každého z výrobců. Výsledky byly vepsány do níže uvedené tabulky. Ověřte homoskedasticitu objemové hmotnosti EPS jednotlivých výrobců.




H$_0$: $\sigma_A^2 = \sigma_B^2 = \sigma_C^2 = \sigma_D^2$,
H$_A$: $\neg H_0$ (alespoň jedna dvojice rozptylů se liší)

Vzhledem k tomu, že nelze předpokládat normalitu jednotlivých výběrů, byl pro ověření homoskedasticity použit Leveneův test. Na hladině významnosti 0,05 nezamítáme nulovou hypotézu (p-hodnota=0,798), tj. předpoklad homoskedasticity nelze zamítnout.

# ANOVA

# Co je to ANOVA?

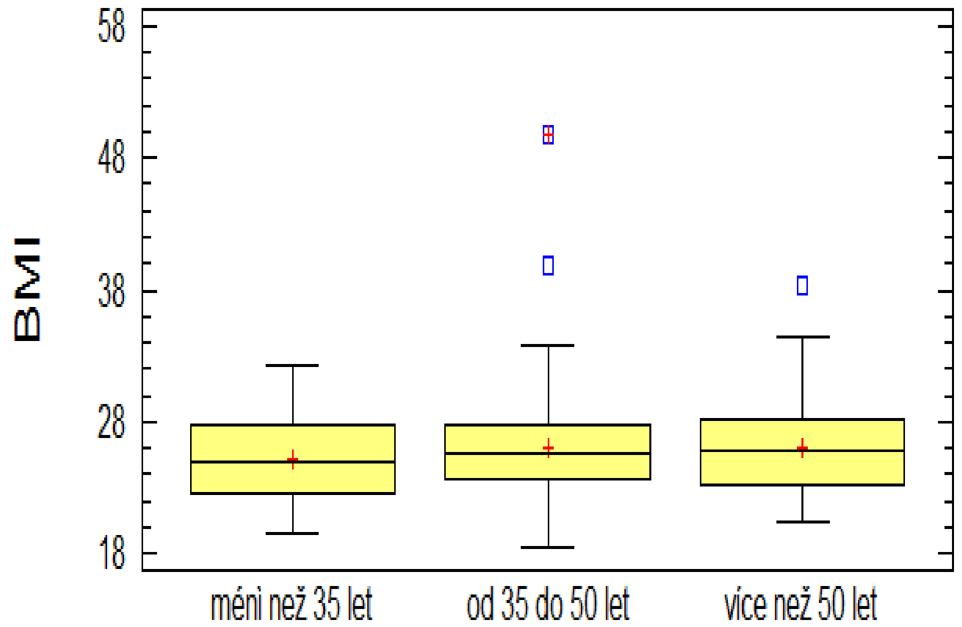
- test umožňující srovnání průměrů více než dvou výběrových souborů (akronym z angl. **AN**alysis **O**f **VA**riance, autor: R. A. Fisher, 1925)

Můžeme například zkoumat, zda

- typ absolvované střední školy ovlivňuje počet bodů dosažených studenty u přijímací zkoušky z matematiky,
- použitá medikace ovlivňuje krevní tlak pacientů,
- typ použitého hnojiva ovlivňuje výnosy určité plodiny,
- pracovní výkon dělníka závisí na umístění stroje, apod.

# Motivační příklad

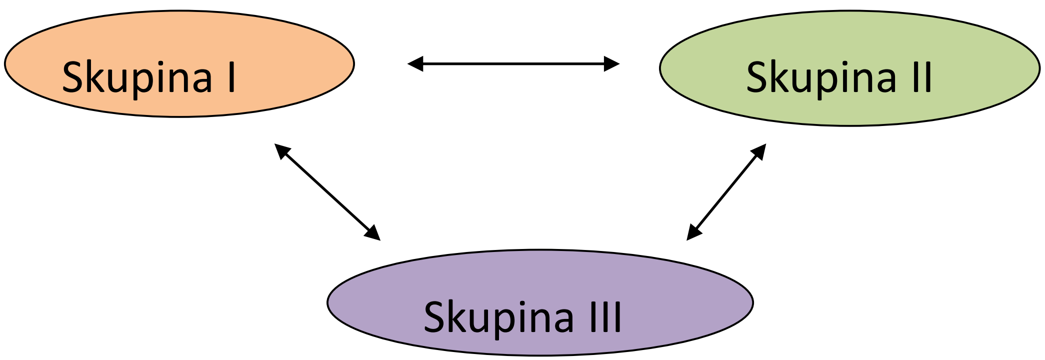
| BMI | | |
|---|---|---|
| méně než 35 let | 35 let - 50 let | více než 50 let |
| 27,8 | 23,9 | 21,3 |
| 26,8 | 23,7 | 27,4 |
| 28,4 | 25,1 | 28,1 |
| 25,8 | 25,3 | 29,8 |
| 30,8 | 25,8 | 29,0 |
| … | … | … |
| | | |
| | | |




Liší se hodnoty BMI v jednotlivých věkových kategoriích?

# Jak lze ověřit, zda jsou odchylky průměrů u $k$ $(k > 2)$ výběrů statisticky významné?

1. Můžeme zvolit vhodný test shody středních hodnot (např. t-test, Aspinové-Welchův test, …) a testovat jak se liší průměry skupin I a II, I a III, II a III, …, tj. provést celkem $\binom{k}{2}$ testů.





2. Lze použít speciální test pro srovnání středních hodnot více než dvou tříd.

# Problém násobného testování hypotéz

- Se zvyšujícím se počtem testů roste pravděpodobnost získání falešně pozitivního výsledku (chyba I. druhu), tedy pravděpodobnost toho, že se při našem testování zmýlíme a ukážeme na statisticky významný rozdíl tam, kde ve skutečnosti žádný neexistuje.

