Osnova přednášky Vícerozměrné analogie t-testů

I. Úlohy o jednom náhodném výběru z vícerozměrného rozložení

1. Test hypotézy o vektoru středních hodnot
2. Příklad na vícerozměrný jednovýběrový t-test
3. Test hypotézy o úplné nezávislosti sledovaných proměnných
4. Příklad na test hypotézy o úplné nezávislosti sledovaných proměnných

II. Úlohy o dvou nezávislých náhodných výběrech z vícerozměrného rozložení

1. Test hypotézy o rozdílu vektorů středních hodnot
2. Test hypotézy o shodě variančních matic
3. Příklad na Hotellingův T^2 test

I. Úlohy o jednom náhodném výběru z vícerozměrného rozložení

1. Test hypotézy o vektoru středních hodnot

Tento test je p-rozměrnou analogií jednovýběrového t-testu. Pro připomenutí:

Náhodný výběr $X_1, \ldots, X_n$ pochází z rozložení $N(\mu, \sigma^2)$, kde parametry μ, σ^2 neznáme. Na hladině významnosti α testujeme hypotézu $H_0 : \mu = c$ proti alternativě $H_1 : \mu \neq c$.

Testová statistika: $T_0 = \dfrac{M - c}{\dfrac{S}{\sqrt{n}}}$ se za platnosti H_0 řídí rozložením $t(n-1)$.

Kritický obor: $W = (-\infty, -t_{1-\alpha/2}(n-1)\rangle \cup \langle t_{1-\alpha/2}(n-1), \infty)$.

Jestliže $t_0 \in W$, H_0 zamítáme na hladině významnosti α.

Poznámka: Vzhledem k tomu, že platí tvrzení: $X \sim t(n) \Rightarrow Y = X^2 \sim F(1, n)$, můžeme H_0 zamítnout na hladině významnosti α, když $t_0^{\,2} \in \langle F_{1-\alpha}(1, n-1), \infty)$.

p-rozměrný případ:

Náhodný výběr $\mathbf{X_1}, \ldots, \mathbf{X}_n$ pochází z rozložení $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, kde parametry $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ neznáme. Na hladině významnosti α testujeme hypotézu $H_0 : \boldsymbol{\mu} = \mathbf{c}$ proti alternativě $H_1 : \boldsymbol{\mu} \neq \mathbf{c}$, kde $\mathbf{c} = (c_1, \ldots, c_p)^T$ je vektor reálných konstant. (Alternativa vlastně tvrdí, že aspoň jedna složka vektoru středních hodnot neodpovídá ověřovanému předpokladu.)

Testová statistika $T_0 = \frac{n(n-p)}{p(n-1)}(\mathbf{M}-\mathbf{c})^T \mathbf{S}^{-1}(\mathbf{M}-\mathbf{c})$ se za platnosti H_0 řídí rozložením $F(p, n-p)$.
Kritický obor: $W = \langle F_{1-\alpha}(p, n-p), \infty)$.
Jestliže $t_0 \in W$, H_0 zamítáme na hladině významnosti α.

Poznámka: Test $H_0 : \boldsymbol{\mu} = \mathbf{c}$ proti $H_1 : \boldsymbol{\mu} \neq \mathbf{c}$ nelze nahradit p jednorozměrnými t-testy $H_{0j} : \mu_j = c_j$ proti $H_{1j} : \mu_j \neq c_j$, $j = 1, \ldots, p$, protože při tomto postupu by pravděpodobnost chyby 1. druhu byla větší než α, dokonce až $1-(1-\alpha)^p$.

Pokud na dané hladině významnosti α zamítneme vícerozměrnou hypotézu $H_0 : \boldsymbol{\mu} = \mathbf{c}$ ve prospěch alternativy $H_1 : \boldsymbol{\mu} \neq \mathbf{c}$, zjistíme, vzhledem ke kterým složkám vektoru $\boldsymbol{\mu}$ byla nulová hypotéza zamítnuta.

K tomu lze použít p jednorozměrných t-testů $H_{0j} : \mu_j = c_j$ proti $H_{1j} : \mu_j \neq c_j$, $j = 1, \ldots, p$, u nichž hladinu významnosti α upravíme pomocí Bonferroniho korekce:

H_{0j} zamítneme na hladině významnosti α, když vypočtená p-hodnota bude $\leq \frac{\alpha}{p}$.

