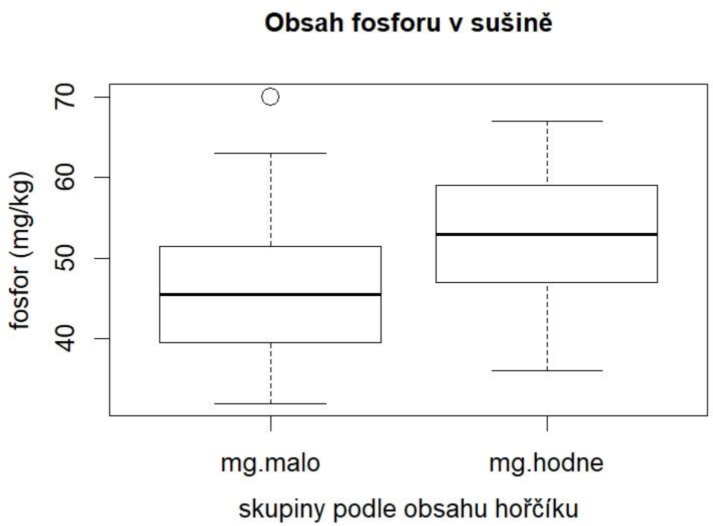
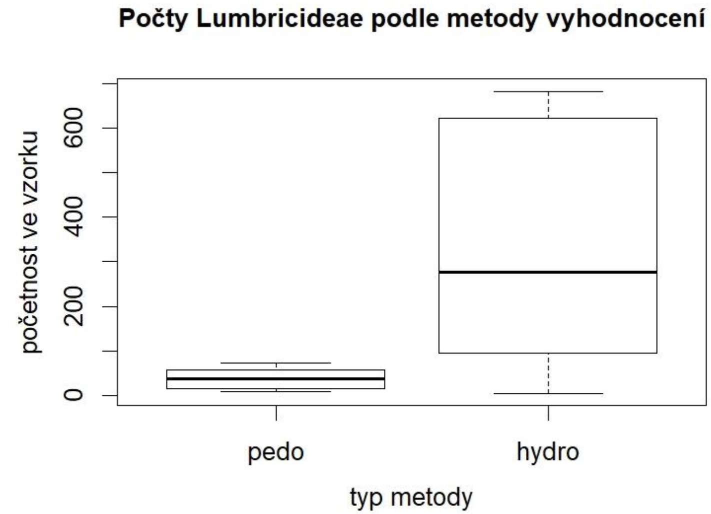
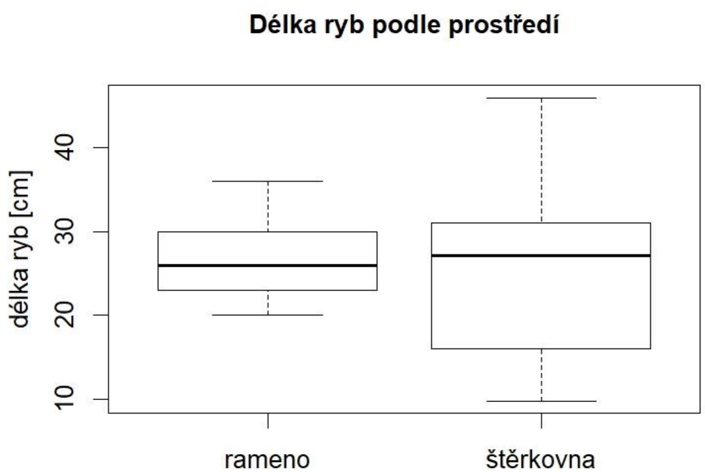
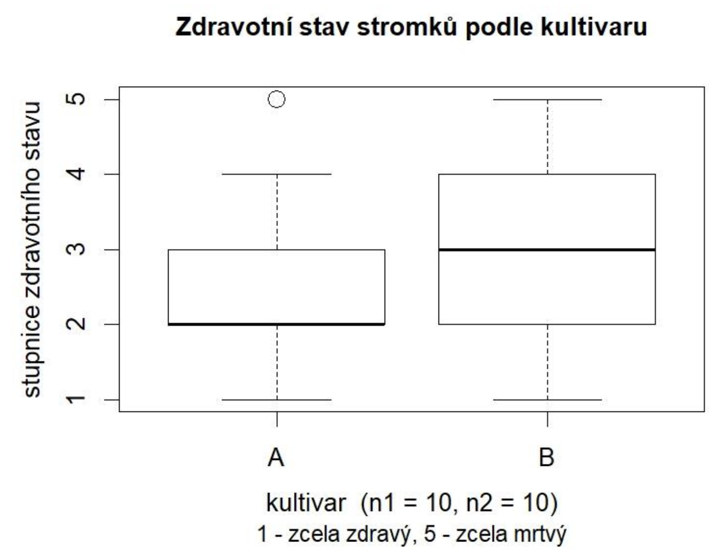
Dva výběry


Obsah fosforu v sušině
fosfor (mg/kg)
40
50
60
70
mg.malo
mg.hodne
skupiny podle obsahu hořčíku


Počty Lumbricideae podle metody vyhodnocení
početnost ve vzorku
0
200
400
600
pedo
hydro
typ metody


Délka ryb podle prostředí
délka ryb [cm]
10
20
30
40
rameno
štěrkovna


Zdravotní stav stromků podle kultivaru
stupnice zdravotního stavu
1
2
3
4
5
A
B
kultivar (n1 = 10, n2 = 10)
1 - zcela zdravý, 5 - zcela mrtvý

Dva výběry: F-test shody variancí [homogeneity of variances]

Předpoklady testu:

- (X_1, X_2, …, X_k) a (Y_1, Y_2, …, Y_m) všechno nezávislé; $n_X = k, n_Y = m$
- $\boldsymbol{X_i \sim N(\mu_X, \sigma_X^2), Y_i \sim N(\mu_Y, \sigma_Y^2)}$, parametry neznáme.

