MANOVA - neřešený příklad - výsledky

Zdeňka Geršlová

2024-04-29

V souboru Howell.csv máme k dispozici kraniometrické údaje z různých populací. Nás zajímají muži ze 4 populací (prom. Population) - ZALAVAR, BERG, ESKIMO a NORSE, a tyto proměnné (vše je v milimetrech): • BPL - délka obličejové části lebky, • NPH - výška horní části obličejového skeletu, • OBH - výška očnice levé strany.

Načtení dat: Použijeme read.csv s argumentem na.strings, který říká, jaké hodnoty v datech mají být brány jako NA - nastavíme na "0".

```
##    Population      BPL             NPH             OBH       
##  BERG   :56   Min.   : 81.00   Min.   :56.00   Min.   :27.00  
##  ESKIMO :53   1st Qu.: 93.00   1st Qu.:67.00   1st Qu.:33.00  
##  NORSE  :55   Median : 98.00   Median :69.00   Median :34.00  
##  ZALAVAR:53   Mean   : 97.65   Mean   :69.23   Mean   :34.09  
##               3rd Qu.:102.00   3rd Qu.:72.00   3rd Qu.:36.00  
##               Max.   :114.00   Max.   :80.00   Max.   :40.00
```

1. Pro každou populaci vypočítejte počet pozorování, vektor výběrových průměrů a výběrovou varianční matici.

```
##
##    BERG  ESKIMO   NORSE ZALAVAR 
##      56      53      55      53
```

```
##      BPL      NPH      OBH 
## 97.05660 68.45283 32.67925
```

```
##      BPL      NPH      OBH 
## 93.75000 67.89286 33.75000
```

```
##       BPL       NPH       OBH 
## 103.05660  71.73585  36.22642
```

```
##      BPL      NPH      OBH 
## 96.96364 68.92727 33.74545
```

```
##           BPL       NPH      OBH
## BPL 24.362119  3.916183 1.326197
## NPH  3.916183 17.867925 5.628810
## OBH  1.326197  5.628810 3.952830
```

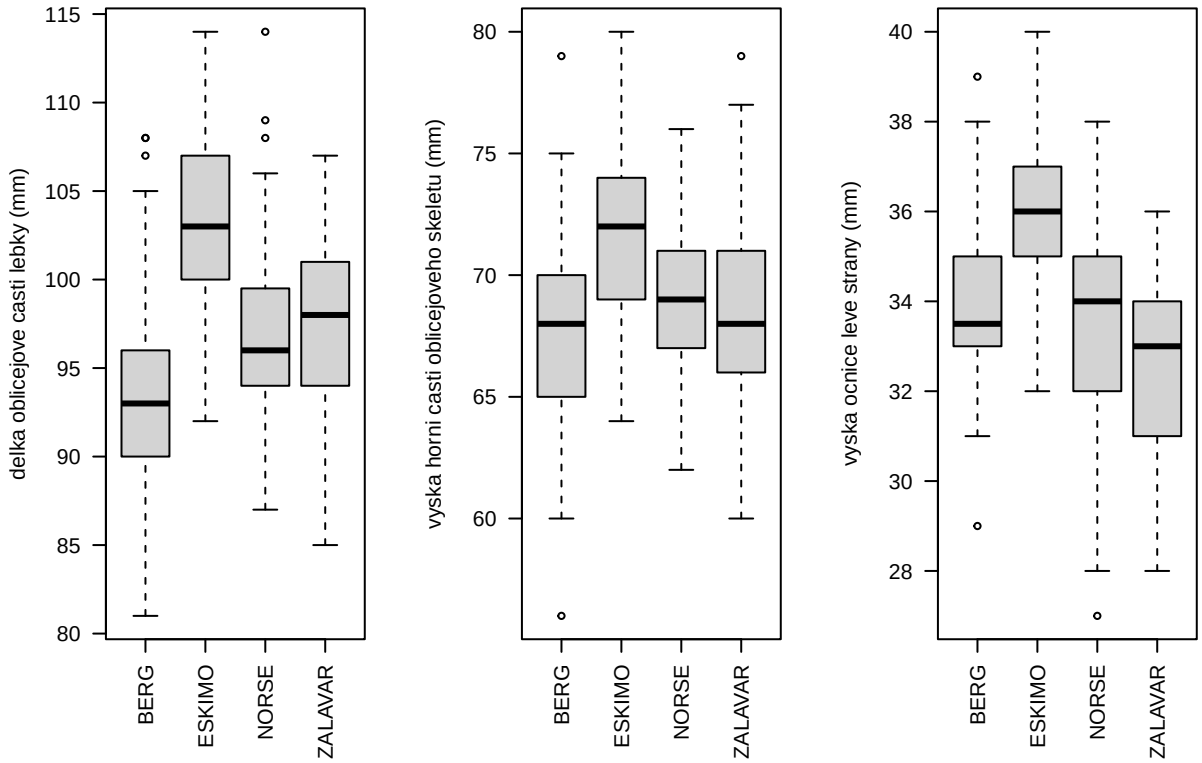
```
##             BPL         NPH       OBH
## BPL 32.44545455 -0.04545455 -2.081818
## NPH -0.04545455 17.37012987  3.300000
## OBH -2.08181818  3.30000000  3.354545
```

```
##            BPL        NPH        OBH
## BPL 21.0159652  0.8037010 -0.8207547
## NPH  0.8037010 13.3519594  0.2340348
```