2. Příklad na vícerozměrný jednovýběrový t-test

Výrobce určitého typu součástek uvádí, že nejdůležitější čtyři rozměry nabývají těchto hodnot: 9,50 mm, 6,35 mm, 5,98 mm a 4,40 mm. Náhodně bylo vybráno 15 součástek, byly u nich zjištěny hodnoty těchto rozměrů a zapsány do proměnných X1, X2, X3, X4. Údaje jsou uloženy v souboru soucastky.sta.

Za předpokladu, že data pocházejí ze čtyřrozměrného normálního rozložení s neznámým vektorem středních hodnot $\boldsymbol{\mu} = (\mu_1 \quad \mu_2 \quad \mu_3 \quad \mu_4)^T$ a neznámou varianční maticí

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{pmatrix},$$

na hladině významnosti 0,05 testujte hypotézu, že tvrzení výrobce je pravdivé. V případě zamítnutí nulové hypotézy zjistěte, které rozměry přispěly k jejímu zamítnutí.

Řešení:

Na hladině významnosti 0,05 testujeme hypotézu H_0: $\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix} = \begin{pmatrix} 9{,}50 \\ 6{,}35 \\ 5{,}98 \\ 4{,}40 \end{pmatrix}$

proti alternativě H_1: $\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix} \neq \begin{pmatrix} 9{,}50 \\ 6{,}35 \\ 5{,}98 \\ 4{,}40 \end{pmatrix}$.

Hodnotu testové statistiky $T_0 = \frac{n(n-p)}{p(n-1)}(\mathbf{M}-\mathbf{c})^T \mathbf{S}^{-1}(\mathbf{M}-\mathbf{c})$ a odpovídající p-hodnotu vypočteme pomocí statistického software.

Proměnná	Test průměrů vůči referenční konstantě (hodnotě) (soucastky.sta) T2(celé případy ChD)=19,2432 F(4,11)=3,7799 p<,03597							
	Průměr	Sm.odch.	N	Sm.chyba	Referenční konstanta	t	SV	p
X1	9,491833	0,010695	15	0,002761	9,500000	-2,95748	14	0,010391
X2	6,357433	0,011481	15	0,002964	6,350000	2,50752	14	0,025099
X3	5,981467	0,011129	15	0,002873	5,980000	0,51043	14	0,617706
X4	4,400327	0,007024	15	0,001814	4,400000	0,18011	14	0,859646

Testová statistika vícerozměrného jednovýběrového t-testu se realizuje hodnotou 3,7799, odpovídající p-hodnota je 0,03597, tedy s rizikem omylu nejvýše 5 % považujeme za prokázané, že rozměry součástky neodpovídají deklarovaným hodnotám.

Protože jsme zamítli nulovou hypotézu, v dalším kroku zjistíme, které rozměry přispěly k jejímu zamítnutí. Budeme tedy simultánně testovat hypotézy $H_{01}: \mu_1 = 9,5$, $H_{02}: \mu_2 = 6,35$, $H_{03}: \mu_3 = 5,98$, $H_{04}: \mu_4 = 4,4$ proti $H_{11}: \mu_1 \neq 9,5$, $H_{12}: \mu_2 \neq 6,35$, $H_{13}: \mu_3 \neq 5,98$, $H_{14}: \mu_4 \neq 4,4$. H_{0j} zamítneme na hladině významnosti $\alpha = 0,05$, když vypočtená p-hodnota bude menší nebo rovna $\frac{\alpha}{\text{počet testů}} = \frac{0,05}{4} = 0,0125$. Vidíme, že vícerozměrná hypotéza byla zamítnuta kvůli X1.