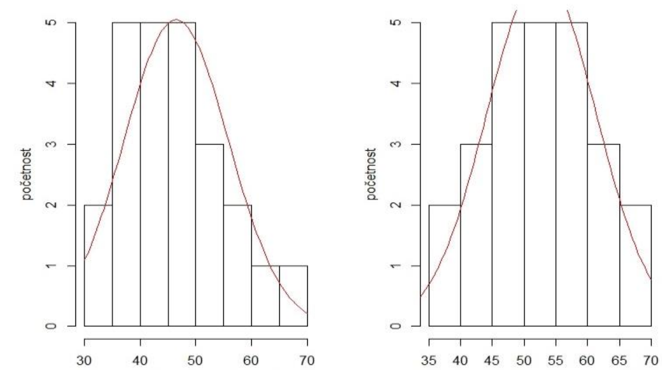
Hypotéza: $\boldsymbol{H_0: \sigma_X^2 = \sigma_Y^2}$

ale testujeme $\boldsymbol{\frac{\sigma_X^2}{\sigma_Y^2}} = 1$

alternativa $\boldsymbol{H_1: \sigma_X^2 \neq \sigma_Y^2}$

Příklad fosfor:

Málo Mg hodně Mg



SD(fosforu): 9.4mg/kg 8.6 mg/kg

Testová statistika:

$$\boldsymbol{F = \frac{S_X^2}{S_Y^2}}$$ má za platnosti $\boldsymbol{H_0}$ rozdělení $\boldsymbol{F_{(n_X-1, n_Y-1)}}$

Doplňkově podrobnější odvození testové statistiky F:

Teoreticky má být $\boldsymbol{F} = \dfrac{\dfrac{\sum_{i=1}^{k} V_i^{\,2}}{k}}{\dfrac{\sum_{j=1}^{m} W_j^{\,2}}{m}}$

$V_i \sim N(0,1)$ a tedy $\sum_{i=1}^{k} V_i^{\,2} \sim \chi^2{}_k$

$W_j \sim N(0,1)$ a tedy $\sum_{j=1}^{m} W_j^{\,2} \sim \chi^2{}_m$

Máme: $\dfrac{S_X{}^2}{\sigma_X{}^2} = \dfrac{1}{\sigma_X{}^2} \cdot \dfrac{\sum_{i=1}^{k}(X_i-\bar{X})^2}{k-1} = \dfrac{1}{k-1} \cdot \dfrac{\sum_{i=1}^{k}(X_i-\bar{X})^2}{\sigma_X{}^2} = \dfrac{1}{k-1} \cdot \sum_{i=1}^{k}\left(\dfrac{X_i-\bar{X}}{\sigma_X}\right)^2$

Jeden stupeň volnosti ztrácím odhadem $\mu_X = \bar{X}$

$\sim N(0,1)$

Podobně pro $S^2{}_Y$.

Dohromady:

$$\boldsymbol{F} = \frac{\dfrac{S_X^2}{\sigma_X^2}}{\dfrac{S_Y^2}{\sigma_Y^2}} = \frac{S_X^2}{\sigma_X^2} \cdot \frac{\sigma_Y^2}{S_Y^2} = \frac{S_X^2}{S_Y^2} \cdot \frac{\sigma_Y^2}{\sigma_X^2} =_{H_0} \frac{S_X^2}{S_Y^2} \sim_{H_0} F_{(n_X-1,\,n_Y-1)}$$

Za platnosti hypotézy je zlomek roven 1

Dva výběry: F-test shody variancí

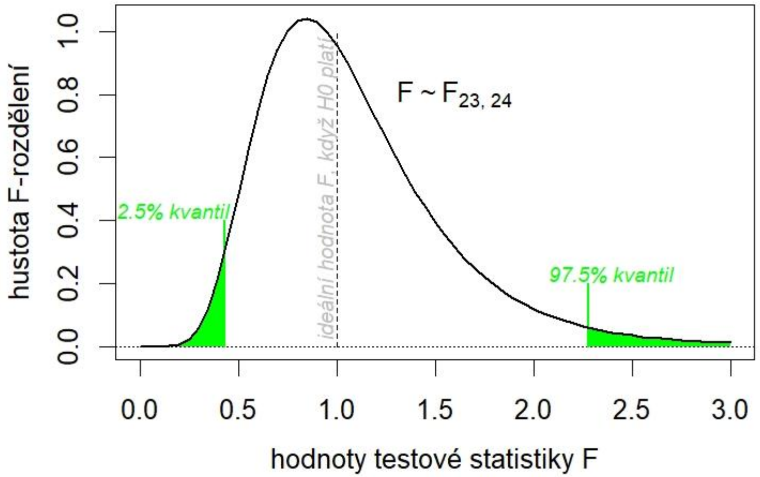
Testová statistika: $F = \frac{S_X^2}{S_Y^2} \sim_{H_0} F_{(n_X-1, n_Y-1)}$

Kritérium: typicky uvažujeme oboustrannou alternativu $H_1: \sigma^2{}_X \neq \sigma^2{}_Y$, zajímá nás jen otázka shody či neshody variancí.