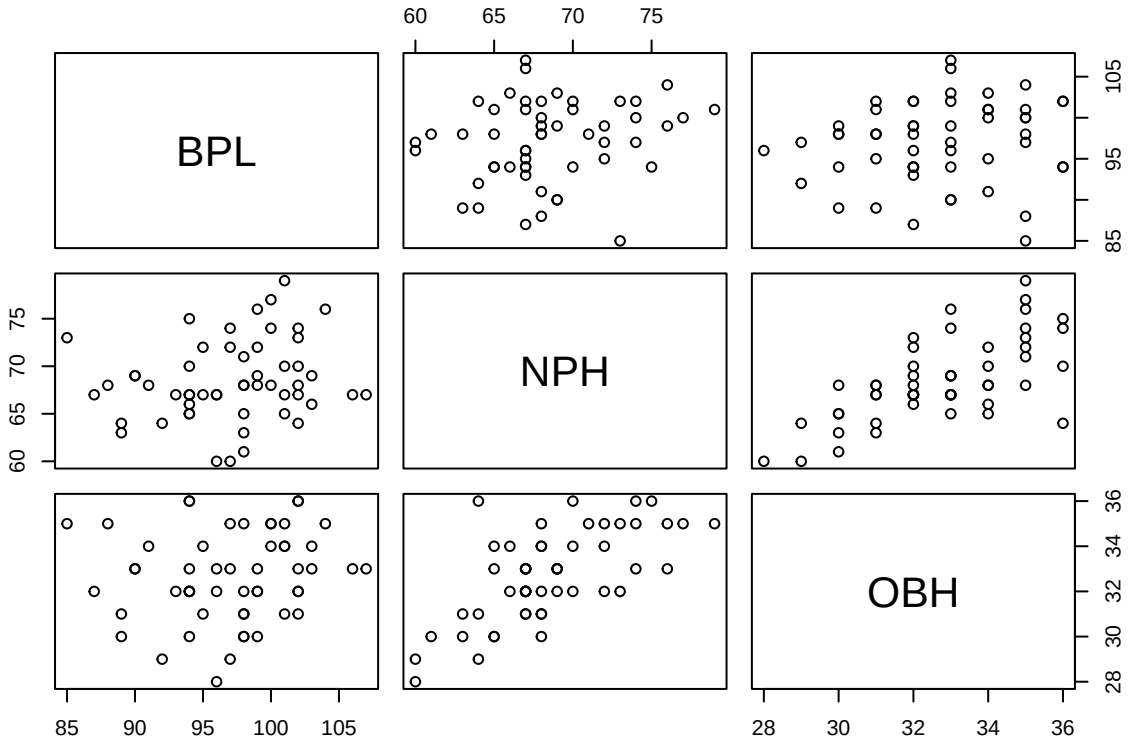
```
## OBH -0.8207547  0.2340348  2.9862119
```

