

9 Jednofaktorová MANOVA

Příklad 1. V souboru Howell.csv máme k dispozici kraniometrické údaje z různých populací. Nás zajímají muži (kategorie M proměnné Sex) ze 3 populací (proměnná Population) - AINU, N JAPAN a PERU. Konkrétně máme tyto kraniometrické (vše v milimetrech):

- XFB - maximální transversální šířka čela,
- ZYB - bizygomatická šířka,
- ZMB - zygomaticomaxilární šířka.

Načteme datový soubor. Protože v databázi jsou chybějící pozorování kódovány jako 0, je potřeba při načítání zadat, aby se 0 braly jako NA. Vybereme pozorování a proměnné, které nás zajímají, a zbavíme se nyní prázdných kategorií proměnné Population.

```
cranio <- read.csv('DATA/Howell.csv',header=T, na.strings='0')
data <- cranio[cranio$Sex == 'M' & cranio$Population %in% c('AINU', 'N JAPAN', 'PERU'),
              c('Population', 'XFB', 'ZYB', 'ZMB')]
data$Population <- factor(data$Population)
```

Vypočítáme počet pozorování, vektor výběrových průměrů a výběrovou varianční matici pro každou populaci.

```
table(data$Population)

##
##      AINU N JAPAN      PERU
##      48      55      55

colMeans(data[data$Population=='AINU',-1])

##      XFB      ZYB      ZMB
## 119.64583 138.93750 99.16667

colMeans(data[data$Population=='N JAPAN',-1])

##      XFB      ZYB      ZMB
## 116.63636 135.72727 98.16364

colMeans(data[data$Population=='PERU',-1])

##      XFB      ZYB      ZMB
## 115.23636 134.92727 96.90909

var(data[data$Population=='AINU',-1])

##      XFB      ZYB      ZMB
## XFB 18.0208333 4.019947 -0.5141844
## ZYB 4.0199468 26.655585 12.0106383
## ZMB -0.5141844 12.010638 23.0354610

var(data[data$Population=='N JAPAN',-1])

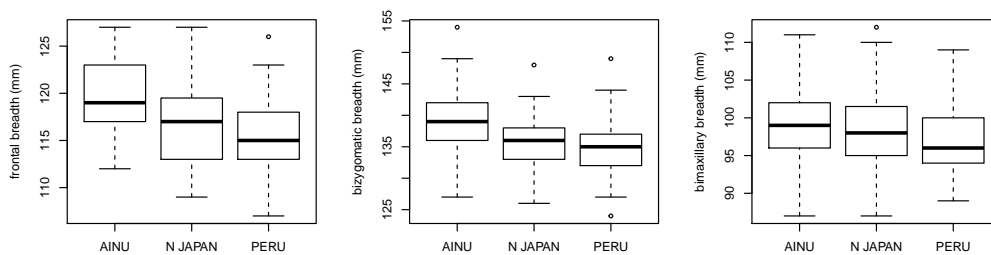
##      XFB      ZYB      ZMB
## XFB 20.3838384 3.010101 0.8013468
## ZYB 3.0101010 19.609428 14.7491582
## ZMB 0.8013468 14.749158 26.9171717
```

```
var(data[data$Population=='PERU',-1])
```

```
##           XFB           ZYB           ZMB
## XFB 19.109764  4.536027  3.836700
## ZYB  4.536027 18.327946  8.734007
## ZMB  3.836700  8.734007 17.454545
```