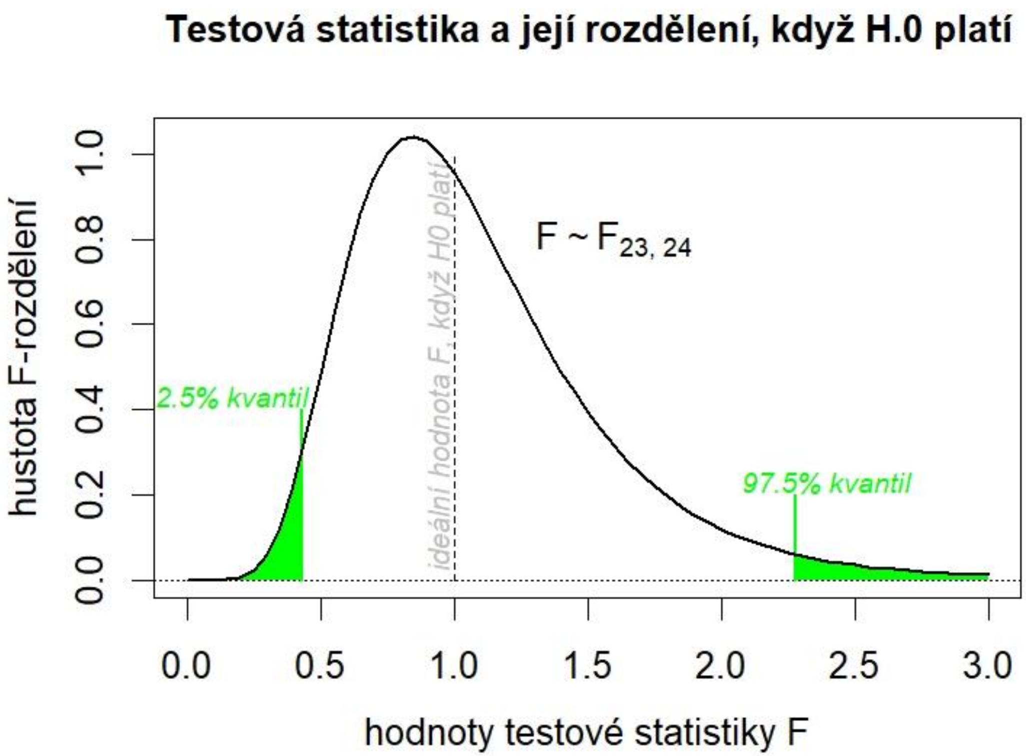
Proti nulové hypotéze svědčí dvě situace:

Buď $S_X^2 \gg S_Y^2$ a $F = \frac{S_X^2}{S_Y^2}$ leží na pravém chvostu, srovnám s $F_{f1,f2}\left(1-\frac{\alpha}{2}\right)$

nebo $S_X^2 \ll S_Y^2$ a $F = \frac{S_X^2}{S_Y^2}$ leží na levém chvostu, srovnám s $F_{f1,f2}\left(\frac{\alpha}{2}\right)$



Dva výběry: F-test shody variancí



Doplňkově:

F-rozdělení <u>není souměrné</u>, ale platí, že

$$P\left(F \leq F_{f1,f2}\left(\frac{\alpha}{2}\right)\right) = P\left(\frac{1}{F} \geq F_{f2,f1}\left(1-\frac{\alpha}{2}\right)\right) = \frac{\alpha}{2}$$

Proto <u>někdy</u> kritérium zní: $F = \frac{větší\ z\ odhadů\ rozptylu}{menší\ z\ odhadů\ rozptylu} \geq F_{čitat,jmenov}\left(1-\frac{\alpha}{2}\right)$

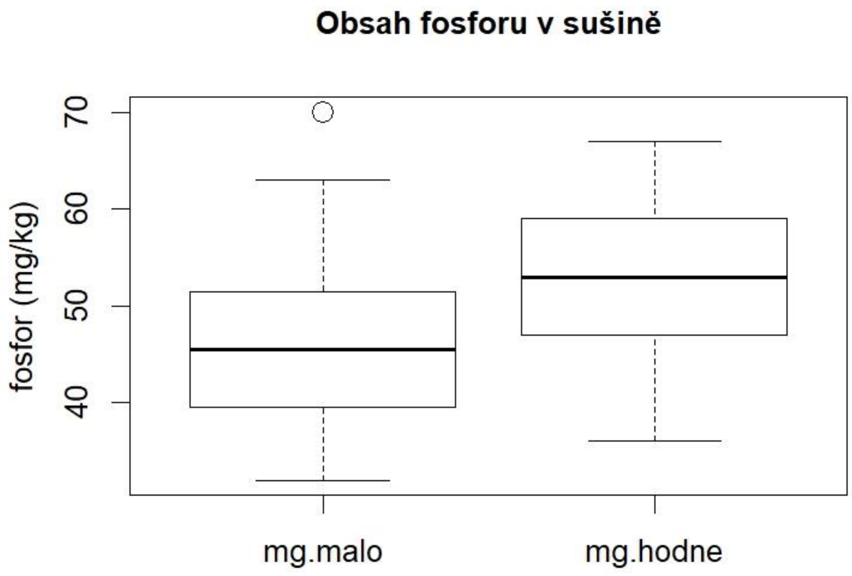
Dva výběry: F-test shody variancí

Příklad: obsah fosforu v listech pšenice

Rozdělených podle obsahu hořčíku.

H$_0$: $\sigma^2_{mg.malo} = \sigma^2_{mg.hodne}$

H$_1$: $\sigma^2_{mg.malo} \neq \sigma^2_{mg.hodne}$



R:
```
var.test(x, y, ratio=1,
  alternative=c("two.sided","less",
  "greater"), conf.level=0.95, ...)
```

- F test to compare two variances

  data: fosfor$mg.malo and fosfor$mg.hodne

  F = 1.1933, **num df** = 23, **denom df** = 24, p-value = 0.6696

  alternative hypothesis: true ratio of variances is not equal to 1

  95 percent confidence interval: 0.5229178 2.7433662

  sample estimates ratio of variances = 1.193335

Numerator **d**egrees of **f**reedom
Stupně volnosti v čitateli (nahoře)
Denominator **d**egr. of **f**reedom
St. volnosti ve jmenovateli (dole).