Výpočet pomocí systému R

Načteme data:

```
soucastky<-read.excel('soucastky.csv')
```

Vypočteme vektor výběrových průměrů a výběrovou varianční matici:

```
colMeans(soucastky)
      X1       X2       X3       X4
9.491833 6.357433 5.981467 4.400327
var(soucastky)
              X1            X2            X3            X4
X1  1.143767e-04  8.879524e-06  2.229119e-05 -2.579524e-06
X2  8.879524e-06  1.318167e-04 -2.551810e-05 -8.874524e-06
X3  2.229119e-05 -2.551810e-05  1.238481e-04  4.012381e-06
X4 -2.579524e-06 -8.874524e-06  4.012381e-06  4.934210e-05
```

Ověříme vícerozměrnou normalitu dat. Načteme knihovnu mvnTest:

```
library(mvnTest)
```

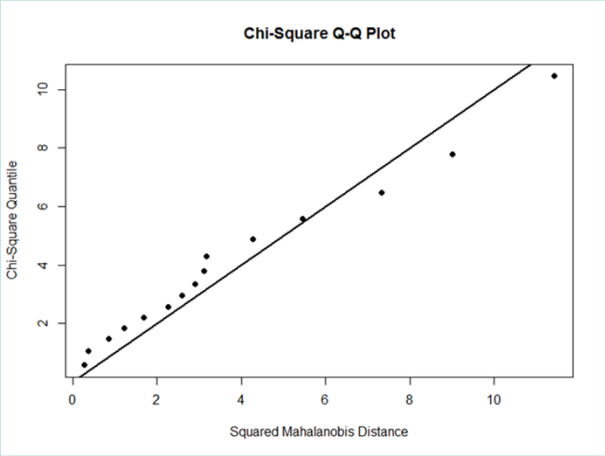
Provedeme Henzeův - Zirklerův test a nakreslíme chí-kvadrát diagram:

```
HZ.test(soucastky,qqplot=T)
            Henze-Zirkler test for Multivariate Normality

  data : soucastky

  HZ              : 0.8046086
  p-value         : 0.101327

  Result  : Data are multivariate normal (sig.level = 0.05)
```



Pomocí jednovýběrového Hotellingova testu otestujeme hypotézu, že vektor středních hodnot je roven zadanému vektoru **c**:

```
c<-c(9.5, 6.35, 5.98, 4.4)
```

Načteme knihovnu ICSNP:

```
library(ICSNP)
```

Provedeme Hotellingův jednovýběrový test:

```
HotellingsT2(soucastky,mu=c)

      Hotelling's one sample T2-test

data:  soucastky
T.2 = 3.7799, df1 = 4, df2 = 11, p-value = 0.03597
alternative hypothesis: true location is not equal to c(9.5,6.35,5.98,4.4)
```

Protože jsme na hladině významnosti 0,05 zamítli hypotézu, že vektor středních hodnot je roven danému vektoru, zjistíme nyní, které rozměry součástek k tomu přispěly. Provedeme jednovýběrové t-testy s Bonferroniho korekcí.

Vypočteme korigovanou hladinu významnosti:

```
(alfa.korig<-0.05/4)
[1] 0.0125
```

Postupně pro každou proměnnou provedeme jednovýběrový t-test. Vypočtenou p-hodnotu porovnáme s 0,0125 a nulovou hypotézu zamítneme, když $p \leq 0{,}0125$.

Test pro proměnnou X1:

```
t.test(soucastky$X1,mu=c[1])
      One Sample t-test
data:  soucastky$X1
t = -2.9575, df = 14, p-value = 0.01039
alternative hypothesis: true mean is not equal to 9.5
95 percent confidence interval:
 9.485911 9.497756
sample estimates:
mean of x
 9.491833
```

Test pro proměnnou X2:

```
t.test(soucastky$X2,mu=c[2])
      One Sample t-test
data:  soucastky$X2
t = 2.5075, df = 14, p-value = 0.0251
alternative hypothesis: true mean is not equal to 6.35
95 percent confidence interval:
 6.351075 6.363791
sample estimates:
mean of x
 6.357433
```

Test pro proměnnou X3:

```
t.test(soucastky$X3,mu=c[3])
      One Sample t-test
data:  soucastky$X3
t = 0.51043, df = 14, p-value = 0.6177
alternative hypothesis: true mean is not equal to 5.98
95 percent confidence interval:
 5.975304 5.987630
sample estimates:
mean of x
 5.981467
```

Test pro proměnnou X4:

```
t.test(soucastky$X4,mu=c[4])
      One Sample t-test
data:  soucastky$X4
t = 0.18011, df = 14, p-value = 0.8596
alternative hypothesis: true mean is not equal to 4.4
95 percent confidence interval:
 4.396437 4.404217
sample estimates:
mean of x
 4.400327
```

Vidíme tedy, že vícerozměrná hypotéza byla zamítnuta kvůli proměnné X1.