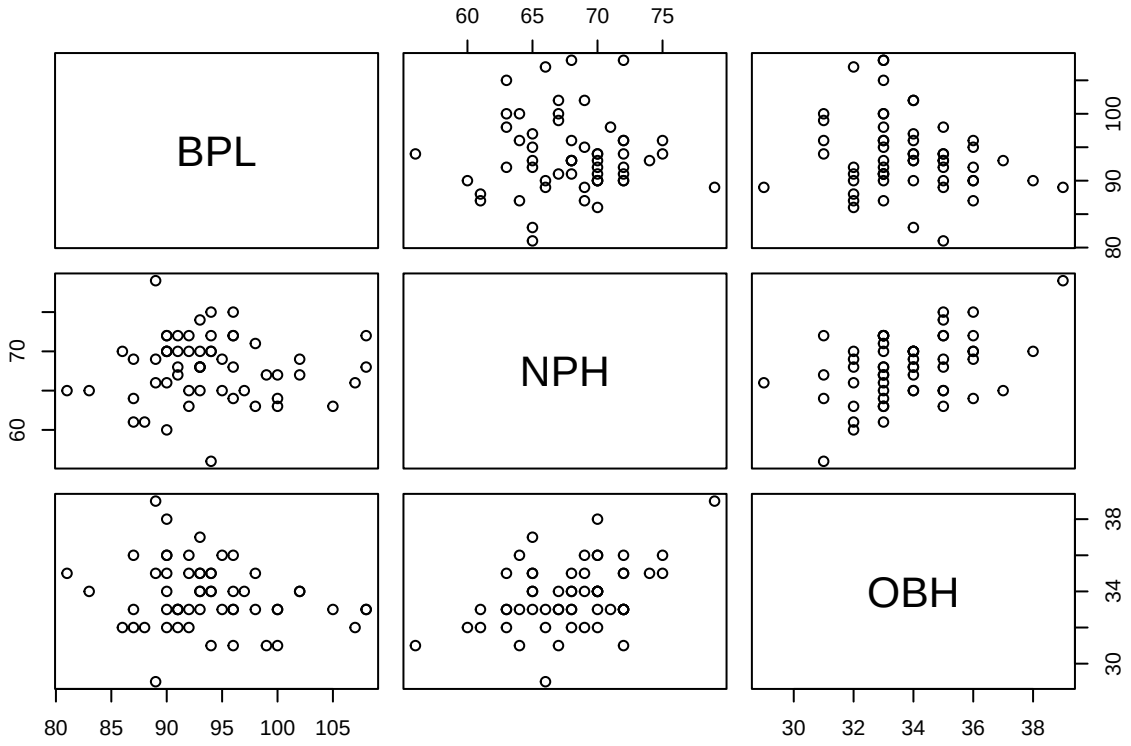
```
##            BPL        NPH       OBH
## BPL 26.5542088 -0.5952862 -3.435354
## NPH -0.5952862 11.3279461  2.832997
## OBH -3.4353535  2.8329966  4.896970
```

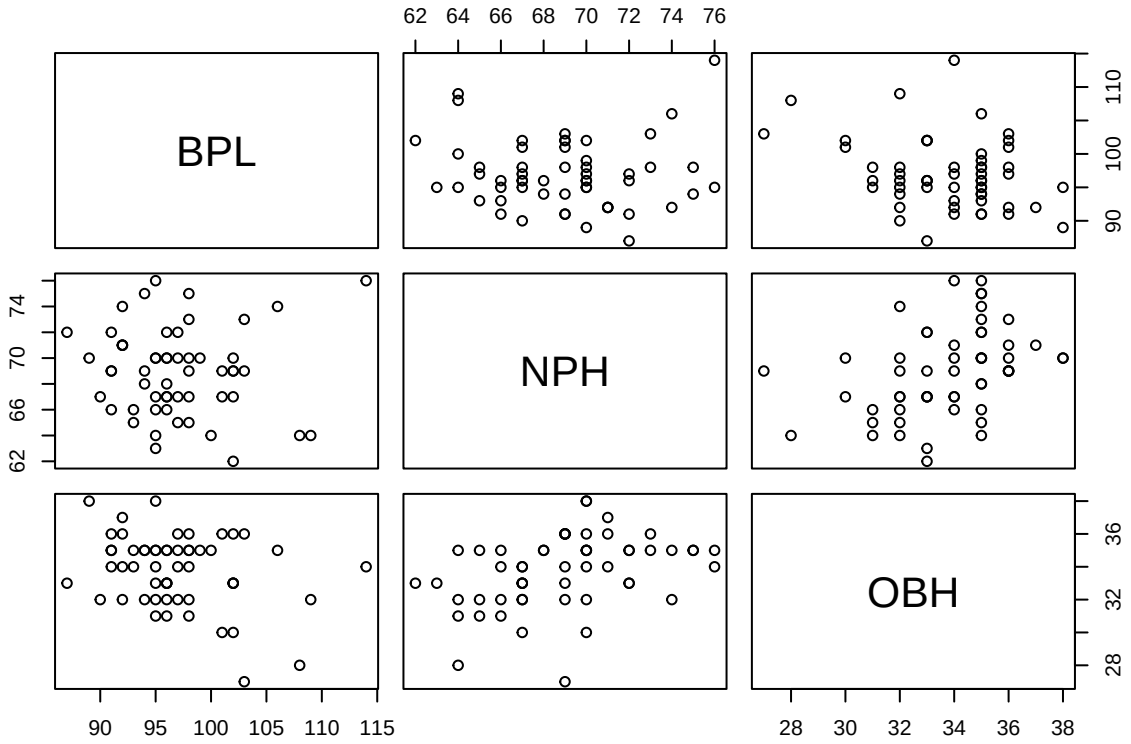
2. Vykreslete krabicové diagramy pro jednotlivé proměnné podle populací.
   ! nastavení popisku osy x svisle - argument “las”