Odhady rozptylů: $S^2_{mg.málo} = 88.26$, $S^2_{mg.hodně} = 73.96$

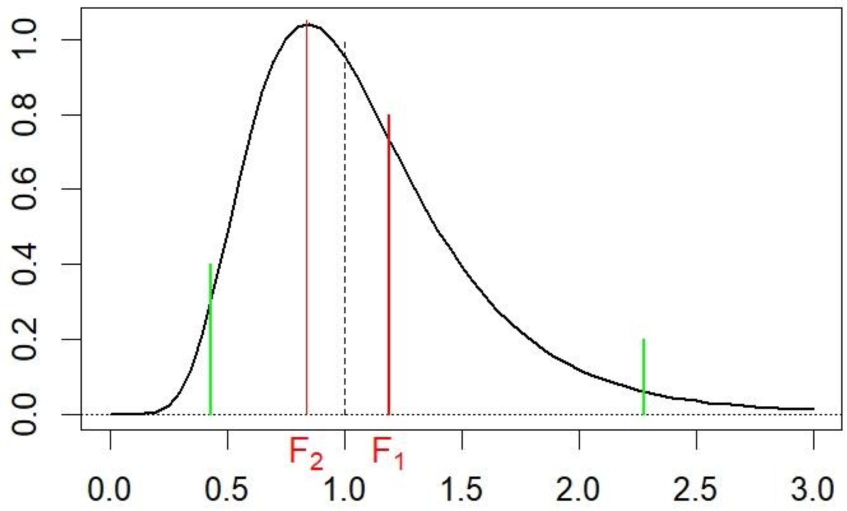
Příklad: F-test shody variancí

Jaký výsledek dostaneme, když zadáme proměnné v opačném pořadí?

```
> var.test(mg.malo,mg.hodne)
        F test to compare two variances
data:  mg.malo and mg.hodne
F = 1.1933, num df = 23, denom df = 24,
        p-value = 0.6696
95 percent confidence interval:
        (0.5229178 ;  2.7433662)
sample estimates: ratio of variances = 1.193335
```

$$F_1 = 1.19$$
$$F_2 = 0.84$$
$$\frac{1}{1.1933} = 0.8380$$
$$F_{23,\,24}(0.975) = 2.28$$
$$F_{24,\,23}(0.025) = 0.44$$

```
> var.test(mg.hodne,mg.malo)
        F test to compare two variances
data:  mg.hodne and mg.malo
F = 0.83799, num df = 24, denom df = 23,
        p-value = 0.6696
95 percent confidence interval:
        (0.3645157; 1.9123465)
sample estimates: ratio of variances = 0.838
```



Konfidenční interval pro poměr variancí $\frac{\sigma_X^2}{\sigma_Y^2}$

Je nesouměrný, protože F-rozdělení je nesouměrné:

$$F_{n_X-1,n_Y-1}\left(\frac{\alpha}{2}\right) \leq \frac{S^2{}_X}{S^2{}_Y}\cdot\frac{\sigma^2{}_Y}{\sigma^2{}_X} \leq F_{n_X-1,n_Y-1}\left(1-\frac{\alpha}{2}\right)$$

zapracujeme znalost $P\left(F \leq F_{f1,f2}\left(\frac{\alpha}{2}\right)\right) = P\left(\frac{1}{F} \geq F_{f2,f1}\left(1-\frac{\alpha}{2}\right)\right)$

$$F_{n_Y-1,n_X-1}\left(1-\frac{\alpha}{2}\right) \geq \frac{S^2{}_Y}{S^2{}_X}\cdot\frac{\sigma^2{}_X}{\sigma^2{}_Y} \geq F_{n_Y-1,n_X-1}\left(\frac{\alpha}{2}\right)$$

Uspořádáme logicky menší < větší a převedeme poměr odhadů rozptylů:

$$F_{n_Y-1,n_X-1}\left(\frac{\alpha}{2}\right)\cdot\frac{S^2{}_X}{S^2{}_Y} \leq \frac{\sigma^2{}_X}{\sigma^2{}_Y} \leq F_{n_Y-1,n_X-1}\left(1-\frac{\alpha}{2}\right)\cdot\frac{S^2{}_X}{S^2{}_Y}$$

F-test shody variancí – další poznámky

- Pokud nezamítám H_0 (rozptyly jsou shodné), počítám <u>odhad společného rozptylu</u> [pooled variance] takto:

$$S^2 = \frac{\sum_{i=1}^{n_X}(X_i - \bar{X})^2 + \sum_{j=1}^{n_Y}\left(Y_j - \bar{Y}\right)^2}{n_X + n_Y - 2} = \frac{(n_X - 1){S_X}^2 + (n_Y - 1){S_Y}^2}{n_X + n_Y - 2}$$

- <u>**F-test je dost slabý test**</u>, zvláště při malých četnostech výběrů. Uvažte, že pro výběry velikosti 10 (běžné počty v biologii) srovnáváme statistiku F s kvantilem $F_{9,\,9}(0.975) = 4.026$, tzn. že musí být $S_X^2 > 4 * S_Y^2$, čtyřikrát větší, aby test zamítnul H_0. Proto je pravděpod. β chyby 2. druhu velká.

Další testy shody variancí

Leveneův test – bude u analýzy rozptylu; pracuje s odchylkami od průměrů (nebo lépe mediánů). Předpokládá normální rozdělení dat.

Brown & Forsythe test – modifikace Leveneova testu na mediány, takto je test robustnější vůči odchylce od normálního rozdělení dat.

R: balík **`car`** (**C**ompanion to **A**pplied **R**egression), funkce **`leveneTest`**`(x, y, …)`

Zahrnuje již modifikaci na mediány, je to tedy Brown & Forsythův test.

Ansari-Bradleyův dvouvýběrový test - pořadový test, neparametrický. Testuje hodnotu ***s***, což je zde poměr směrodatných odchylek (scales).