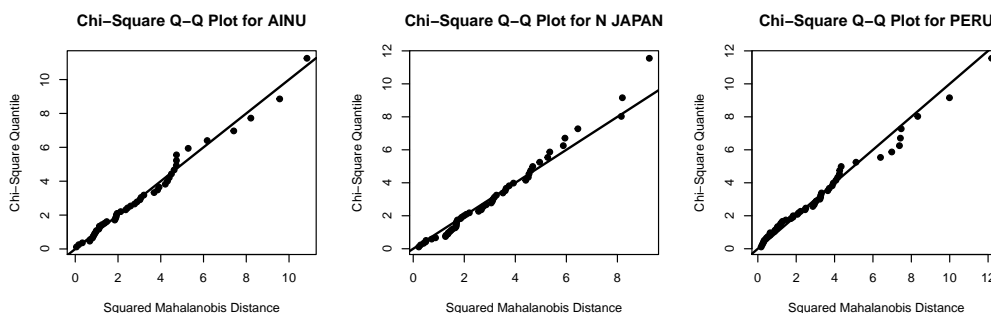
Vykreslíme si krabicové diagramy jednotlivých proměnných rozdělených podle populací.

```
par(mfrow=c(1,3))
boxplot(data$XFB ~ data$Population, ylab='frontal breadth (mm)')
boxplot(data$ZYB ~ data$Population, ylab='bizygomatic breadth (mm)')
boxplot(data$ZMB ~ data$Population, ylab='bimaxillary breadth (mm)')
```



Je potřeba ověřit předpoklad, že data pro jednotlivé populace pocházejí z třírozměrného normálního rozdělení.

```
library(MVN)
par(mfrow=c(1,3))
mvn(data, subset='Population', mvnTest = 'mardia', multivariatePlot = 'qq')$multivariateNormality
```



```
## $AINU
##           Test           Statistic           p value Result
## 1 Mardia Skewness    3.1150193672714 0.978586110255638   YES
## 2 Mardia Kurtosis   -0.170360771983077 0.864726420549113   YES
## 3 MVN                <NA>                <NA>                YES
##
## $`N JAPAN`
##           Test           Statistic           p value Result
## 1 Mardia Skewness    6.82574420606157 0.741786998088034   YES
## 2 Mardia Kurtosis   -1.2260185014743 0.220191712732232   YES
```

```
## 3          MVN          <NA>          <NA>      YES
##
## $PERU
##          Test          Statistic          p value Result
## 1 Mardia Skewness 10.7454781697583 0.377688433405176    YES
## 2 Mardia Kurtosis 0.609308387020002 0.54232004892899    YES
## 3          MVN          <NA>          <NA>      YES

mvn(data, subset='Population', mvnTest = 'hz')$multivariateNormality

## $AINU
##          Test          HZ    p value MVN
## 1 Henze-Zirkler 0.6376286 0.5145807 YES
##
## $`N JAPAN`
##          Test          HZ    p value MVN
## 1 Henze-Zirkler 0.5761641 0.7244284 YES
##
## $PERU
##          Test          HZ    p value MVN
## 1 Henze-Zirkler 0.8932563 0.07586977 YES
```

Populace Ainu:

Mardiův test pro šikmost:

Hodnota testovací statistiky

 p -hodnota

Mardiův test pro špičatost:

Hodnota testovací statistiky

 p -hodnota

Závěr

Henzeův-Zirklerův test:

Hodnota testovací statistiky

 p -hodnota

Závěr

Populace severojaponská:

Mardiův test pro šikmost:

Hodnota testovací statistiky

 p -hodnota

Mardiův test pro špičatost:

Hodnota testovací statistiky

 p -hodnota

Závěr

Henzeův-Zirklerův test:

Hodnota testovací statistiky

 p -hodnota

Závěr

Populace peruánská:

Mardiův test pro šikmost:

Hodnota testovací statistiky

 p -hodnota

Mardiův test pro špičatost:

Hodnota testovací statistiky

p -hodnota

Závěr

Henzeův-Zirklerův test:

Hodnota testovací statistiky

p -hodnota

Závěr

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

```
library(biotools)
boxM(data[,-1], grouping=data$Population)

##
## Box's M-test for Homogeneity of Covariance Matrices
##
## data: data[, -1]
## Chi-Sq (approx.) = 8.883, df = 12, p-value = 0.7129
```