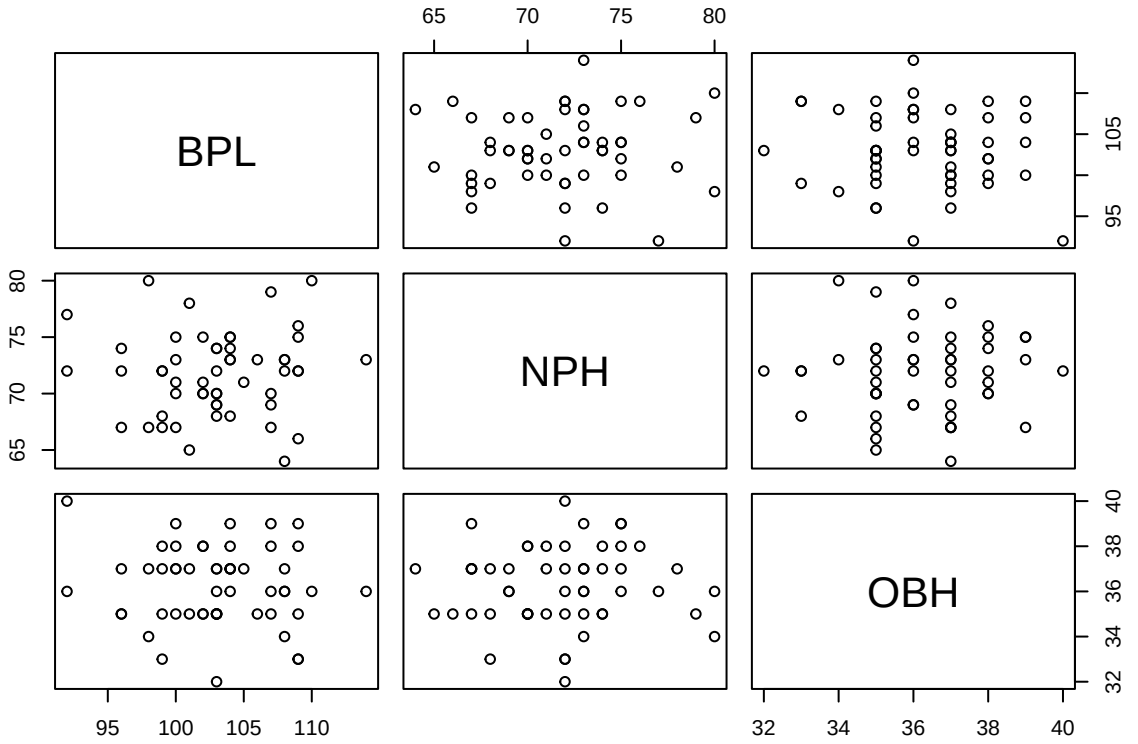


Orientační ověření linearity vztahů mezi proměnnými - vykreslíme bodové grafy proměnných pro každou populaci