R: **`ansari.test`**`(x, y, …)`

R obsahuje také **Moodyho test**, který je ovšem komentován takto:

„existují užitečnější testy pro tento problém“ ☺

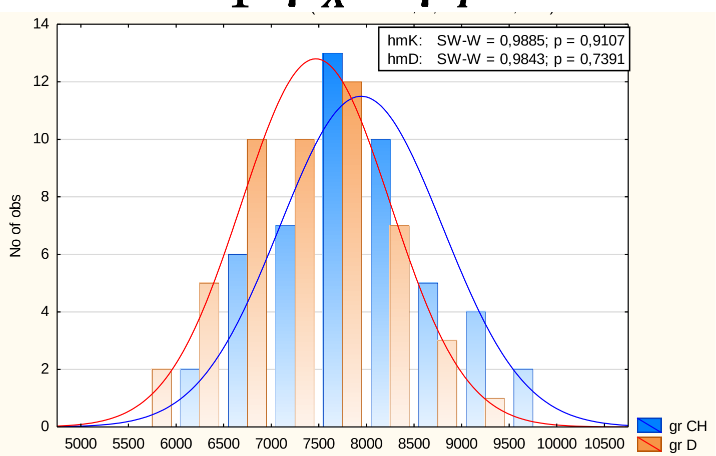
Dva výběry: t-test shody průměrů

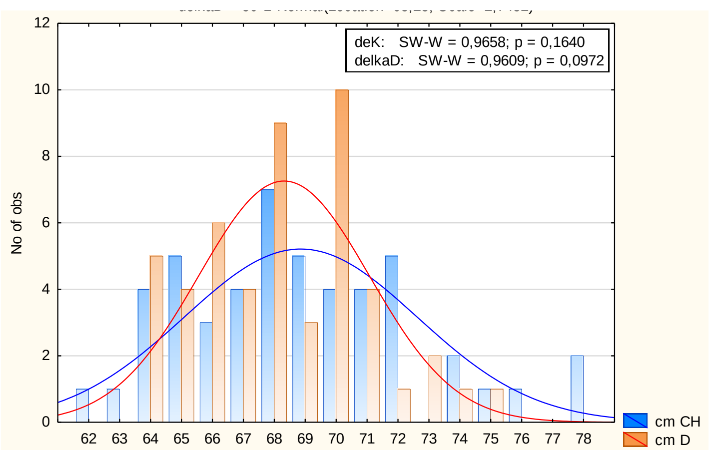
Předpoklady testu:

- $(X_1, X_2, \ldots, X_k)$ a $(Y_1, Y_2, \ldots, Y_m)$ všechno nezávislé
- $\boldsymbol{X_i \sim N(\mu_X, {\sigma_X}^2), Y_i \sim N(\mu_Y, {\sigma_Y}^2)}$, parametry neznáme.
- výběry mají stejnou varianci, liší se tedy jen posunutím střední hodnoty; $\boldsymbol{{\sigma_X}^2 = {\sigma_Y}^2}$ ozn. $\boldsymbol{\sigma^2}$
- pokud předpoklad stejných rozptylů není splněn, máme Welchův test (dále)

Hypotéza: $\boldsymbol{H_0: \mu_X = \mu_Y}$ testujeme $\mu_X - \mu_Y = 0$

alternativa $\boldsymbol{H_1: \mu_X \neq \mu_Y}$





Dva výběry: t-test shody průměrů

Hypotéza: $\boldsymbol{H_0}$: $\boldsymbol{\mu_X = \mu_Y}$ testujeme $\mu_X - \mu_Y = 0$
alternativa $\boldsymbol{H_1}$: $\boldsymbol{\mu_X \neq \mu_Y}$

Použijeme odhad pro společný rozptyl [pooled variance]

$$S^2 = \frac{\sum_{i=1}^{n_X}(X_i - \bar{X})^2 + \sum_{j=1}^{n_Y}\left(Y_j - \bar{Y}\right)^2}{n_X + n_Y - 2} = \frac{(n_X - 1){S_X}^2 + (n_Y - 1){S_Y}^2}{n_X + n_Y - 2}$$

Testová statistika:

$$\boldsymbol{T = \frac{\bar{X} - \bar{Y} - 0}{S.E.(\bar{X} - \bar{Y})} = \frac{\bar{X} - \bar{Y}}{S}\sqrt{\frac{n_X n_Y}{n_X + n_Y}}} \sim_{H_0} \boldsymbol{t_{n_X + n_Y - 2}}$$

čti: *má za platnosti hypotézy H_0 rozdělení …*

Kritéria podle H_1:

$$\mu_X \neq \mu_Y \;\rightarrow\; \boldsymbol{|T| \geq t_{n_X+n_Y-2}(1 - {}^{\alpha}/_{2})}$$

$$\mu_X > \mu_Y \;\rightarrow\; \boldsymbol{T \geq t_{n_X+n_Y-2}(1 - \alpha)}$$

$$\mu_X < \mu_Y \;\rightarrow\; \boldsymbol{T \leq t_{n_X+n_Y-2}(\alpha)} = -t_{n_X+n_Y-2}(1 - \alpha)$$

Pro zájemce odvození rovnosti

$S.E.(\bar{X}-\bar{Y}) = S \cdot \sqrt{\frac{n_X+n_Y}{n_X n_Y}}$, tedy $\boldsymbol{var}\,(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}}) = \left(\frac{\boldsymbol{n_X+n_Y}}{\boldsymbol{n_X n_Y}}\right)\boldsymbol{S^2}$

- $\boldsymbol{var}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}}) = E\{(\bar{X}-\bar{Y}) - E(\bar{X}-\bar{Y})\}^2 = E\{\bar{X}-E\bar{X}-\bar{Y}+E\bar{Y}\}^2 = E\{(\bar{X}-E\bar{X})-(\bar{Y}-E\bar{Y})\}^2 = E\{(\bar{X}-E\bar{X})^2 - 2\cdot(\bar{X}-E\bar{X})(\bar{Y}-E\bar{Y}) +$

- Všimněte si rozdílu proti párovému t-testu:

Párový t-test **H₀**: $\boldsymbol{\mu_X = \mu_Y}$ a počítám $\boldsymbol{X_i = U_i - V_i}$ … průměr rozdílů, tj. má smysl počítat rozdíl v páru pozorování

Dvouvýběrový **H₀**: $\boldsymbol{\mu_X = \mu_Y}$ (H_0 stejná) $\boldsymbol{\bar{X} - \bar{Y}}$ … rozdíl průměrů
Zde není žádný vztah mezi X_i a Y_i, navíc počet $\boldsymbol{n_X}$ a $\boldsymbol{n_Y}$ se může lišit.