Určete výslednou pravděpodobnost chyby I. druhu, které byste se dopustili při ověřování shody středních hodnot tří skupin, použili-li byste přístup (1), tj. opakované testy shody stř.  hodnot, každý z nich na hladině významnosti 0,05.

**Řešení:**

- Potřebujeme ověřit shodu středních hodnot skupin I a II, I a III, II a III, tj. celkem 3 testy.
- Každý z testů má pravděpodobnost, že neuděláme chybu I. druhu 0,95, tj. celková pravděpodobnost, že neuděláme chybu I. druhu je $0{,}95 \cdot 0{,}95 \cdot 0{,}95 == 0{,}857$.
- Celková pravděpodobnost, že uděláme chybu I. druhu je tedy $1 - 0{,}857 =$

  $= \mathbf{0,143}$.

# Problém násobného testování hypotéz

- Se zvyšujícím se počtem testů roste pravděpodobnost získání falešně pozitivního výsledku (chyba I. druhu), tedy pravděpodobnost toho, že se při našem testování zmýlíme a ukážeme na statisticky významný rozdíl tam, kde ve skutečnosti žádný neexistuje.

| Počet tříd | Hladina významnosti používaná v t-testech | | |
|---|---|---|---|
| | 0,01 | 0,05 | 0,1 |
| 2 | 0,01 | 0,05 | 0,10 |
| 3 | 0,03 | 0,14 | 0,27 |
| 5 | 0,10 | 0,40 | 0,65 |
| 10 | 0,36 | 0,90 | 0,99 |
| 20 | 0,85 | 1,00 | 1,00 |

*Celková hladina významnosti při mnohonásobném testování*

$\Longrightarrow$ Mnohem vhodnější je použít speciální test shody středních hodnot více než dvou skupin, tzv. test **ANOVA**.

# ANOVA

Mějme *k* **nezávislých** realizací náhodného výběru z **normálních rozdělení** s **identickými rozptyly**:

| Číslo skupiny | 1 | 2 | … | $k$ |
|---|---|---|---|---|
| Náhodný výběr | $X_{11}$ <br> $\vdots$ <br> $X_{1n_1}$ | $X_{21}$ <br> $\vdots$ <br> $X_{2n_2}$ | $\vdots$ | $X_{k1}$ <br> $\vdots$ <br> $X_{kn_k}$ |

Předpoklady testu ANOVA:

- nezávislost výběrů,
- normalita rozdělení (POZOR - nutno ověřit pro každý výběr zvlášť!!!),
- homoskedasticita (tj. identické rozptyly, pro ověření lze použít např. Bartlettův nebo Leveneův test)

**Poznámka:** ANOVA byla původně navržena pro vyvážené třídění $(n_1 = n_2 =$

# ANOVA

Mějme *k* **nezávislých** realizací náhodného výběru z **normálních rozdělení** s **identickými rozptyly**:

| Číslo skupiny | 1 | 2 | … | $k$ |
|---|---|---|---|---|
| Náhodný výběr | $X_{11}$ | $X_{21}$ | | $X_{k1}$ |
| | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| | $X_{1n_1}$ | $X_{2n_2}$ | | $X_{kn_k}$ |

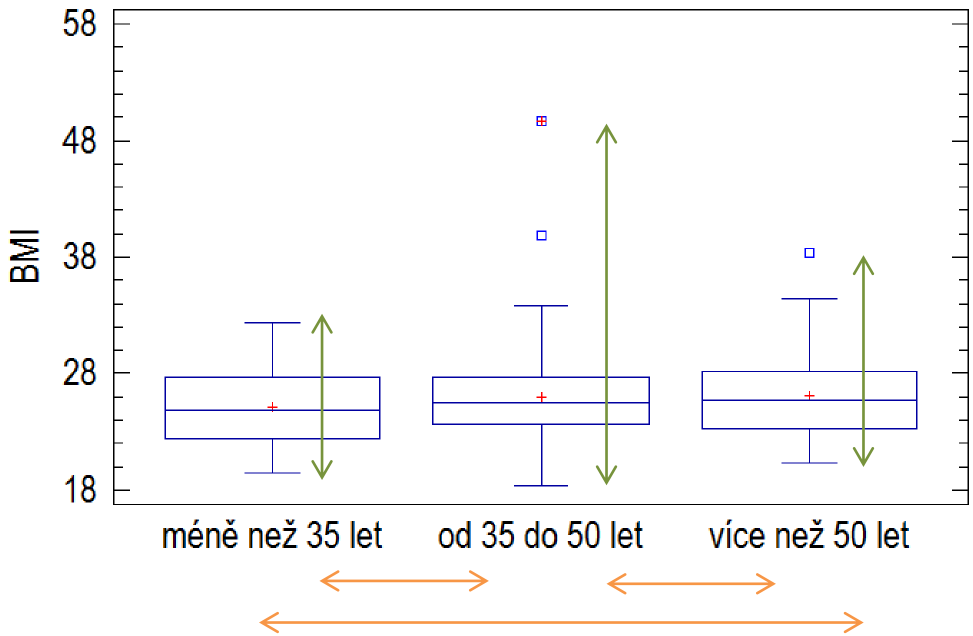
Předpoklady testu ANOVA:

- nezávislost výběrů,
- normalita rozdělení (POZOR - nutno ověřit pro každý výběr zvlášť!!!),
- homoskedasticita (tj. identické rozptyly, pro ověření lze použít např. Bartlettův nebo Leveneův test)

Nulová a alternativní hypotéza:

$H_0$: $\mu_1 = \mu_2 = \ldots = \mu_k$, $H_A$: $\neg H_0$ (alespoň jedna dvojice středních hodnot se liší)

# Princip výpočtu aneb
# co je příčinou rozdílných pozorovaných hodnot?




1. Vliv sledovaného faktoru (věk) - způsobuje rozdíly mezi třídami

2. Reziduální vlivy (životospráva, …) – způsobují rozdíly uvnitř tříd

Liší-li se průměry jednotlivých skupin vlivem různých středních hodnot příslušných populací, pak musí být rozptyl mezi třídami dostatečně velký vzhledem k rozptylu uvnitř tříd.