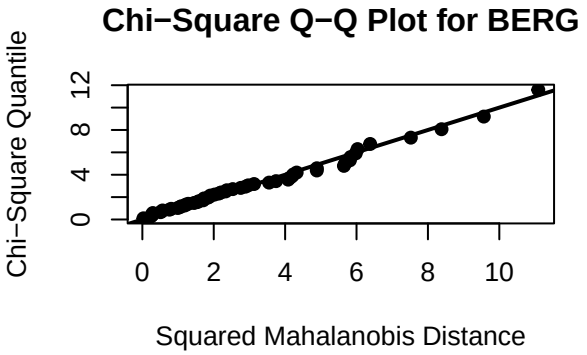

BPL
NPH
OBH
60
65
70
75
85
90
95
100
105
28
30
32
34
36

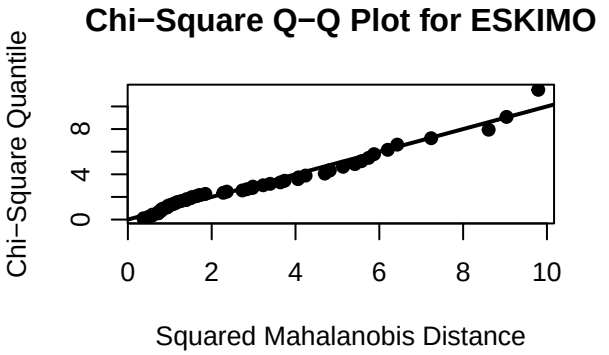

BPL
NPH
OBH
60
65
70
75
80
90
100
60
70
80
85
90
95
100
105
30
32
34
36
38
30
34
38

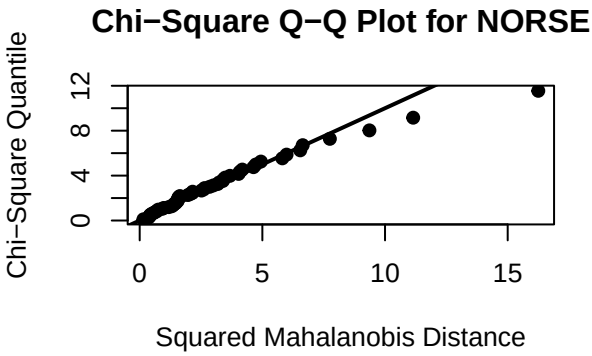

BPL
NPH
OBH
62 64 66 68 70 72 74 76
90 100 110
62 66 70 74
28 32 36
90 95 100 105 110 115
28 30 32 34 36 38

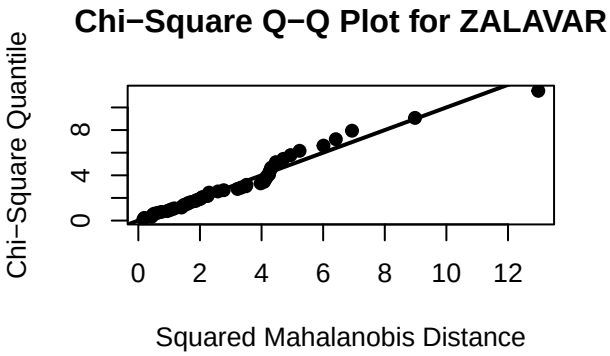


3. Ověřte pro každou populaci, že data pocházejí z trojrozměrného normálního rozdělení.










```
## $BERG
##            Test        HZ   p value MVN
## 1 Henze-Zirkler 0.7223814 0.3299512 YES
##
## $ESKIMO
##            Test        HZ   p value MVN
## 1 Henze-Zirkler 0.4939625 0.8994991 YES
##
## $NORSE
##            Test        HZ   p value MVN
## 1 Henze-Zirkler 0.8087569 0.1651607 YES
##
## $ZALAVAR
##            Test        HZ   p value MVN
## 1 Henze-Zirkler 0.6747828 0.4369919 YES
```

Nezamítli jsme vícerozměrnou normalitu.
4. Ověřte další předpoklad pro MANOVU - shodnost variančních matic použijeme Boxův M-test z knihovny biotools.

