Vzorový příklad na vícerozměrné obdoby t- testů

Příklad na vícerozměrný jednovýběrový t-test: V souboru mlrm-fat.txt máme k dispozici antropometrická data mladých zdravých dospělých žen (převážně studentek vysokých škol z Brna). Zajímají nás proměnné tělesná hmotnost (proměnná body.W), tělesná výška (body.H), tloušťka kožní řasy ve výši 10. žebra (rib.F), tloušťka kožní rasy na břiše (abdo.F), tloušťka kožní rasy na boku (hip.F) a tloušťka kožní řasy nad čtyřhlavým svalem stehenním (quad.H). Hmotnost byla měřena v kg, tělesná výška v cm, ostatní veličiny v mm. Chceme otestovat hypotézu H_0: $(\mu_1\ \mu_2\ \mu_3\ \mu_4\ \mu_5\ \mu_6)^T = (60{,}8\ \ 167{,}9\ \ 13{,}0\ \ 21{,}5\ \ 22{,}0\ \ 25{,}0)^T$ proti alternativní hypotéze H_1: $(\mu_1\ \mu_2\ \mu_3\ \mu_4\ \mu_5\ \mu_6)^T \neq (60{,}8\ \ 167{,}9\ \ 13{,}0\ \ 21{,}5\ \ 22{,}0\ \ 25{,}0)^T$.
S proměnnou BMI se v tomto příkladu nepracuje! Ze souboru odstraníme pozorování 36, které jsme v předchozích cvičeních identifikovali jako odlehlé.
Načteme datový soubor a vynecháme z něj proměnnou, která nás nezajímá, a odlehlé pozorování.

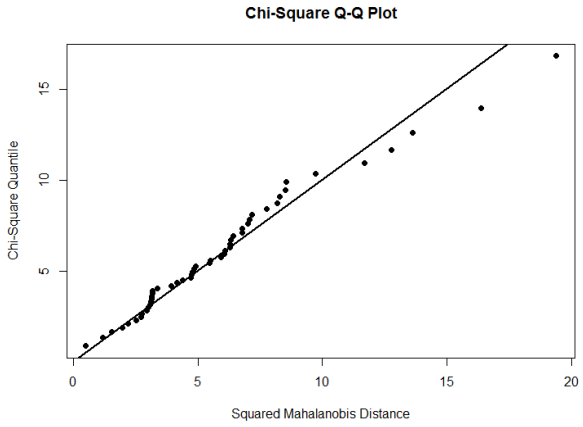
```
fat <- read.table('mlrm-fat.txt', header=T)
str(fat)
'data.frame':  51 obs. of  7 variables:
 $ body.W: num  53.3 49.3 53.3 61.2 65.4 64.3 62.4 60.2 54.3 58.6 ...
 $ body.H: num  165 162 179 171 174 ...
 $ BMI   : num  19.6 18.8 16.6 20.9 21.6 ...
 $ rib.F : num  10.2 12.8 9.2 13.8 19.6 14.2 17.2 16.8 9.2 12.6 ...
 $ abdo.F: num  17 17.8 13.4 16.6 24.8 29 25.8 25.2 17 23.4 ...
 $ hip.F : num  24.8 20.4 9.2 19.4 25.2 29.2 25.8 27.2 10.4 26.2 ...
 $ quad.H: num  22.4 25.8 25.4 24.2 27.8 27.2 31.2 18.4 15.8 28.4 ...
fat2 <- fat[-36, -3]
```

Vypočítáme vektor výběrových průměrů a výběrovou varianční matici.

```
colMeans(fat2)
 body.W  body.H   rib.F  abdo.F   hip.F  quad.H
 58.452 167.350  12.684  20.662  19.898  23.478

var(fat2)
          body.W     body.H     rib.F     abdo.F     hip.F     quad.H
body.W 28.582547  9.7222449 11.396359 11.8873224 17.619698 16.7972898
body.H  9.722245 34.5841837 -2.434082  0.5319388 -2.258469 -0.2441837
rib.F  11.396359 -2.4340816 12.550351 10.4855020 15.249355 12.9431102
abdo.F 11.887322  0.5319388 10.485502 20.7652612 19.670331 11.3466980
hip.F  17.619698 -2.2584694 15.249355 19.6703306 39.452037 23.4248531
quad.H 16.797290 -0.2441837 12.943110 11.3466980 23.424853 33.3033837
```