# Jak kvantifikovat rozdíly mezi třídami a rozdíly uvnitř tříd?

- meziskupinový součet čtverců (angl. sum of squares between groups),

$$SS_B = \sum_{i=1}^{k} n_i \left(\bar{X}_i - \bar{\bar{X}}\right)^2,$$

resp. rozptyl mezi skupinami

$$MS_B = \frac{SS_B}{k-1},$$

kde $k - 1$ je odpovídající počet stupňů volnosti $df_B$.

Kvantifikace rozdílů mezi třídami

- reziduální součet čtverců $SS_e$ (angl. sum of squares – errors)

$$SS_e = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_i\right)^2 = \sum_{i=1}^{k} (n_i - 1) s_i^2,$$

resp. reziduální rozptyl

$$MS_e = \frac{SS_e}{n-k},$$

kde $n - k$ je odpovídající počet stupňů volnosti $df_e$.

Kvantifikace rozdílů uvnitř tříd

# Rozklad celkové variability

Variabilitu jednotlivých pozorování kolem celkového průměru charakterizuje celkový součet čtverců (angl. total sum of squares),

$$SS_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{\bar{X}}\right)^2,$$

resp. celkový rozptyl (angl. „mean of squares“)

$$MS_T = \frac{SS_T}{n-1},$$

kde $n - 1$ je odpovídající počet stupňů volnosti $df_T$ (z angl. degreese of freedom).

Lze dokázat, že

$$SS_T = SS_B + SS_e.$$

V appletu ANOVA (java) sledujte vliv poměru rozptylu mezi třídami a rozptylu uvnitř tříd na rozhodnutí v testu ANOVA.

**Závěr**:

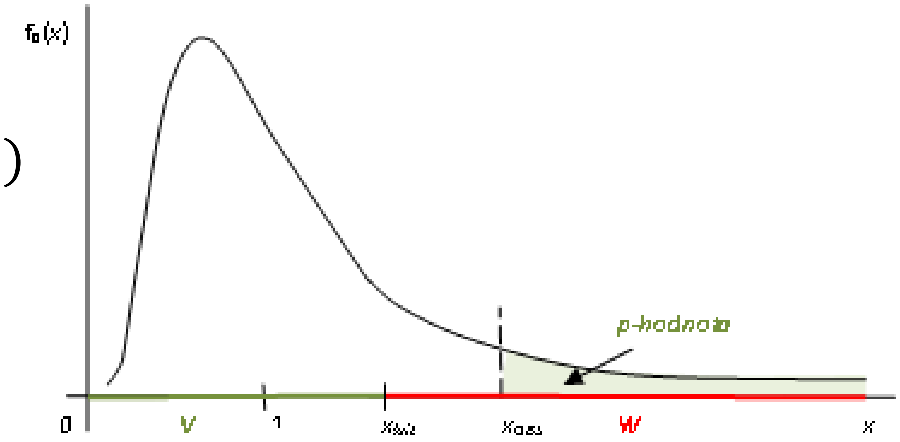
Čím je rozptyl mezi třídami větší než rozptyl uvnitř tříd,
tím silnější je naše přesvědčení o platnosti nulové hypotézy

Testovací kritérium pro test ANOVA:

$$F - poměr = \frac{MS_B}{MS_e}$$

- Platí-li $H_0$, $F - poměr \sim F_{k-1;n-k}$

*p*-hodnota: $p - hodnota = 1 - F_0(x_{OBS})$



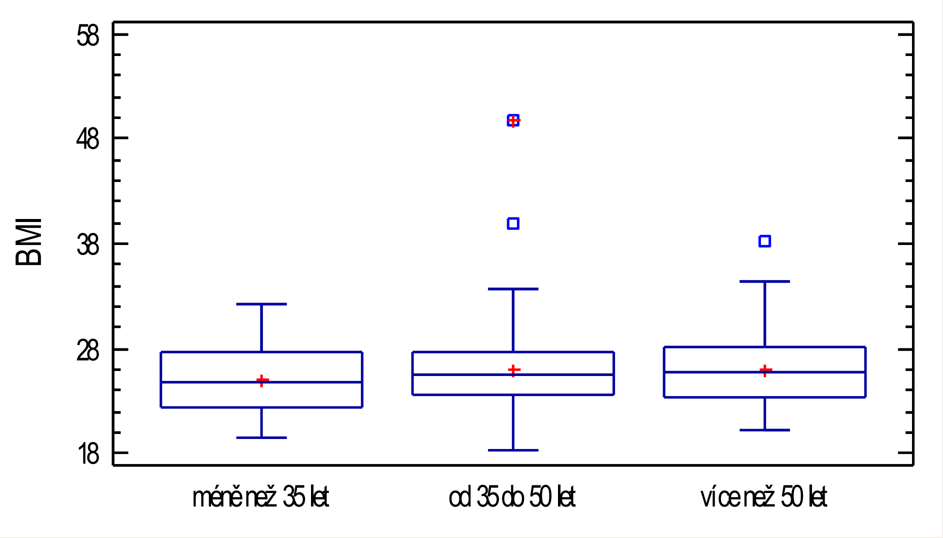
# Tabulka ANOVA

- dílčí i celkové výsledky testu ANOVA se zapisují do tabulky

| Zdroj variability | Součet čtverců | Počet stupňů volnosti | Rozptyl (prům. součet čtverců) | $F\text{–}poměr$ | $p\text{–}hodnota$ |
|---|---|---|---|---|---|
| Skupinový (faktor) | $SS_B = \sum_{i=1}^{k} n_i \left(\bar{X}_i - \bar{\bar{X}}\right)^2$ | $df_B = k - 1$ | $MS_B = \frac{SS_B}{df_B}$ | $\frac{MS_B}{MS_e}$ | $1 - F_0(x_{OBS})$ |
| Reziduální | $SS_e = \sum_{i=1}^{k} (n_i - 1) s_i^2$ | $df_e = n - k$ | $MS_e = \frac{SS_e}{df_e}$ | --- | --- |
| Celkový | $SS_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{\bar{X}}\right)^2$ | $df_T = n - 1$ | --- | --- | --- |