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  data[, -1]
## Chi-Sq (approx.) = 28.262, df = 18, p-value = 0.05817
```

Nezamítáme hypotézu o shodnosti variančních matic.

5. Na hladině významnosti 0.05 otestujte hypotézu, že vektory středních hodnot jsou ve všech populacích stejné. Použijte Wilksův, Pillaiův, Hotellingův-Lawleyho a Royův test.
   Sestavíme model mnohorozměrné analýzy rozptylu (MANOVA), použijeme funkci manova(), do formule zadáváme jako závislou proměnnou numerické proměnné, ale musí vstupovat jako matice, proto použijeme as.matrix(). Jako nezávislá proměnná (tj. vpravo od vlnky) bude proměnná Population. Potom zadáváním argumentu "test" v summary() provedeme jednotlivé testy.

```
##                   Df   Wilks approx F num Df den Df    Pr(>F)
## data$Population    3 0.47734   20.266      9 513.67 < 2.2e-16 ***
## Residuals        213
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##                   Df  Pillai approx F num Df den Df    Pr(>F)
## data$Population    3 0.56416   16.444      9    639 < 2.2e-16 ***
## Residuals        213
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##                   Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
## data$Population    3            1.008   23.483      9    629 < 2.2e-16 ***
## Residuals        213
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##                   Df     Roy approx F num Df den Df    Pr(>F)
## data$Population    3 0.91278   64.807      3    213 < 2.2e-16 ***
## Residuals        213
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Všechny testy zamítají hypotézu o shodě stř. hodnot. 6. Pomocí simultánního testu založeného na Wilksově statistice zjistěte, které proměnné způsobují rozdíly mezi populacemi. (statistika K, přednáška podkapitola 3)

```
##           BPL       NPH      OBH
## BPL 2445.5898 1045.9534 743.8228
## NPH 1045.9534  469.9791 373.0420
## OBH  743.8228  373.0420 360.3901
```

```
##           BPL       NPH       OBH
## BPL 5578.0877  210.7885 -273.7261
## NPH  210.7885 3190.5002  639.3497
## OBH -273.7261  639.3497  809.7666
```

```
##           BPL      NPH       OBH
## BPL 8023.6774 1256.742  470.0968
## NPH 1256.7419 3660.479 1012.3917
## OBH  470.0968 1012.392 1170.1567
```

```
##      BPL      NPH      OBH
## 77.25455 29.20098 78.23122
```

```
## [1] 16.91898
```

Rozdíly způsobují všechny proměnné (statistiky spadají do kritického oboru).

7. Na celkové hladině významnosti 0.05 proveďte obdobu mnohonásobného porovnávání, tj. zjistěte, které dvojice populací se liší. U jednotlivých testů je tedy potřeba upravit hladinu významnosti!
   Využijeme Hotellingovy testy (testujeme hypotézu o shodě středních hodnot pro 2 výběry), pro úpravu hladiny významnosti musíme původní alpha vydělit kombinačním číslem "počet populací nad 2", tzv.

Bonferroniho korigovaná hl. významnosti - viz přednáška str. 14. Populace porovnáváme mezi sebou po dvojicích, při 4 populacích tak máme 6 dvojic.

```
## [1] 0.008333333
```

```
##
##  Hotelling's two sample T2-test
##
## data:  data[data$Population == "ZALAVAR", -1] and data[data$Population == "BERG", -1]
## T.2 = 7.8504, df1 = 3, df2 = 105, p-value = 8.915e-05
## alternative hypothesis: true location difference is not equal to c(0,0,0)
```

```
##
##  Hotelling's two sample T2-test
##
## data:  data[data$Population == "ZALAVAR", -1] and data[data$Population == "ESKIMO", -1]
## T.2 = 44.061, df1 = 3, df2 = 102, p-value < 2.2e-16
## alternative hypothesis: true location difference is not equal to c(0,0,0)
```

```
##
##  Hotelling's two sample T2-test
##
## data:  data[data$Population == "ZALAVAR", -1] and data[data$Population == "NORSE", -1]
## T.2 = 2.5381, df1 = 3, df2 = 104, p-value = 0.06063
## alternative hypothesis: true location difference is not equal to c(0,0,0)
```

```
##
##  Hotelling's two sample T2-test
##
## data:  data[data$Population == "BERG", -1] and data[data$Population == "ESKIMO", -1]
## T.2 = 56.69, df1 = 3, df2 = 105, p-value < 2.2e-16
## alternative hypothesis: true location difference is not equal to c(0,0,0)
```

```
##
##  Hotelling's two sample T2-test
##
## data:  data[data$Population == "BERG", -1] and data[data$Population == "NORSE", -1]
## T.2 = 3.911, df1 = 3, df2 = 107, p-value = 0.01077
## alternative hypothesis: true location difference is not equal to c(0,0,0)
```

```
##
##  Hotelling's two sample T2-test
##
## data:  data[data$Population == "ESKIMO", -1] and data[data$Population == "NORSE", -1]
## T.2 = 37.046, df1 = 3, df2 = 104, p-value = 2.22e-16
## alternative hypothesis: true location difference is not equal to c(0,0,0)
```

! nezapomeňte, že porovnáváme s korigovanou hl. významnosti !