3. Test hypotézy o úplné nezávislosti sledovaných proměnných

Řada statistických úloh vede na zkoumání závislosti mezi p sledovanými proměnnými. Nejdříve by se mělo zjistit, zda se nejedná o systém nezávislých proměnných. V takovém případě by bylo zbytečné pokračovat v analýze závislostí.

Na hladině významnosti 0,05 testujeme $H_0 : \mathrm{cor}\mathbf{X} = \mathbf{I}$ proti $H_0 : \mathrm{cor}\mathbf{X} \neq \mathbf{I}$ ($\mathbf{I}$ je jednotková matice řádu p).

Testová statistika $T_0 = -n\left(1 - \frac{2p+11}{6n}\right)\ln|\mathbf{R}|$ se za platnosti H_0 asymptoticky řídí rozložením $\chi^2\left(\frac{p(p-1)}{2}\right)$.

Kritický obor: $W = \left\langle \chi^2_{1-\alpha}\left(\frac{p(p-1)}{2}\right), \infty\right)$

Jestliže $t_0 \in W$, H_0 zamítáme na hladině významnosti α.

4. Příklad: Na základě dat z příkladu o rozměrech součástek testujte hypotézu, že mezi sledovanými čtyřmi rozměry není žádná závislost.

Řešení:

Logaritmus determinantu výběrové korelační matice je číslo $\ln|\mathbf{R}|$ = -0,10371221.

Testová statistika $t_0 = -n\left(1-\frac{2\cdot p+11}{6n}\right)\ln|\mathbf{R}| = -15\left(1-\frac{2\cdot 4+11}{6\cdot 15}\right)(-0{,}10371221) = 1{,}2273$.

Kritický obor $W = \langle \chi^2_{0,95}(6), \infty) = \langle 12{,}5916; \infty)$

Protože testová statistika 1,2273 nepatří do kritického oboru $\langle 12{,}5916; \infty)$, hypotézu o úplné nezávislosti čtyř rozměrů součástek nezamítáme na hladině významnosti 0,05.

Výpočet pomocí systému R

Vypočteme výběrovou korelační matici:

```
R<-cor(soucastky)
```

Načteme knihovnu psych:

```
library(psych)
```

Provedeme Bartlettův test nezávislosti:

```
cortest.bartlett(R,n=15,diag=T)
$chisq
[1] 1.227261

$p.value
[1] 0.9755168

$df
[1] 6
```

Protože p-hodnota je 0,9755, což je větší než hladina významnosti 0,05, hypotézu o úplné nezávislosti čtyř rozměrů součástek nezamítáme na hladině významnosti 0,05.

II. Úlohy o dvou nezávislých náhodných výběrech z vícerozměrného rozložení

1. Test hypotézy o rozdílu vektorů středních hodnot

Tento test je p-rozměrnou analogií dvouvýběrového t-testu. Pro připomenutí:
Náhodný výběr $X_{11},\dots,X_{1n_1}$ pochází z rozložení $N(\mu_1,\sigma^2)$, na něm nezávislý náhodný výběr $X_{21},\dots,X_{2n_2}$ pochází z rozložení $N(\mu_2,\sigma^2)$, přičemž parametry μ_1,μ_2,σ^2 neznáme. Označíme M_1, M_2 výběrové průměry, $S_1^{\,2}, S_2^{\,2}$ výběrové rozptyly, $S_*^{\,2} = \dfrac{(n_1-1)S_1^{\,2}+(n_2-1)S_2^{\,2}}{n_1+n_2-2}$ vážený průměr výběrových rozptylů. Na hladině významnosti α testujeme hypotézu $H_0 : \mu_1 = \mu_2$ proti alternativě $H_1 : \mu_1 \neq \mu_2$.

Testová statistika: $T_0 = \dfrac{M_1 - M_2}{S_*\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$ se za platnosti H_0 řídí rozložením $t(n_1+n_2-2)$.