Pomocí výsledků explorační analýzy a testu ANOVA ověřte, zda má věk statisticky významný vliv na hodnoty BMI.

| | Count | Average | Variance |
|---|---|---|---|
| méně než 35 let | 53 | 25,0796 | 10,3825 |
| od 35 do 50 let | 123 | 25,9492 | 16,2775 |
| více než 50 let | 76 | 26,0982 | 12,3393 |
| Total | 252 | 25,8113 | 13,8971 |

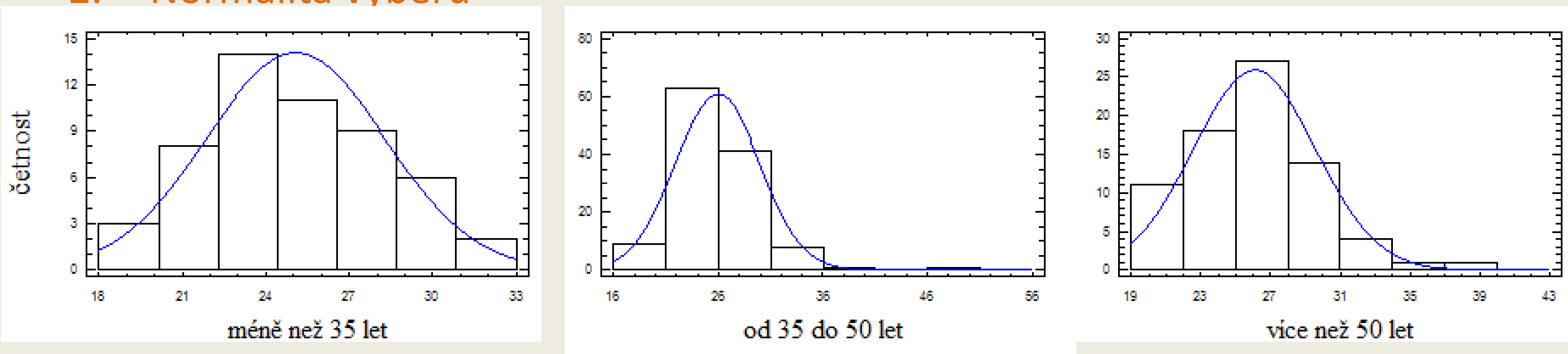


Pomocí výsledků explorační analýzy a testu ANOVA ověřte, zda má věk statisticky významný vliv na hodnoty BMI.

<u>Ověření předpokladů testu</u>:

1. Nezávislost výběrů – účastníky studie bylo 252 náhodně vybraných pacientů.

2. Normalita výběrů




Na základě exploratorní analýzy lze předpokládat, že předpoklad normality je splněn. (Později budeme pro ověření normality používat exaktní testy.)

3. Homoskedasticita

$H_0$: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$, $H_A$: $\neg H_0$, $p - hodnota = 0{,}129$ (Bartlettův test)

Na hladině významnosti 0,05 nelze předpoklad homoskedasticity zamítnout.

Pomocí výsledků explorační analýzy a testu ANOVA ověřte, zda má věk statisticky významný vliv na hodnoty BMI.

Nulová a alternativní hypotéza:

$H_0$: $\mu_1 = \mu_2 = \mu_3$, $H_A$: $\overline{H_0}$ (alespoň jedna dvojice středních hodnot se liší)

Výpočet *p*-hodnoty - doplnění tabulky ANOVA:

| | Count | Average | Variance |
|---|---|---|---|
| méně než 35 let | 53 | 25,0796 | 10,3825 |
| od 35 do 50 let | 123 | 25,9492 | 16,2775 |
| více než 50 let | 76 | 26,0982 | 12,3393 |
| Total | 252 | 25,8113 | 13,8971 |

$$SS_B = \sum_{i=1}^{k} n_i\left(\bar{X}_i - \bar{\bar{X}}\right)^2 = 53 \cdot (25{,}1 - 25{,}8)^2 + 123 \cdot (25{,}9 - 25{,}8)^2 + \\ +76 \cdot (26{,}1 - 25{,}8)^2 = 34{,}0$$

$$SS_e = \sum_{i=1}^{k} (n_i - 1)s_i^2 = 52 \cdot 10{,}4 + 122 \cdot 16{,}3 + 75 \cdot 12{,}3 = 3451{,}9$$

Pomocí výsledků explorační analýzy a testu ANOVA ověřte, zda má věk statisticky významný vliv na hodnoty BMI.

Nulová a alternativní hypotéza:

$H_0$: $\mu_1 = \mu_2 = \mu_3$, $H_A$: $\neg H_0$ (alespoň jedna dvojice středních hodnot se liší)

Výpočet *p*-hodnoty - doplnění tabulky ANOVA:

| Zdroj variability | Součet čtverců | Počet stupňů volnosti | Rozptyl (prům. součet čtverců) | $F$–poměr | $p$–hodnota |
|---|---|---|---|---|---|
| Skupinový (faktor) | $SS_B = \sum_{i=1}^{k} n_i\left(\bar{X}_i - \bar{\bar{X}}\right)^2$ | $df_B = k - 1$ | $MS_B = \frac{SS_B}{df_B}$ | $\frac{MS_B}{MS_e}$ | $1 - F_0(x_{OBS})$ |
| Reziduální | $SS_e = \sum_{i=1}^{k} (n_i - 1)s_i^2$ | $df_e = n - k$ | $MS_e = \frac{SS_e}{df_e}$ | --- | --- |
| Celkový | $SS_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{\bar{X}}\right)^2$ | $df_T = n - 1$ | --- | --- | --- |

Pomocí výsledků explorační analýzy a testu ANOVA ověřte, zda má věk statisticky významný vliv na hodnoty BMI.