Je potřeba ověřit předpoklad, že data pocházejí z šestirozměrného normálního rozdělení. Za tímto účelem můžeme použít Henzeův-Zirklerův test nebo Cramérův-von Misesův test dobré shody, v obou můžeme volbou qqplot=T vykreslit kvantil-kvantilový graf.

```
library(mvnTest)

HZ.test(fat2, qqplot=T)
            Henze-Zirkler test for Multivariate Normality
  data : fat2
  HZ              : 0.8464247
  p-value         : 0.7212326
  Result  : Data are multivariate normal (sig.level = 0.05)

CM.test(fat2)
            Cramer-von Mises test for Multivariate Normality
  data : fat2
  CM              : 0.05588075
  p-value         : 0.6551345
  Result  : Data are multivariate normal (sig.level = 0.05)
```



Hodnota testové statistiky pro
H-Z test =
C-M test =

p-hodnota pro H-Z test =
p-hodnota pro C-M test =

Na hladině významnosti 0,05
H-Z test šestirozměrnou normalitu
…….

Na hladině významnosti 0,05
C-M test šestirozměrnou normalitu
…….

Pomocí jednovýběrového Hotellingova testu otestujeme hypotézu, že vektor středních hodnot je roven zadanému vektoru.

```
mu0 <- c(60.8, 167.9, 13, 21.5, 22, 25)
library("ICSNP")
HotellingsT2(fat2, mu=mu0)
        Hotelling's one sample T2-test
data:  fat2
T.2 = 2.5518, df1 = 6, df2 = 44, p-value = 0.03302
alternative hypothesis: true location is not equal to
c(60.8,167.9,13,21.5,22,25)
```

Hodnota testové statistiky = , p-hodnota = , závěr:

Protože jsme na hladině významnosti 0,05 ..... hypotézu, že vektor středních hodnot je roven $(60{,}8\ \ 167{,}9\ \ 13{,}0\ \ 21{,}5\ \ 22{,}0\ \ 25{,}0)^T$, chceme zjistit, které proměnné to způsobují. Provedeme proto jednorozměrné t-testy, u nichž musíme upravit hladinu významnosti pomocí Bonferroniho korekce (hladinu významnosti dělíme počtem proměnných):

```
alpha.korig <- 0.05 / 6
> alpha.korig
[1] 0.008333333
```

U jednorozměrných t-testů tedy budeme zamítat hypotézu v případě, že p-hodnota bude menší než 0,00833.

```
t.test(fat2$body.W, mu=mu0[1])
        One Sample t-test
data:  fat2$body.W
t = -3.1055, df = 49, p-value = 0.003155
alternative hypothesis: true mean is not equal to 60.8
95 percent confidence interval:56.93261 59.97139
sample estimates: mean of x 58.452

t.test(fat2$body.H, mu=mu0[2])
        One Sample t-test
data:  fat2$body.H
t = -0.66132, df = 49, p-value = 0.5115
alternative hypothesis: true mean is not equal to 167.9
95 percent confidence interval: 165.6787 169.0213
sample estimates: mean of x 167.35
```

```
t.test(fat2$rib.F, mu=mu0[3])
        One Sample t-test
data:  fat2$rib.F
t = -0.63073, df = 49, p-value = 0.5311
alternative hypothesis: true mean is not equal to 13
95 percent confidence interval: 11.67719 13.69081
sample estimates: mean of x 12.684

t.test(fat2$abdo.F, mu=mu0[4])
        One Sample t-test
data:  fat2$abdo.F
t = -1.3004, df = 49, p-value = 0.1996
alternative hypothesis: true mean is not equal to 21.5
95 percent confidence interval: 19.36695 21.95705
sample estimates: mean of x 20.662

t.test(fat2$hip.F, mu=mu0[5])
        One Sample t-test
data:  fat2$hip.F
t = -2.3664, df = 49, p-value = 0.02196
alternative hypothesis: true mean is not equal to 22
95 percent confidence interval: 18.11294 21.68306
sample estimates: mean of x 19.898

t.test(fat2$quad.H, mu=mu0[6])
        One Sample t-test
data:  fat2$quad.H
t = -1.8649, df = 49, p-value = 0.06819
alternative hypothesis: true mean is not equal to 25
95 percent confidence interval: 21.83793 25.11807
sample estimates: mean of x 23.478
```