Kritický obor: $W = (-\infty, -t_{1-\alpha/2}(n_1+n_2-2)\rangle \cup \langle t_{1-\alpha/2}(n_1+n_2-2), \infty)$.

Jestliže $t_0 \in W$, H_0 zamítáme na hladině významnosti α.

Upozornění: Předpoklad, že rozptyly obou rozložení jsou shodné (tj. test nulové hypotézy $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ proti alternativě $H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$) ověřujeme F-testem.

Testová statistika $T_0 : \frac{S_1^2}{S_2^2}$ se v případě platnosti H_0 řídí rozložením $F(n_1 - 1, n_2 - 1)$.

Kritický obor: $W = \langle 0, F_{\alpha/2}(n_1 - 1, n_2 - 1)\rangle \cup \langle F_{1-\alpha/2}(n_1 - 1, n_2 - 1), \infty)$.

Jestliže $t_0 \in W$, H_0 zamítáme na hladině významnosti α.

p- rozměrný případ (Hotellingův T^2 test)

Máme náhodný výběr $\mathbf{X}_{11},\ldots,\mathbf{X}_{1n_1}$ (kde $\mathbf{X}_{1i}=\left(X_{1i1},\ldots,X_{1ip}\right)^T$, $i=1,\ldots,n_1$) z $N_p\left(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}\right)$ a dále na něm nezávislý náhodný výběr $\mathbf{X}_{21},\ldots,\mathbf{X}_{2n_2}$ (kde $\mathbf{X}_{2i}=\left(X_{2i1},\ldots,X_{2ip}\right)^T$, $i=1,\ldots,n_2$) z $N_p\left(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}\right)$, přičemž parametry $\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\boldsymbol{\Sigma}$ neznáme.

Zavedeme označení:

$n=n_1+n_2$ … celkový rozsah obou výběrů

$M_{hj}=\frac{1}{n_h}\sum_{i=1}^{n_h}X_{hij}$ … výběrový průměr j-té proměnné v h-tém výběru, $h=1,2,\ j=1,\ldots,p$

$\mathbf{M}_h=\left(M_{h1}\quad\ldots\quad M_{hp}\right)^T$ … vektor výběrových průměrů v h-tém výběru, $h=1,2$

$\mathbf{S}_h=\frac{1}{n_h-1}\sum_{i=1}^{n_h}\left(\mathbf{X}_{hi}-\mathbf{M}_h\right)\left(\mathbf{X}_{hi}-\mathbf{M}_h\right)^T$ … výběrová varianční matice v h-tém výběru, $h=1,2$

$\mathbf{S}=\frac{\left(n_1-1\right)\mathbf{S}_1+\left(n_2-1\right)\mathbf{S}_2}{n-2}$ … společná výběrová varianční matice

Na hladině významnosti α testujeme hypotézu $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ proti alternativě $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$.

Testová statistika $T_0 = \frac{n-p-1}{p(n-2)} \cdot \frac{n_1 n_2}{n} (\mathbf{M}_1 - \mathbf{M}_2)^T \mathbf{S}^{-1} (\mathbf{M}_1 - \mathbf{M}_2)$ se za platnosti H_0 řídí rozložením $F(p, n-p-1)$.

Kritický obor: $W = \langle F_{1-\alpha}(p, n-p-1), \infty)$.

Jestliže $t_0 \in W$, H_0 zamítáme na hladině významnosti α.

2. Test hypotézy o shodě variančních matic

Předpoklad o shodě variančních matic můžeme ověřit pomocí Boxova M-testu.
Na hladině významnosti α testujeme hypotézu $H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ proti alternativě $H_1 : \boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$.

Testová statistika má tvar: $T_0 = \frac{1}{C_p}\left[(n-2)\ln|\mathbf{S}| - (n_1-1)\ln|\mathbf{S}_1| - (n_2-1)\ln|\mathbf{S}_2|\right]$, kde

$C_p = 1 + \frac{2p^2+3p-1}{6(p+1)}\left(\frac{1}{n_1-1} + \frac{1}{n_2-1} - \frac{1}{n-2}\right)$ je konstanta zlepšující aproximaci.