Nulová a alternativní hypotéza:

$H_0$: $\mu_1 = \mu_2 = \mu_3$, $H_A$: $\overline{H_0}$ (alespoň jedna dvojice středních hodnot se liší)

Výpočet *p*-hodnoty - doplnění tabulky ANOVA:

| Zdroj variability | Součet čtverců | Počet stupňů volnosti | Rozptyl (prům. součet čtverců) | $F\text{–}poměr$ | $p\text{–}hodnota$ |
|---|---|---|---|---|---|
| Skupinový (faktor) | 34,0 | $df_B = k - 1$ | $MS_B = \frac{SS_B}{df_B}$ | $\frac{MS_B}{MS_e}$ | $1 - F_0(x_{OBS})$ |
| Reziduální | 3451,9 | $df_e = n - k$ | $MS_e = \frac{SS_e}{df_e}$ | --- | --- |
| Celkový | $SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{\bar{X}})^2$ | $df_T = n - 1$ | --- | --- | --- |

Pomocí výsledků explorační analýzy a testu ANOVA ověřte, zda má věk statisticky významný vliv na hodnoty BMI.

Nulová a alternativní hypotéza:

$H_0$: $\mu_1 = \mu_2 = \mu_3$, $H_A$: $\overline{H_0}$ (alespoň jedna dvojice středních hodnot se liší)

Výpočet *p*-hodnoty - doplnění tabulky ANOVA:

| Zdroj variability | Součet čtverců | Počet stupňů volnosti | Rozptyl (prům. součet čtverců) | $F$−poměr | $p$−hodnota |
|---|---|---|---|---|---|
| Skupinový (faktor) | 34,0 | $df_B = k - 1$ | $MS_B = \frac{SS_B}{df_B}$ | $\frac{MS_B}{MS_e}$ | $1 - F_0(x_{OBS})$ |
| Reziduální | 3451,9 | $df_e = n - k$ | $MS_e = \frac{SS_e}{df_e}$ | --- | --- |
| Celkový | 3485,9 | $df_T = n - 1$ | --- | --- | --- |

Pomocí výsledků explorační analýzy a testu ANOVA ověřte, zda má věk statisticky významný vliv na hodnoty BMI.

Nulová a alternativní hypotéza:

$H_0$: $\mu_1 = \mu_2 = \mu_3$, $H_A$: $\neg H_0$ (alespoň jedna dvojice středních hodnot se liší)

Výpočet *p*-hodnoty - doplnění tabulky ANOVA:

| Zdroj variability | Součet čtverců | Počet stupňů volnosti | Rozptyl (prům. součet čtverců) | $F\text{–}poměr$ | $p\text{–}hodnota$ |
|---|---|---|---|---|---|
| Skupinový (faktor) | 34,0 | $df_B = k - 1$ | $MS_B = \frac{SS_B}{df_B}$ | $\frac{MS_B}{MS_e}$ | $1 - F_0(x_{OBS})$ |
| Reziduální | 3451,9 | $df_e = n - k$ | $MS_e = \frac{SS_e}{df_e}$ | --- | --- |
| Celkový | 3485,9 | $df_T = n - 1$ | --- | --- | --- |

*k* … počet tříd, tj. počet porovnávaných výběrů ($k = 3$)
*n* … celkový rozsah všech výběrů ($n = 252$)

Pomocí výsledků explorační analýzy a testu ANOVA ověřte, zda má věk statisticky významný vliv na hodnoty BMI.

Nulová a alternativní hypotéza:

H$_0$: $\mu_1 = \mu_2 = \mu_3$, H$_A$: $\neg H_0$ (alespoň jedna dvojice středních hodnot se liší)

Doplnění tabulky ANOVA:

| Zdroj variability | Součet čtverců | | Počet stupňů volnosti | | Rozptyl (prům. součet čtverců) | $F$–poměr | $p$–hodnota |
|---|---|---|---|---|---|---|---|
| Skupinový (faktor) | 34,0 | / | 2 | = | $MS_B = \frac{SS_B}{df_B}$ | $\frac{MS_B}{MS_e}$ | $1 - F_0(x_{OBS})$ |
| Reziduální | 3451,9 | / | 249 | = | $MS_e = \frac{SS_e}{df_e}$ | --- | --- |
| Celkový | 3485,9 | | 251 | | --- | --- | --- |

Pomocí výsledků explorační analýzy a testu ANOVA ověřte, zda má věk statisticky významný vliv na hodnoty BMI.

Nulová a alternativní hypotéza:

$H_0$: $\mu_1 = \mu_2 = \mu_3$, $H_A$: $\neg H_0$ (alespoň jedna dvojice středních hodnot se liší)

Doplnění tabulky ANOVA:

| Zdroj variability | Součet čtverců | Počet stupňů volnosti | Rozptyl (prům. součet čtverců) | $F$–poměr | $p$–hodnota |
|---|---|---|---|---|---|
| Skupinový (faktor) | 34,0 | 2 | 17,0 | $\frac{MS_B}{MS_e}$ | $1 - F_0(x_{OBS})$ |
| Reziduální | 3451,9 | 249 | 13,9 | --- | --- |
| Celkový | 3485,9 | 251 | --- | --- | --- |

Pomocí výsledků explorační analýzy a testu ANOVA ověřte, zda má věk statisticky významný vliv na hodnoty BMI.