- **t-test je celkem robustní vůči narušení předpokladů**, zvlášť pokud máme dostatek pozorování **a** výběry jsou zhruba stejně početné (uplatní se CLV).

PŘESTO je-li podezření na nestejnost variancí (normálního rozdělení) **nebo** se $\boldsymbol{n_X}$ a $\boldsymbol{n_Y}$ značně liší, použijeme lépe

Welchův přibližný t-test:

$$T = \frac{\bar{X} - \bar{Y} - 0}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}} \sim_{H_0} t_f \quad , kde \ f = \frac{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)^2}{\frac{\left(\frac{S_X^2}{n_X}\right)^2}{n_X - 1} + \frac{\left(\frac{S_Y^2}{n_Y}\right)^2}{n_Y - 1}}$$

Počet stupňů volnosti $\boldsymbol{f}$ může být i desetinné číslo! Rozdělení testové statistiky známe jen přibližně, proto také p-hodnota je jen přibližná.

T-test - zadání v Rku:

R: `t.test(x, y, alternative=c("two.sided","less", "greater"), mu=0, paired=FALSE, var.equal=FALSE, conf.level=0.95, ...)`

Přednastavené nestejné rozptyly, počítá rovnou Welchův přibližný test.

tady každá skupina ve vlastním sloupečku
př. fosfor: *mg.malo* a *mg.hodne*

Zadání pomocí formule:

`t.test(měření ~ skupiny, data, subset, …)`

tady jednom sloupečku kódování příslušnosti ke skupině (faktor *skupiny*)
a ve druhém sloupečku naměřené hodnoty (*měření*)
př. fosfor.kody: *P ~ skup*
datová tabulka musí být „*attach*“ nebo ji musím zadat jako *data = fosfor.kody*

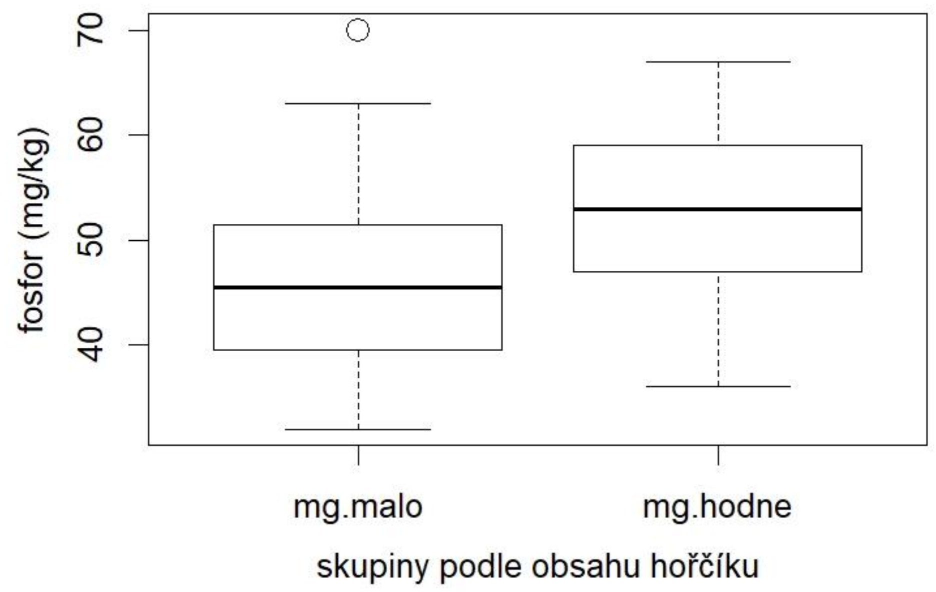
t-test – příklad fosfor:

Obsah fosforu v listech pšenice
Skupiny podle obsahu hořčíku.
H$_0$: $\mu_1 = \mu_2$ H$_1$: $\mu_1 \neq \mu_2$
H$_0$: průměrný populační obsah fosforu je v obou skupinách srovnatelný.
Předpoklady: $n_1 = 24$, $n_2 = 25$
Normalita: histogramy, Q-Q ploty

```
> shapiro.test(fosfor$mg.malo)
data:  fosfor$mg.malo
W = 0.96787, p-value = 0.6147
> shapiro.test(fosfor$mg.hodne)
data:  fosfor$mg.hodne
W = 0.97993, p-value = 0.8838
```

Normalita splněna,
počet pozorování rozumný.