V případě platnosti H_0 se statistika T_0 asymptoticky řídí rozložením $\chi^2\left(\frac{p(p+1)}{2}\right)$. Pokud

$t_0 \in \left\langle \chi^2_{1-\alpha}\left(\frac{p(p+1)}{2}\right), \infty\right)$, hypotézu o shodě variančních matic zamítneme na asymptotické hladině významnosti α. Aproximace je vyhovující, když rozsahy výběrů jsou aspoň 20 a počet proměnných je nejvýše 5.

V případě, že rozsahy výběrů jsou shodné, nemusíme Boxův test provádět.

Simultánní t-testy:

Pokud na dané hladině významnosti α zamítneme hypotézu $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ ve prospěch alternativy $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, zjistíme, které proměnné jsou příčinou jejího zamítnutí.
V této situaci provedeme p simultánních testů $H_{0j} : \mu_{1j} = \mu_{2j}$ proti $H_{1j} : \mu_{1j} \neq \mu_{2j}$, $j = 1, \ldots, p$ pomocí testové statistiky $T_{0j} = \frac{n-p-1}{p(n-2)} \cdot \frac{n_1 n_2}{n} \cdot \frac{(M_{1j} - M_{2j})^2}{S_{*j}^{\,2}}$, která se za platnosti H_{0j} řídí rozložením $F(p, n-p-1)$.

Kritický obor: $W = \langle F_{1-\alpha}(p, n-p-1), \infty)$.
Jestliže $t_{0j} \in W$, H_{0j} zamítáme na hladině významnosti α.

3. Příklad na Hotellingův T^2 test

23 náhodně vybraných mužů a 22 náhodně vybraných žen mělo posoudit podobné výrobky od tří firem – označme je A, B, C – na škále 0 bodů (naprosto nevyhovující) až 10 bodů (zcela vyhovující). Výsledky jsou uloženy v souboru hodnoceni_vyrobku.sta.
Za předpokladu, že data tvoří realizace dvou nezávislých náhodných výběrů ze dvou třírozměrných normálních rozložení se stejnými variančními maticemi, Hotellingovým T^2 testem ověřte na hladině významnosti 0,05 hypotézu, že hodnocení mužů a žen se neliší. Pokud dojde k zamítnutí nulové hypotézy, zjistěte, které firmy se v hodnocení mužů a žen liší.

Řešení:

Na hladině významnosti 0,05 testujeme hypotézu

$$H_0: \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \end{pmatrix} = \begin{pmatrix} \mu_{21} \\ \mu_{22} \\ \mu_{23} \end{pmatrix} \text{ proti alternativě } H_1: \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \end{pmatrix} \neq \begin{pmatrix} \mu_{21} \\ \mu_{22} \\ \mu_{23} \end{pmatrix}.$$

Hodnotu testové statistiky $T_0 = \frac{n-p-1}{p(n-2)} \cdot \frac{n_1 n_2}{n} (\mathbf{M}_1 - \mathbf{M}_2)^T \mathbf{S}^{-1} (\mathbf{M}_1 - \mathbf{M}_2)$ a odpovídající p-hodnotu vypočteme pomocí statistického software.

Proměnná	t-testy; grupováno: ID: pohlaví respondenta (hodnoceni_vyrobku.sta) Skup. 1: muž; Skup. 2: žena Hotellingovo 15,5599 F(3,41)=4,9454 p<,00506										
	Průměr muž	Průměr žena	t	sv	p	Poč.plat muž	Poč.plat. žena	Sm.odch. muž	Sm.odch. žena	F-poměr Rozptyly	p Rozptyly
X1	5,086957	4,545455	0,697666	43	0,489142	23	22	2,574579	2,631807	1,044950	0,917081
X2	5,434783	3,818182	2,098562	43	0,041766	23	22	2,642762	2,519190	1,100510	0,829044
X3	5,304348	3,045455	3,117687	43	0,003246	23	22	2,770540	2,011332	1,897411	0,147512

Testová statistika Hotellingova testu nabývá hodnoty 4,9454, odpovídající p-hodnota je menší než 0,00506, tedy na hladině významnosti 0,05 zamítáme hypotézu, že vektory středních hodnot proměnných X1, X2, X3 jsou v obou skupinách shodné. S rizikem omylu nejvýše 5 % jsme tedy prokázali, že mezi muži a ženami existuje rozdíl v hodnocení výrobků tří firem. (Vidíme, že hodnocení mužů je příznivější než hodnocení žen.)