8. Na celkové hladině významnosti 0.05 zjistěte, které proměnné způsobují rozdíly mezi jednotlivými dvojicemi. U jednotlivých testů je tedy potřeba upravit hladinu významnosti!
   Využijeme dvouvýběrový (Studentův) t-test, v němž je ovšem nutné korigovat hladinu významnosti počtem prováděných testů, v našem případě jich je 18 (mám 6 dvojic populací a pro každou 3 proměnné).

```
## [1] 0.002777778
```

```
##
##  F test to compare two variances
##
```

```
## data:  data[data$Population == "ZALAVAR", 2] and data[data$Population == "BERG", 2]
## F = 0.75086, num df = 52, denom df = 55, p-value = 0.2998
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4376457 1.2936191
## sample estimates:
## ratio of variances
##          0.7508639
```

```
##
##  F test to compare two variances
##
## data:  data[data$Population == "ZALAVAR", 3] and data[data$Population == "BERG", 3]
## F = 1.0287, num df = 52, denom df = 55, p-value = 0.9161
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5995598 1.7722144
## sample estimates:
## ratio of variances
##           1.028658
```

```
##
##  F test to compare two variances
##
## data:  data[data$Population == "ZALAVAR", 4] and data[data$Population == "BERG", 4]
## F = 1.1784, num df = 52, denom df = 55, p-value = 0.5486
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.686809 2.030111
## sample estimates:
## ratio of variances
##            1.17835
```

Pro ostatní analogicky. Žádný test rovnost rozptylů nezamítl, v t-testech tedy využijeme variantu var.equal=T. (v nich už používáme alfa korigované = 0.002777778)

```
##
##  Two Sample t-test
##
## data:  data[data$Population == "ZALAVAR", 2] and data[data$Population == "BERG", 2]
## t = 3.2311, df = 107, p-value = 0.001639
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.277886 5.335322
## sample estimates:
## mean of x mean of y
##   97.0566   93.7500
```

```
##
##  Two Sample t-test
##
## data:  data[data$Population == "ZALAVAR", 3] and data[data$Population == "BERG", 3]
## t = 0.69628, df = 107, p-value = 0.4878
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.034341  2.154287
```

```
## sample estimates:
## mean of x mean of y
##  68.45283  67.89286
```

```
##
##  Two Sample t-test
##
## data:  data[data$Population == "ZALAVAR", 4] and data[data$Population == "BERG", 4]
## t = -2.9265, df = 107, p-value = 0.004188
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.7960847 -0.3454247
## sample estimates:
## mean of x mean of y
##  32.67925  33.75000
```

Analogicky pro všechny další dvojice a proměnné.
Výsledky: Populace ZALAVAR a BERG se liší v proměnné BPL, v ostatních jsme nezamítli shodnost. Populace ZALAVAR vs. ESKIMO, BERG vs. ESKIMO, NORSE vs. ESKIMO se liší ve všech proměnných. ZALAVAR vs. NORSE se neliší (to nám vyšlo už předtím při porovnávání celých populací). BERG vs. NORSE se liší v proměnné BLP.

Antropologický závěr: Na základě mnohorozměrné analýzy rozptylu tří rozměrů lebky u čtyř populací, z nichž tři jsou záměrně zvoleny evropského původu a jedna výrazně odlišná – Eskymáci, jsme prokázali odlišnost těchto populací a pak jednotlivě zejména rozdíly mezi populací eskymáckou a všemi evropskými. V jednom rozměru nacházíme rozdíl i mezi některými evropskými populacemi, ale celkově není zdaleka tak výrazný. Tento příklad demonstruje možnosti kraniometrického odlišení lidských skupin, kdy populačně vzdálené skupiny jsou jednoznačně odlišeny, zatímco u blízkých populací (při těchto počtech případů) nemusíme shodu statisticky zamítnout.