Nulová a alternativní hypotéza:

$H_0$: $\mu_1 = \mu_2 = \mu_3$, $H_A$: $\neg H_0$ (alespoň jedna dvojice středních hodnot se liší)

Doplnění tabulky ANOVA:

| Zdroj variability | Součet čtverců | Počet stupňů volnosti | Rozptyl (prům. součet čtverců) | $F$–poměr | $p$–hodnota |
|---|---|---|---|---|---|
| Skupinový (faktor) | 34,0 | 2 | 17,0 | 1,23 | 0,294 |
| Reziduální | 3451,9 | 249 | 13,9 | --- | --- |
| Celkový | 3485,9 | 251 | --- | --- | --- |

$$p - hodnota = 1 - F(1{,}23) = 0{,}294 \text{ ,}$$

kde *F*(x) je distr. f-ce Fisher-Snedecorova rozdělení s 2 st. volnosti v čitateli a 249 st. volnosti ve jmenovateli

Pomocí appletu Vybraná rozdělení pravděpodobnosti (excel).

Pomocí výsledků explorační analýzy a testu ANOVA ověřte, zda má věk statisticky významný vliv na hodnoty BMI.

Nulová a alternativní hypotéza:

$H_0$: $\mu_1 = \mu_2 = \mu_3$, $H_A$: $\neg H_0$ (alespoň jedna dvojice středních hodnot se liší)

Doplnění tabulky ANOVA:

| Zdroj variability | Součet čtverců | Počet stupňů volnosti | Rozptyl (prům. součet čtverců) | $F$–poměr | $p$–hodnota |
|---|---|---|---|---|---|
| Skupinový (faktor) | 34,0 | 2 | 17,0 | 1,23 | 0,294 |
| Reziduální | 3451,9 | 249 | 13,9 | --- | --- |
| Celkový | 3485,9 | 251 | --- | --- | --- |

Vyhodnocení testu:

Na hladině významnosti 0,05 nezamítáme nulovou hypotézu, tj. pozorované rozdíly mezi průměry BMI v jednotlivých věkových třídách nejsou statisticky významné.

# Post hoc analýza
# aneb mnohonásobné porovnávání

Post hoc analýza - dojde–li u vícevýběrových testů k zamítnutí nulové hypotézy, zajímá nás většinou co je příčinou tohoto rozhodnutí, tj. které dvojice skupin se statisticky významně liší.

Pro každou dvojici skupin *I* a *J* $(I \neq J)$ testujeme

$H_0$: $\mu_I = \mu_J$ vůči alternativě $H_A$: $\mu_I \neq \mu_J$.

# Metody pro post hoc analýzu

## Fisherovo LSD

Nulovou hypotézu zamítáme pokud

$$\left|\bar{x}_I - \bar{x}_J\right| \geq LSD_{IJ},$$

kde $LSD_{IJ}$ nazýváme nejmenší signifikantní diferencí (angl. **L**east **S**ignificant **D**ifference) a určíme ji jako

$$LSD_{IJ} = t_{n-k}\left(1 - \frac{\alpha}{2}\right)\sqrt{MS_e}\sqrt{\frac{1}{n_I} + \frac{1}{n_J}},$$

kde $t_{1-\frac{\alpha}{2}}(n-k)$ je $\left(1 - \frac{\alpha}{2}\right)$ kvantil Studentova rozdělení s $n-k$ st. volnosti.

**Nevýhoda**: Celková pravděpodobnost chyby I. druhu je vyšší (obvykle podstatně vyšší) než hladina významnosti $\alpha$ zvolená pro jednotlivá dílčí porovnávání dvojic. (viz slide 18)

# Metody pro post hoc analýzu

## Bonferroniho metoda (LSD s Bonferroniho korekcí)

Nulovou hypotézu zamítáme, pokud

$$\left|\bar{x}_I - \bar{x}_J\right| \geq t_{n-k}\left(1 - \frac{\alpha^*}{2}\right)\sqrt{MS_e}\sqrt{\frac{1}{n_I} + \frac{1}{n_J}},$$

kde $\alpha^*$ je upravená hladina významnosti, $\alpha^* = \frac{\alpha}{\binom{k}{2}}$,

$t_{1-\frac{\alpha^*}{2}}(n-k)$ je $\left(1 - \frac{\alpha^*}{2}\right)$ kvantil Studentova rozdělení s $n-k$ stupni volnosti.

Snížení hladiny významnosti v dílčích testech zajišťuje celkovou pravděpodobnost chyby I. druhu $\alpha$.

# Metody pro post hoc analýzu

## Schéffého metoda

Nulovou hypotézu zamítáme, pokud

$$\left|\bar{x}_I - \bar{x}_J\right| \geq \sqrt{MS_e}\sqrt{F_{1-\alpha}(k-1, n-k)(k-1)\left(\frac{1}{n_I} + \frac{1}{n_J}\right)},$$

kde $F_{1-\alpha}(k-1, n-k)$ je $(1-\alpha)$ kvantil Fisher-Snedecorova rozdělení s $k-1$ stupni volnosti v čitateli a $n-k$ stupni volnosti ve jmenovateli.

**Poznámka**: V současnosti jedna z nejpreferovanějších metod post hoc analýzy pro ANOVu.

# Metody pro post hoc analýzu

Tukeyho metoda (pouze pro vyvážené třídění, citlivější než Schéffého metoda)

Nulovou hypotézu zamítáme, pokud

$$\left|\bar{x}_I - \bar{x}_J\right| \geq q_\alpha(k, n-k)\sqrt{MS_e}\sqrt{\frac{1}{n_I}},$$

kde $q_\alpha(k, n-k)$ je $\alpha$ kvantil studentizovaného rozpětí, který je tabelován (tabulka T10).

V případě **nevyváženého třídění** lze použít modifikovaný Tukeyho test známý pod názvem **Tukey HSD**.

Nulovou hypotézu pak zamítáme, pokud

$$\left|\bar{x}_I - \bar{x}_J\right| \geq q_\alpha(k, n-k)\sqrt{MS_e}\sqrt{\frac{1}{2}\left(\frac{1}{n_I} + \frac{1}{n_J}\right)},$$

kde $q_\alpha(k, n-k)$ je $\alpha$ kvantil studentizovaného rozpětí, který je tabelován v T10.