Hodnota testovací statistiky

p -hodnota

Závěr

Sestavíme model mnohorozměrné analýzy rozptylu a podíváme se na hodnoty testovacích statistik.

```
model <- manova(as.matrix(data[,-1]) ~ data$Population)
summary(model, test='Wilks')

##              Df  Wilks approx F num Df den Df   Pr(>F)
## data$Population  2 0.78552   6.5429     6   306 1.629e-06 ***
## Residuals      155
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(model, test='Pillai')

##              Df  Pillai approx F num Df den Df   Pr(>F)
## data$Population  2 0.21667   6.2368     6   308 3.381e-06 ***
## Residuals      155
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(model, test='Hotelling-Lawley')

##              Df Hotelling-Lawley approx F num Df den Df   Pr(>F)
## data$Population  2          0.27025   6.8464     6   304 7.902e-07 ***
## Residuals      155
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(model, test='Roy')
```

```
##          Df      Roy approx F num Df den Df      Pr(>F)
## data$Population  2 0.25951  13.322      3   154 8.902e-08 ***
## Residuals      155
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Wilksova testovací statistika
 p -hodnota
 Závěr

Pillaiova testovací statistika
 p -hodnota
 Závěr

Hotellingova-Lawleyho testovací statistika
 p -hodnota
 Závěr

Royova testovací statistika
 p -hodnota
 Závěr

Protože jsme hypotézu, že vektory středních hodnot jsou u všech populací stejné, pokusíme se zjistit, které proměnné způsobují rozdíly. K tomu využijeme simultánní test založený na Wilksově statistice.

```
SSB <- summary(model, test='Wilks')$SS[[1]]
SSE <- summary(model, test='Wilks')$SS[[2]]
SST <- SSB + SSE

n <- nrow(data)
k <- 3 # pocet promennych
r <- 3 # pocet skupin
const <- (n - (k+r)/2 - 1)
K <- -const * log(diag(SSE)/diag(SST))
K

##          XFB          ZYB          ZMB
## 24.496114 19.807912  5.739724

kvantil <- qchisq(0.95, df=k*(r-1))
kvantil

## [1] 12.59159
```

Rozdíly mezi skupinami způsobují proměnné

Dále provedeme mnohorozměrnou obdobu mnohonásobného porovnávání. Provedeme dvouvýběrové Hotellingovy testy pro dvojice populací, ale je potřeba upravit hladinu významnosti $\frac{\alpha}{(\text{počet populací})}$.

```

alpha.korig <- 0.05/ choose(r,2)
alpha.korig

## [1] 0.01666667

library(ICSNP)
HotellingsT2(data[data$Population=='AINU',-1], data[data$Population=='N JAPAN',-1])

##
## Hotelling's two sample T2-test
##
## data: data[data$Population == "AINU", -1] and data[data$Population == "N JAPAN", -1]
## T.2 = 6.8092, df1 = 3, df2 = 99, p-value = 0.0003209
## alternative hypothesis: true location difference is not equal to c(0,0,0)

HotellingsT2(data[data$Population=='AINU',-1], data[data$Population=='PERU',-1])

##
## Hotelling's two sample T2-test
##
## data: data[data$Population == "AINU", -1] and data[data$Population == "PERU", -1]
## T.2 = 12.393, df1 = 3, df2 = 99, p-value = 5.992e-07
## alternative hypothesis: true location difference is not equal to c(0,0,0)

HotellingsT2(data[data$Population=='N JAPAN',-1], data[data$Population=='PERU',-1])

##
## Hotelling's two sample T2-test
##
## data: data[data$Population == "N JAPAN", -1] and data[data$Population == "PERU", -1]
## T.2 = 1.3815, df1 = 3, df2 = 106, p-value = 0.2525
## alternative hypothesis: true location difference is not equal to c(0,0,0)

```

Korigovaná hladina významnosti:

Srovnání Ainu a severojaponské populace

p-hodnota

Závěr

Srovnání Ainu a peruánské populace

p-hodnota

Závěr

Srovnání severojaponské a peruánské populace

p-hodnota

Závěr

Chceme zjistit, které proměnné způsobují rozdíly mezi jednotlivými dvojicemi populací. K tomu využijeme dvouvýběrový Studentův *t*-test, ale opět je třeba korigovat hladinu významnosti počtem prováděných testů!