Variance jsou srovnatelné.




```
> var.test(fosfor$mg.malo, fosfor$mg.hodne)

F test to compare two variances
F = 1.1933, num df = 23, denom df = 24,
        p-value = 0.6696
```

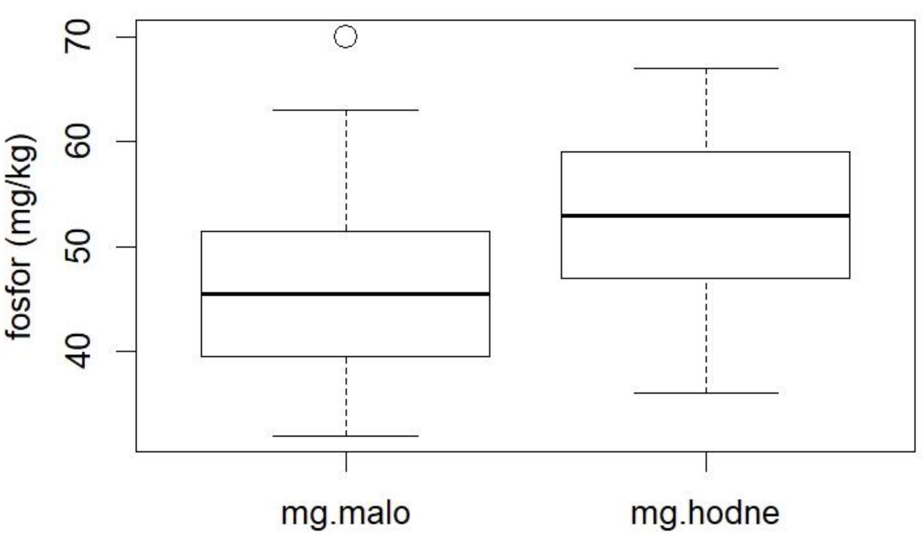
Nezamítám hypotézu o shodných rozptylech.

t-test – příklad fosfor:

Obsah fosforu v listech pšenice

Skupiny podle obsahu hořčíku.

H$_0$: $\mu_1 = \mu_2$ H$_1$: $\mu_1 \neq \mu_2$



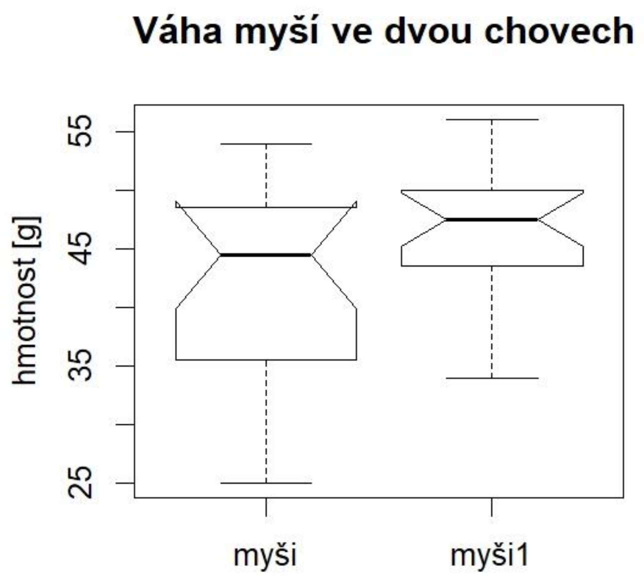

```
> t.test(fosfor$mg.malo, fosfor$mg.hodne, var.equal = T)

      Two Sample t-test

data:  fosfor$mg.malo and fosfor$mg.hodne
t = -2.4352, df = 47, p-value = 0.01873
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval: ( -11.434431;  -1.088902 )
sample estimates:    mean of x     mean of y
                      46.45833      52.72000
```

Zamítám hypotézu o shodných populačních průměrech na hladině 5 %.

t-test – příklad myši:



Jsou hmotnosti myší v základním a „prvním“ chovu srovnatelné?

H_0: $\mu_X = \mu_Y$ H_1: $\mu_X \neq \mu_Y$

Předpoklady:
Normalita dobrá, víme z DÚ.

```
> var.test(mysi, mysi1)
  F test to compare two variances
F = 2.2339, num df = 19, denom df = 19,
      p-value = 0.08788
alternative hypothesis: true ratio of variances is not equal to 1
sample estimates:ratio of variances = 2.233919
```

Rozdíl v rozptylech dvojnásobný => Welchův přibližný *t*-test.

```
> t.test(mysi, mysi1)
      Welch Two Sample t-test
data:  mysi and mysi1
t = -1.9661, df = 33.171, p-value = 0.0577
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval: -8.8504732  0.1504732
sample estimates:mean of x mean of y
    42.45     46.80
```

Těsně nezamítám hypotézu o srovnatelných populačních průměrech.

Neparametrický test rovnosti středních hodnot

Mannův-Whitneyův test alias dvouvýběrový Wilcoxonův test

Předpoklady:

$(X_1, X_2, \ldots, X_k)$ a $(Y_1, Y_2, \ldots, Y_m)$ všechno nezávislé ze spojitého rozdělení.

Nulová hypotéza: rozdělení pravděpodobností náh. veličin ***X*** a ***Y*** je stejné (tedy ani posunutí, ani velký rozdíl ve variabilitě, stejné histogramy)

→ Za platnosti nulové hypotézy jsou stejné i populační mediány

Princip Wilcoxonova pořadového testu:

Sesypeme oba výběry dohromady a hodnoty seřadíme (neřešíme +/- jako u párového testu). Když platí H_0, měly by se zhruba pravidelně střídat hodnoty z ***X*** a z ***Y***. Součet pořadí by tedy měl být srovnatelný, zhruba polovina z $\frac{(n_x+n_Y)\cdot(n_x+n_Y+1)}{2}$, tedy $\frac{(n_x+n_Y)\cdot(n_x+n_Y+1)}{4}$.