# Metody prezentace výsledků post hoc analýzy

- Znaménkové schéma

tj. tabulka $k \times k$, ve které každé porovnávané skupině odpovídá jeden řádek a jeden sloupec. V příslušném poli tabulky lze dohodnutým symbolem (tečka, křížek, hvězdička, zabarvení) označit ty dvojice skupin, pro něž byl identifikován statisticky významný rozdíl mezi průměry.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | \|tr-ts\| (kritická hodnota) | |
| r\s | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | |
| 1 | - | 7,55 (4,44) | 9,55 (4,44) | | | | | | |
| 2 | 7,55 (4,44) | - | 2 (4,925) | | | | | | |
| 3 | 9,55 (4,44) | 2 (4,925) | - | | | | | | |
| 4 | | | | - | | | | | |

| | $\left\lvert \bar{x}_I - \bar{x}_J \right\rvert$ | $LSD_{IJ}$ |
|---|---|---|
| Gymnázium – SPŠ* | 6,7 | 3,898 |
| Gymnázium – OU* | 8,3 | 3,539 |
| SPŠ - OU | 1,6 | 4,003 |

# Metody prezentace výsledků post hoc analýzy

- Homogenní skupiny

tj. skupiny, pro něž by v jednofaktorové analýze rozptylu nebyla zamítnuta hypotéza o shodě středních hodnot

Method: 95,0 percent Tukey HSD

| | Count | Mean | Homogeneous Groups |
|---|---|---|---|
| vietnamský | 9 | 58,3333 | X |
| Japonský | 12 | 60,9167 | X |
| indický | 8 | 68,875 | XX |
| gruzínský | 8 | 69,75 | XX |
| Čínský | 10 | 72,4 | X |

Skupiny jsou seřazeny vzestupně podle průměrů.

**Poznámka:** Některé homogenní skupiny se mohou překrývat. Znamená to, že některé skupiny mohou mít vlastnosti blízké více homogenním skupinám současně.

# Metody prezentace výsledků post hoc analýzy

- Homogenní skupiny

tj. skupiny, pro něž by v jednofaktorové analýze rozptylu nebyla zamítnuta hypotéza o shodě středních hodnot

Method: 95,0 percent Tukey HSD

| | Count | Mean | Homogeneous Groups |
|---|---|---|---|
| vietnamský | 9 | 58,3333 | X |
| Japonský | 12 | 60,9167 | X |
| indický | 8 | 68,875 | XX |
| gruzínský | 8 | 69,75 | XX |
| Čínský | 10 | 72,4 |  X |

Skupiny jsou seřazeny vzestupně podle průměrů.

**Poznámka:** Některé homogenní skupiny se mohou překrývat. Znamená to, že některé skupiny mohou mít vlastnosti blízké více homogenním skupinám současně. Stanovení hom. skupin nemusí být jednoznačné.

# Jak postupovat při porušení předpokladů?

Porušení nezávislosti výběrů:

**Friedmanův test** (viz Úvod do statistiky, test není vyučován v rámci tohoto předmětu)

Porušení homoskedasticity:

Pokusíme se **stabilizovat rozptyl** pomocí transformací proměnných (logaritmická, Box-Coxova transformace, …). Pokud se nám rozptyl stabilizovat nepodaří, je možné přihlédnout k tomu, že ANOVA není (v případě, že **data jsou vyvážena**) příliš citlivá na porušení předpokladu homoskedasticity. Nejsou-li data vyvážena, můžeme použít **Kruskal-Wallisův test** (tzv. neparametrická ANOVA, vícevýběrový test o shodě mediánů).

Porušení normality:

Pokusíme se **normalizovat data** pomocí transformací proměnných (logaritmická, Box-Coxova transformace, …). Pokud se nám data normalizovat nepodaří, lze použít **Kruskal – Wallisův test.**

# Kruskal-Wallisův test

# Kruskal-Wallisův test

Nechť je dáno $k$ **nezávislých výběrů** z rozdělení se **spojitou distribuční funkci** o rozsazích $n_1, n_2, \dots, n_k$. Označme $n = n_1 + n_2 \; \dots + \; n_k$.

| Číslo skupiny | 1 | 2 | … | $k$ |
|---|---|---|---|---|
| Náhodný výběr | $X_{11}$<br>$\vdots$<br>$X_{1n_1}$ | $X_{21}$<br>$\vdots$<br>$X_{2n_2}$ | $\vdots$ | $X_{k1}$<br>$\vdots$<br>$X_{kn_k}$ |

Nulová hypotéza a alternativa:

H$_0$: $x_{0,5_1} = x_{0,5_2} = \cdots = x_{0,5_k}$, H$_A$:$\neg H_0$(alespoň jedna dvojice mediánů se liší)

# Kruskal-Wallisův test

Výpočet testové statistiky Q:

Všech $n$ pozorovaných hodnot veličiny $X_{ij}$ se seřadí do rostoucí posloupnosti a určí se jejich **pořadí** $R_{ij}$. Tato pořadí uspořádáme do tabulky a určíme tzv. **součty pořadí pro jednotlivé výběry** $T_i$.

| Výběr | Pořadí veličin $X_{ij}$ v uspořádané rostoucí posloupnosti | | | | Součty pořadí |
|---|---|---|---|---|---|
| 1 | $R_{11}$ | $R_{12}$ | | $R_{1n_1}$ | $T_1$ |
| 2 | $R_{21}$ | $R_{22}$ | | $R_{2n_2}$ | $T_2$ |
| | | | | | |
| k | $R_{k1}$ | $R_{k2}$ | | $R_{kn_k}$ | $T_k$ |

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

*p*-hodnota:

Jsou-li rozsahy jednotlivých výběru alespoň 5 prvků, má testová statistika $Q$ v případě platnosti nulové hypotézy přibližně $\chi^2$ rozdělení s $k-1$ stupni volnosti a $p - hodnota = 1 - F_0(x_{OBS})$.