```
alpha.korig2 <- 0.05 / (k*r*(r-1)/2)
alpha.korig2

## [1] 0.005555556

# AINU vs N JAPAN
t.test(data[data$Population=='AINU',2], data[data$Population=='N JAPAN',2])

##
## Welch Two Sample t-test
##
## data: data[data$Population == "AINU", 2] and data[data$Population == "N JAPAN", 2]
## t = 3.4842, df = 100.42, p-value = 0.0007329
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.295919 4.723020
## sample estimates:
## mean of x mean of y
## 119.6458 116.6364

t.test(data[data$Population=='AINU',3], data[data$Population=='N JAPAN',3])

##
## Welch Two Sample t-test
##
## data: data[data$Population == "AINU", 3] and data[data$Population == "N JAPAN", 3]
## t = 3.3618, df = 93.264, p-value = 0.001124
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.314029 5.106425
## sample estimates:
## mean of x mean of y
## 138.9375 135.7273

t.test(data[data$Population=='AINU',4], data[data$Population=='N JAPAN',4])

##
## Welch Two Sample t-test
##
## data: data[data$Population == "AINU", 4] and data[data$Population == "N JAPAN", 4]
## t = 1.0188, df = 100.64, p-value = 0.3107
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.9501059 2.9561665
## sample estimates:
## mean of x mean of y
## 99.16667 98.16364

# AINU vs PERU
t.test(data[data$Population=='AINU',2], data[data$Population=='PERU',2])

##
## Welch Two Sample t-test
##
## data: data[data$Population == "AINU", 2] and data[data$Population == "PERU", 2]
```

```
## t = 5.1862, df = 99.83, p-value = 1.126e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 2.722610 6.096329
## sample estimates:
## mean of x mean of y
## 119.6458 115.2364

t.test(data[data$Population=='AINU',3], data[data$Population=='PERU',3])

##
## Welch Two Sample t-test
##
## data: data[data$Population == "AINU", 3] and data[data$Population == "PERU", 3]
## t = 4.2543, df = 91.617, p-value = 5.054e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 2.137969 5.882485
## sample estimates:
## mean of x mean of y
## 138.9375 134.9273

t.test(data[data$Population=='AINU',4], data[data$Population=='PERU',4])

##
## Welch Two Sample t-test
##
## data: data[data$Population == "AINU", 4] and data[data$Population == "PERU", 4]
## t = 2.5284, df = 93.954, p-value = 0.01313
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 0.4847012 4.0304503
## sample estimates:
## mean of x mean of y
## 99.16667 96.90909

# N JAPAN vs PERU
t.test(data[data$Population=='N JAPAN',2], data[data$Population=='PERU',2])

##
## Welch Two Sample t-test
##
## data: data[data$Population == "N JAPAN", 2] and data[data$Population == "PERU", 2]
## t = 1.6521, df = 107.89, p-value = 0.1014
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.2796888 3.0796888
## sample estimates:
## mean of x mean of y
## 116.6364 115.2364

t.test(data[data$Population=='N JAPAN',3], data[data$Population=='PERU',3])

##
## Welch Two Sample t-test
```



```
##
## data: data[data$Population == "N JAPAN", 3] and data[data$Population == "PERU", 3]
## t = 0.96325, df = 107.88, p-value = 0.3376
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.8462643  2.4462643
## sample estimates:
## mean of x mean of y
## 135.7273 134.9273

t.test(data[data$Population=="N JAPAN",4], data[data$Population=="PERU",4])

##
## Welch Two Sample t-test
##
## data: data[data$Population == "N JAPAN", 4] and data[data$Population == "PERU", 4]
## t = 1.3967, df = 103.3, p-value = 0.1655
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.5267553  3.0358462
## sample estimates:
## mean of x mean of y
## 98.16364 96.90909
```

Korigovaná hladina významnosti:

Srovnání Ainu a severojaponské populace

Proměnná XFB: p -hodnota

Závěr

Proměnná ZYB: p -hodnota

Závěr

Proměnná ZMB: p -hodnota

Závěr

Srovnání Ainu a peruánské populace

Proměnná XFB: p -hodnota

Závěr

Proměnná ZYB: p -hodnota

Závěr

Proměnná ZMB: p -hodnota

Závěr

Srovnání severojaponské a peruánské populace

Proměnná XFB: p -hodnota

Závěr

Proměnná ZYB: p -hodnota

Závěr

Proměnná ZMB: p -hodnota

Závěr