Protože ale velikosti výběrů $\boldsymbol{n_X}$ a $\boldsymbol{n_Y}$ nemusí být stejné, musím spočtený součet porovnávat s poměrnou částí celého součtu, viz dále:

Dvouvýběrový Wilcoxonův test

Princip Wilcoxonova pořadového testu:

- $W_X = \sum_{i=1}^{n_X} R(X_i)$ … součet pořadí hodnot z výběru X
- $W_Y = \sum_{j=1}^{n_Y} R(Y_j)$ … součet pořadí hodnot z výběru Y

~ multinomické rozdělení

- Součet všech pořadí: $W_X + W_Y = \frac{(n_x+n_Y)\cdot(n_x+n_Y+1)}{2}$
- Za platnosti H_0 je očekávaný součet

Poměrná část vůči celkovému počtu hodnot

$$EW_X = \frac{n_x}{n_x + n_Y} \frac{(n_x + n_Y)\cdot(n_x + n_Y + 1)}{2} = \frac{n_x \cdot (n_x + n_Y + 1)}{2}$$

$$var\, W_X = \frac{n_x \cdot n_Y \cdot (n_x + n_Y + 1)}{12}$$

- Při větším množství shod v pořadí ještě úprava výběrového rozptylu…
- Pro menší rozsahy výběrů n_X a n_Y lze počítat přesné pravděpodobnosti, pro větší n se používá aproximace normálním rozdělením.

Dvouvýběrový Mannův-Whitneyův test

Test je založen na zdánlivě jiné myšlence: porovnává všechny možné dvojice (X_i, Y_j) a počítá, v kolika případech je hodnota *X* menší než *Y*.

Jsou-li distribuce veličin srovnatelné (H_0), bude to zhruba polovina dvojic.

- Označme $\boldsymbol{U_X}$ počet dvojic, kde $(\boldsymbol{X_i < Y_j}), \boldsymbol{U_X \sim_{H_0} Bi(n_X \cdot n_Y, 0.5)}$
- Označme $\boldsymbol{U_Y}$ počet dvojic, kde $(\boldsymbol{X_i > Y_j})$
- Případy $(\boldsymbol{X_i = Y_j})$ započítáme polovinou k $\boldsymbol{U_X}$ a polovinou k $\boldsymbol{U_Y}$
- Kritický obor je zpravidla popisován pomocí $U = \min(U_X, U_Y)$.

- Tyto rovnice ukazují souvislost mezi testovými statistikami:

(Wilcoxon) $\boldsymbol{W_X = n_x n_Y + \frac{n_x \cdot (n_x + 1)}{2} - U_X}$ (Mann-Whitney)

(Wilcoxon) $\boldsymbol{W_Y = n_x n_Y + \frac{n_Y \cdot (n_Y + 1)}{2} - U_Y}$ (Mann-Whitney)

Poznámky k Mann-Whitneyovu *U* testu a Wilcoxonovu testu

- Oba testy prověřují nulovou hypotézu o shodě rozdělení, ze kterého pocházejí porovnávané výběry. Pokud testujeme nulovou hypotézu o shodě polohy (mediánu), musíme předpokládat, že se distribuce příliš neliší tvarem.

Mann-Whitney alias Wilcoxon - příklad:

R:
```
wilcox.test(x, y, alternative=c("two.sided","less","greater"),
  mu=0, paired=FALSE, exact=NULL, correct=TRUE, conf.int=FALSE,
  conf.level=0.95, ...)
```

Příklad: stromky

```
> wilcox.test(Health~Cultivar, data=stromky)

        Wilcoxon rank sum test with continuity correction

data:  Health by Cultivar
W = 40, p-value = 0.4619
alternative hypothesis: true location shift is not equal to 0


Warning message:
In wilcox.test.default(x = c(2, 2, 1, 2, 3, 4, 2, 3, 1, 5), y = c(4,  :
  cannot compute exact p-value with ties
```

Mediánový test

Existuje, je však velmi slabý.

Porovnání dvou pravděpodobností

tj. mám data o dvou znacích nominální proměnné a ptám se,
zda se dva výběry shodnou v pravděpodobnosti, že nastane znak A.
Toto odpovídá čtyřpolní kontingenční tabulce.
(je-li sledovaných znaků více, testuji obecnou kontingenční tabulku)

Předpoklad: mám dvě série vzájemně nezávislých pokusů, ve kterých zjišťuji, zda nastal znak (jev) A. Prst. znaku A v jedné sérii je stejný.

- Y_1 = počet pokusů, kdy nastal znak A v první sérii (celkem n_1)
- Y_2 = počet pokusů, kdy nastal znak A ve druhé sérii (celkem n_2)
- $Y_1 \sim Bi(n_1, p_1) \qquad Y_2 \sim Bi(n_2, p_2)$

Nulová hypotéza: obě pravděpodobnosti jsou shodné, $p_1 = p_2 = p$

Odhady: $\hat{p}_1 = \frac{Y_1}{n_1} \qquad \hat{p}_2 = \frac{Y_2}{n_2}$

$$var(\hat{p}_1 - \hat{p}_2) = \frac{p_1 \cdot (1 - p_1)}{n_1} + \frac{p_2 \cdot (1 - p_2)}{n_2} =_{H_0} p \cdot (1 - p) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Porovnání dvou pravděpodobností

$\boldsymbol{Y_1 \sim Bi(n_1, p_1)}$ $\qquad$ $\boldsymbol{Y_2 \sim Bi(n_2, p_2)}$

Nulová hypotéza: obě pravděpodobnosti jsou shodné, $\boldsymbol{p_1 = p_2 = p}$

Testovat můžeme trojím způsobem:

1) Přesný binomický test → R: binom.test
2) Přibližný test přes aproximaci normálním rozdělením

$$Z = \frac{(\hat{\boldsymbol{p}}_1 - \hat{\boldsymbol{p}}_2) - \boldsymbol{0}}{S.E.(\hat{\boldsymbol{p}}_1 - \hat{\boldsymbol{p}}_2)} = \frac{\hat{\boldsymbol{p}}_1 - \hat{\boldsymbol{p}}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{\mathbf{1}}{\boldsymbol{n_1}} + \frac{\mathbf{1}}{\boldsymbol{n_2}}\right)}} \sim N(0,1)$$

3) Chí-kvadrát test

   R: prop.test

   chisq.test