Nyní pomocí simultánních testů zjistíme, které firmy jsou rozdílně hodnoceny muži a ženami. Pro simultánní testy musíme spočítat statistiky

$T_{0j} = \frac{n-p-1}{p(n-2)} \cdot \frac{n_1 n_2}{n} \cdot \frac{(M_{1j} - M_{2j})^2}{S_{*j}^{\ 2}}$, j = 1, 2, 3 a najít kvantil $F_{0,95}(3, 41)$.

V našem případě n = 45, p = 3, n_1 = 23, n_2 = 22, tedy $\frac{n-p-1}{p(n-2)} \cdot \frac{n_1 n_2}{n} = \frac{20746}{5805}$.)

Proměnná	Průměr muž	Průměr žena	Sm.odch. muž	Sm.odch. žena	T0j =(20746/58	kvantil =VF(0,95;3;
X1	5,086957	4,545455	2,574579	2,631807	0,154700	2,832747
X2	5,434783	3,818182	2,642762	2,519190	1,399710	2,832747
X3	5,304348	3,045455	2,770540	2,011332	3,089293	2,832747

Vidíme, že statistika T03 se realizuje v kritickém oboru $W = \langle 2{,}8327; \infty)$. S rizikem omylu nejvýše 5 % jsme tedy prokázali, že výrobky firmy C jsou odlišně hodnoceny muži a ženami.

Výpočet pomocí systému R

Data jsou uložena v excelovském souboru hodnoceni_vyrobku.xls. Načteme je pomocí Environment v prostředí R Studia:

Environment – Import Dataset – From Excel – Browse – najdeme soubor hodnoceni_vyrobku.xls – Otevřít – Import.

Načtený soubor přejmenujeme:

```
data<-hodnoceni_vyrobku
```

Proměnnou ID zavedeme jako faktor a popíšeme její varianty:

```
data$ID<-factor(data$ID,levels=c(1,2),labels=c("muz","zena"))
```

Vypočítáme průměry pro muže a pro ženy:

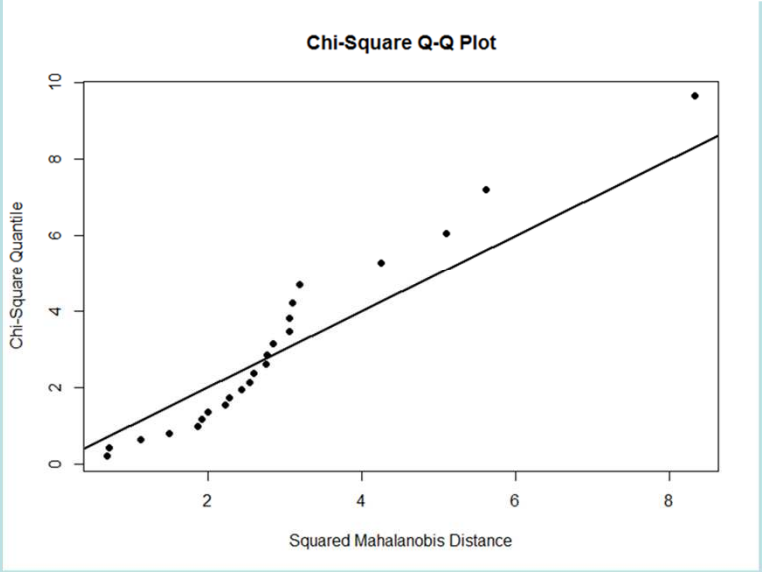
```
colMeans(data[data$ID=="muz",1:3])
      X1       X2       X3
5.086957 5.434783 5.304348
> colMeans(data[data$ID=="zena",1:3])
      X1       X2       X3
4.545455 3.818182 3.045455
```

Vypočítáme výběrové varianční matice pro muže a pro ženy:

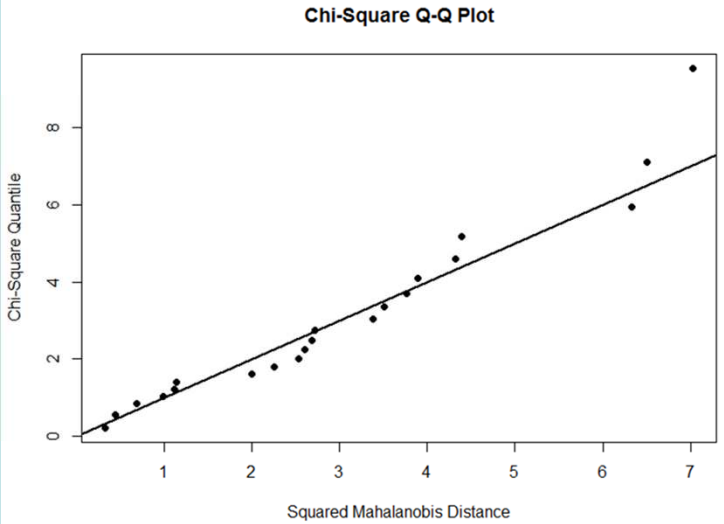
```
var(data[data$ID=="muz",1:3])
         X1       X2       X3
X1 6.628458 5.142292 5.063241
X2 5.142292 6.984190 5.679842
X3 5.063241 5.679842 7.675889
var(data[data$ID=="zena",1:3])
         X1       X2       X3
X1 6.926407 4.532468 4.307359
X2 4.532468 6.346320 4.294372
X3 4.307359 4.294372 4.045455
```

Nyní pomocí H-Z testu ověříme, zda data pro muže a pro ženy pocházejí z třírozměrného normálního rozložení.

```
HZ.test(data[data$ID=="muz",1:3],qqplot=T)
            Henze-Zirkler test for Multivariate Normality
  data : data[data$ID == "muz", 1:3]
  HZ               : 0.5900191
  p-value          : 0.4456571
  Result  : Data are multivariate normal (sig.level = 0.05)
```



```
HZ.test(data[data$ID=="zena",1:3],qqplot=T)
             Henze-Zirkler test for Multivariate Normality
  data : data[data$ID == "zena", 1:3]
  HZ               : 0.4815084
  p-value          : 0.7512396
  Result  : Data are multivariate normal (sig.level = 0.05)
```



K ověření shody variančních matic použijeme Boxův M-test z knihovny biotools:

```
library(biotools)
boxM(data[,1:3],grouping=data$ID)
      Box's M-test for Homogeneity of Covariance Matrices
data:  data[, 1:3]
Chi-Sq (approx.) = 9.3635, df = 6, p-value = 0.1541
```

Předpoklady jsou splněny, přistoupíme k provedení Hotellingova dvouvýběrového testu:

```
library(ICSNP)
HotellingsT2(data[data$ID=="muz",1:3], data[data$ID=="zena",1:3])

     Hotelling's two sample T2-test

data:  data[data$ID == "muz", 1:3] and data[data$ID == "zena", 1:3]
T.2 = 4.9454, df1 = 3, df2 = 41, p-value = 0.005062
alternative hypothesis: true location difference is not equal to c(0,0,0)
```

Provedeme simultánní testy.

```
n1<-table(data$ID)[1]
n2<-table(data$ID)[2]
n<-n1+n2
k<-3
m1<-colMeans(data[data$ID=="muz",1:3])
m2<-colMeans(data[data$ID=="zena",1:3])
var1 <- diag(cov(data[data$ID=="muz",1:3]))
var2 <- diag(cov(data[data$ID=="zena",1:3]))
var <- ( (n1-1)*var1 + (n2-1)*var2 )/(n-2)
F.stat <- n1*n2*(n-k-1) * (m1-m2)^2 /(var*n*k*(n-2))
p.hodnota <- 1-pf(F.stat, k, n-k-1)
kvantil <- qf(0.95, k, n-k-1)
tab <- round(rbind(F.stat,p.hodnota, kvantil),digits=4)
rownames(tab) <- c("F","p-hodnota", "kvantil")
tab
                X1     X2     X3
F          0.1547 1.3997 3.0893
p-hodnota  0.9261 0.2566 0.0375
kvantil    2.8327 2.8327 2.8327
```

Rozdílné hodnocení mužů a žen je na hladině významnosti 0,05 prokazatelné u firmy C.