# Metody post hoc analýzy pro Kruskal-Wallisův test

## Dunnové metoda

Nechť průměrné pořadí *i*-té skupiny je $t_i = \frac{T_i}{n_i}$, $z_p$... *p* kvantil normovaného normálního rozdělení, modifikovaná hladina významnosti je $\alpha^* = \frac{\alpha}{\binom{k}{2}}$. Jestliže

$$|t_I - t_J| \geq \sqrt{\frac{1}{12}\left(\frac{1}{n_I} + \frac{1}{n_J}\right) n(n+1)} z_{1-\alpha^*},$$

pak se mediány *I*-tého a *J*-tého výběru statisticky významně liší.

# Metody post hoc analýzy pro Kruskal-Wallisův test

Neméneiova metoda (pouze pro vyvážené třídění)

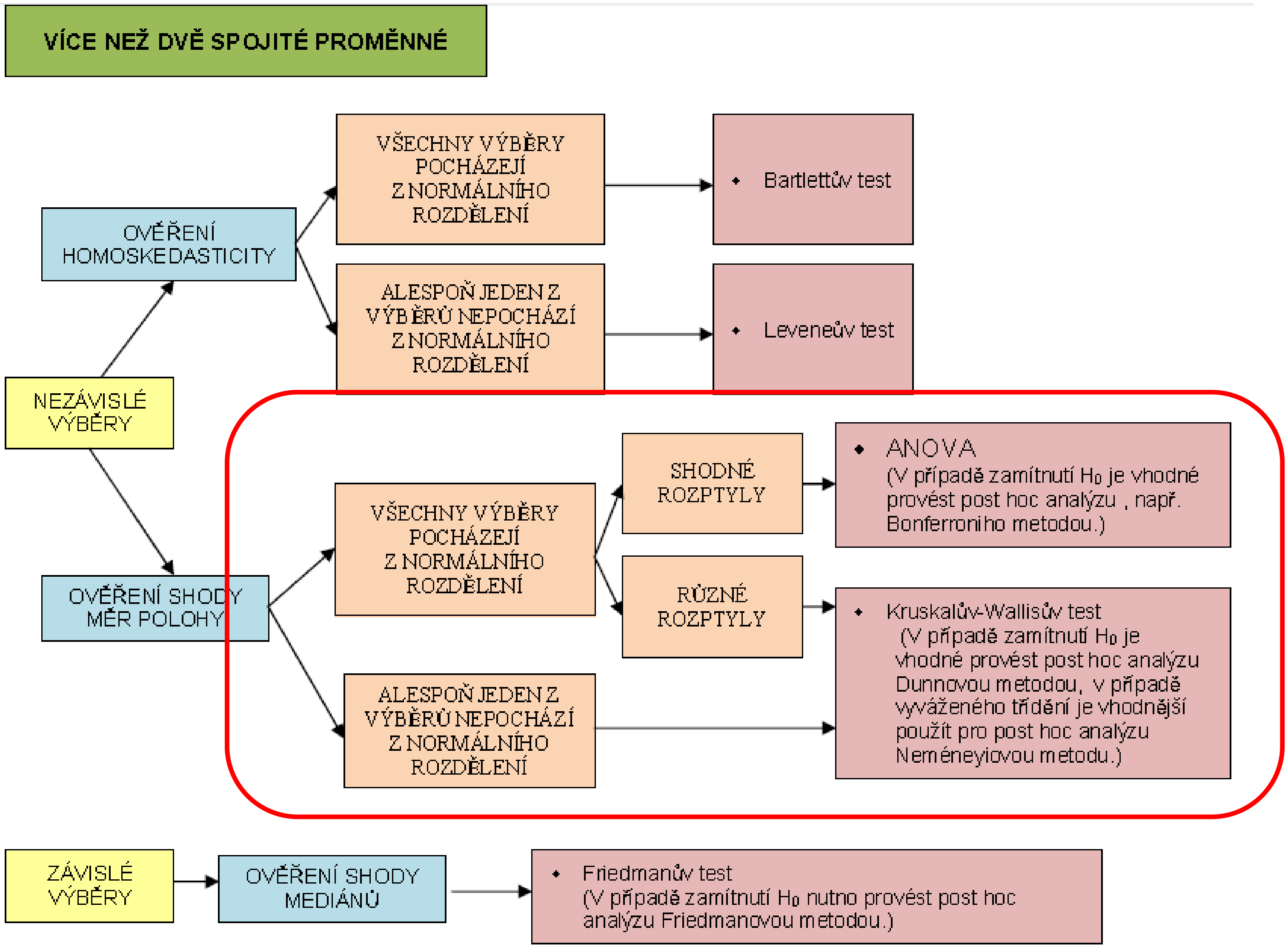
Pro menší počty skupin $k$ a rozsahy jednotlivých výběrů $m$ jsou kritické hodnoty pro $\left|T_I - T_J\right|$ uvedeny v tabulce T11. Je-li počet skupin $k > 10$ nebo rozsahy jednotlivých výběrů $m > 16$, užije se následující postup.

Nechť $q_\alpha(k, \infty)$ je kritická hodnota rozpětí $k$ nezávislých náhodných veličin s rozdělením $N(0; 1)$. Lze ji najít v posledním řádku tabulky T10.
Řekneme, že se mediány *I*-tého a *J*-tého výběru statisticky významně liší, když

$$\left|t_I - t_J\right| \geq q_\alpha(k, \infty)\sqrt{\frac{1}{12}k(km + 1)}.$$

**Poznámka**: *Statgraphics 5.0 neobsahuje metody post hoc analýzy pro Kruskal-Wallisův test. Pro tento test lze použít výpočetní applet Kruskalův-Wallisův test (excel).*

# Ověřování shody měr polohy pro více než dva výběry



Děkuji za pozornost!