Nulová hypotéza	Testová statistika	p-hodnota	závěr
μ_1 = ...			
μ_2 = ...			
μ_3 = ...			
μ_4 = ...			
μ_5 = ...			
μ_6 = ...			

Vidíme, že vícerozměrná hypotéza byla zamítnuta kvůli proměnným ..... .

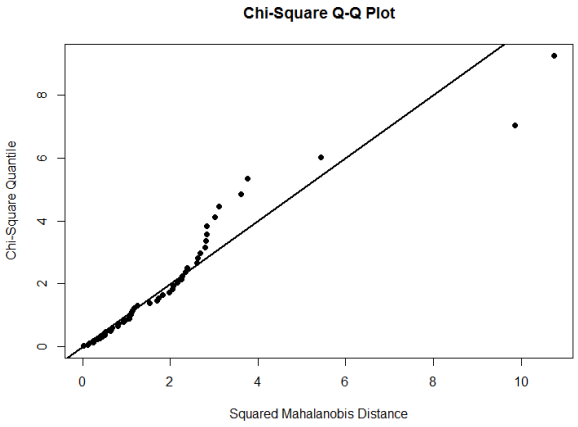
Příklad na vícerozměrný dvouvýběrový t-test: V souboru d2d4.txt máme k dispozici antropometrická data mladých dospělých lidí (převážně studentů z Brna a Ostravy) - tělesnou výšku (proměnná body.H), a poměr délky 2. a 4. prstu (proměnná d2d4). Známe také pohlaví sledovaných jedinců. Chceme otestovat hypotézu, že vektor středních hodnot sledovaných proměnných je stejný pro muže a pro ženy.
Načteme data a vypočítáme vektory výběrových průměrů a výběrové varianční matice zvlášť pro muže a pro ženy.

```
digits <- read.table("d2d4.txt", header=T)
str(digits)
'data.frame':  87 obs. of  4 variables:
 $ id    : int  2 4 6 8 10 12 14 16 18 20 ...
 $ sex   : Factor w/ 2 levels "f","m": 2 1 1 2 1 1 1 2 2 2 ...
 $ body.H: int  1824 1576 1676 1711 1579 1680 1602 1810 1830 1680 ...
```

```
 $ d2d4   : num  0.915 0.939 0.99 0.895 1.012 ...
colMeans(digits[digits$sex=='f', 3:4])
     body.H        d2d4
1658.372549    0.981012

var(digits[digits$sex=='f', 3:4])
             body.H       d2d4
body.H 4756.7184314 0.19805179
d2d4      0.1980518 0.00122236

colMeans(digits[digits$sex=='m', 3:4])
      body.H         d2d4
1780.6388889    0.9624943

var(digits[digits$sex=='m', 3:4])
             body.H         d2d4
body.H 2943.4944444 0.3078064436
d2d4      0.3078064 0.0008390637
```

Dále je potřeba ověřit předpoklady. Začneme předpokladem, že data pocházejí z dvourozměrného normálního rozdělení. Nejprve provedeme H-Z test a C-M test pro ženy:

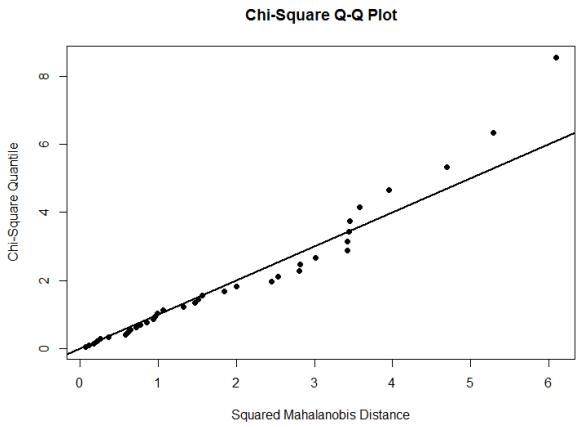
```
library(mvnTest)
HZ.test(digits[digits$sex=="f",3:4], qqplot=T)
            Henze-Zirkler test for Multivariate Normality
  data : digits[digits$sex == "f", 3:4]
  HZ               : 0.4857857
  p-value          : 0.4815272
  Result   : Data are multivariate normal (sig.level = 0.05)
```




```
CM.test(digits[digits$sex=="f",3:4])
            Cramer-von Mises test for Multivariate Normality
  data : digits[digits$sex == "f", 3:4]
  CM               : 0.1085218
  p-value          : 0.2841716
  Result   : Data are multivariate normal (sig.level = 0.05)
```

Pokračujeme H-Z testem a C-M testem pro muže:

```
HZ.test(digits[digits$sex=="m",3:4], qqplot=T)
            Henze-Zirkler test for Multivariate Normality
  data : digits[digits$sex == "m", 3:4]
  HZ               : 0.3238762
  p-value          : 0.7910886
  Result   : Data are multivariate normal (sig.level = 0.05)
```




```
CM.test(digits[digits$sex=="m",3:4])
            Cramer-von Mises test for Multivariate Normality
  data : digits[digits$sex == "m", 3:4]
  CM              : 0.06620061
  p-value         : 0.5794421
  Result  : Data are multivariate normal (sig.level = 0.05)
```

	H-Z test, ženy	C-M test, ženy	H-Z test, muži	C-M test, muži
Testová statistika				
p-hodnota				
závěr				

Dalším předpokladem, který je nutné ověřit, je shoda variančních matic. K tomu použijeme Boxův M test.

```
library("biotools")
boxM(digits[,3:4], grouping=digits$sex)
        Box's M-test for Homogeneity of Covariance Matrices
data:  digits[, 3:4]
Chi-Sq (approx.) = 4.1121, df = 3, p-value = 0.2496
```

Hodnota testovací statistiky .....
p-hodnota .....
Závěr .....

Předpoklady jsou splněny, můžeme tedy přikročit k dvouvýběrovému Hotellingovu T-testu.

```
library("ICSNP")
HotellingsT2(digits[digits$sex=="f",3:4], digits[digits$sex=="m",3:4])
        Hotelling's two sample T2-test
data:  digits[digits$sex == "f", 3:4] and digits[digits$sex == "m", 3:4]
T.2 = 45.553, df1 = 2, df2 = 84, p-value = 3.997e-14
alternative hypothesis: true location difference is not equal to c(0,0)
```

Hodnota testovací statistiky .....
p-hodnota .....
Závěr .....

Protože jsme na hladině významnosti 0,05 ..... hypotézu, že vektory středních hodnot mužů a žen jsou si rovny, provedeme simultánní testy. Využijeme přitom toho, že systém R pracuje s vektory po složkách.

```
n1 <- table(digits$sex)[1]
n2 <- table(digits$sex)[2]
n <- n1 + n2
k <- 2 #pocet promennych
mu1 <- colMeans(digits[digits$sex=="f",3:4])
mu2 <- colMeans(digits[digits$sex=="m",3:4])
var1 <- diag(cov(digits[digits$sex=="f",3:4]))
var2 <- diag(cov(digits[digits$sex=="m",3:4]))
var <- ( (n1-1)*var1 + (n2-1)*var2 )/(n-2)
F.stat <- n1*n2*(n-k-1) * (mu1-mu2)^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
            body.H   d2d4
F          38.8725 3.3589
p-hodnota   0.0000 0.0395
kvantil     3.1052 3.1052
```

	Tělesná výška	Poměr délky 2. a 4. prstu
Testová statistika		
p-hodnota		
Kritický obor		
Závěr		