Statistical inference - R exercises

Ruth Gómez

October 17, 2017

Data exploration

1.Practice data exploration with the class collaborative data set CLASS_TRAITS(a larger dataset with data of previous students is in CLASS_TRAITS large).

```
# The two datasets are downloaded and merged
class_1718 <- read.table("http://bioinformatica.uab.cat/base/documents/bioinformaticsintranet1718/datab
class_old <- read.table("http://bioinformatica.uab.cat/base/documents/bioinformaticsintranet1617/databl
classTraits <- merge(class_1718, class_old, all = T)
rm(class_1718, class_old)

# Variable identification
str(classTraits)
```

```
## 'data.frame':    89 obs. of  10 variables:
##  $ name          : Factor w/ 83 levels "Alejandra","Alex",..: 1 2 3 4 5 6 6 7 8 9 ...
##  $ gender        : Factor w/ 2 levels "F","M": 1 2 2 2 2 2 2 1 2 1 ...
##  $ degree        : Factor w/ 24 levels "Biochemistry",..: 9 1 3 3 8 8 9 9 10 6 ...
##  $ height        : int  159 167 183 169 180 180 180 170 176 177 ...
##  $ weight        : int  49 70 74 63 74 78 74 62 65 55 ...
##  $ ABO           : Factor w/ 5 levels "O","A","AB","B",..: 1 2 3 NA 1 1 1 2 2 3 ...
##  $ Rh            : Factor w/ 2 levels "rh-","rh+": 1 1 2 NA 2 1 1 2 1 2 ...
##  $ hair_color    : Factor w/ 6 levels "black","Black",..: 5 5 3 5 1 1 5 5 4 5 ...
##  $ eye_color     : Factor w/ 8 levels "blue","brown",..: 2 2 1 4 2 4 2 2 4 2 ...
##  $ tongue_rolling: Factor w/ 4 levels "no","No","yes",..: 1 3 3 3 3 3 1 1 3 3 ...
```

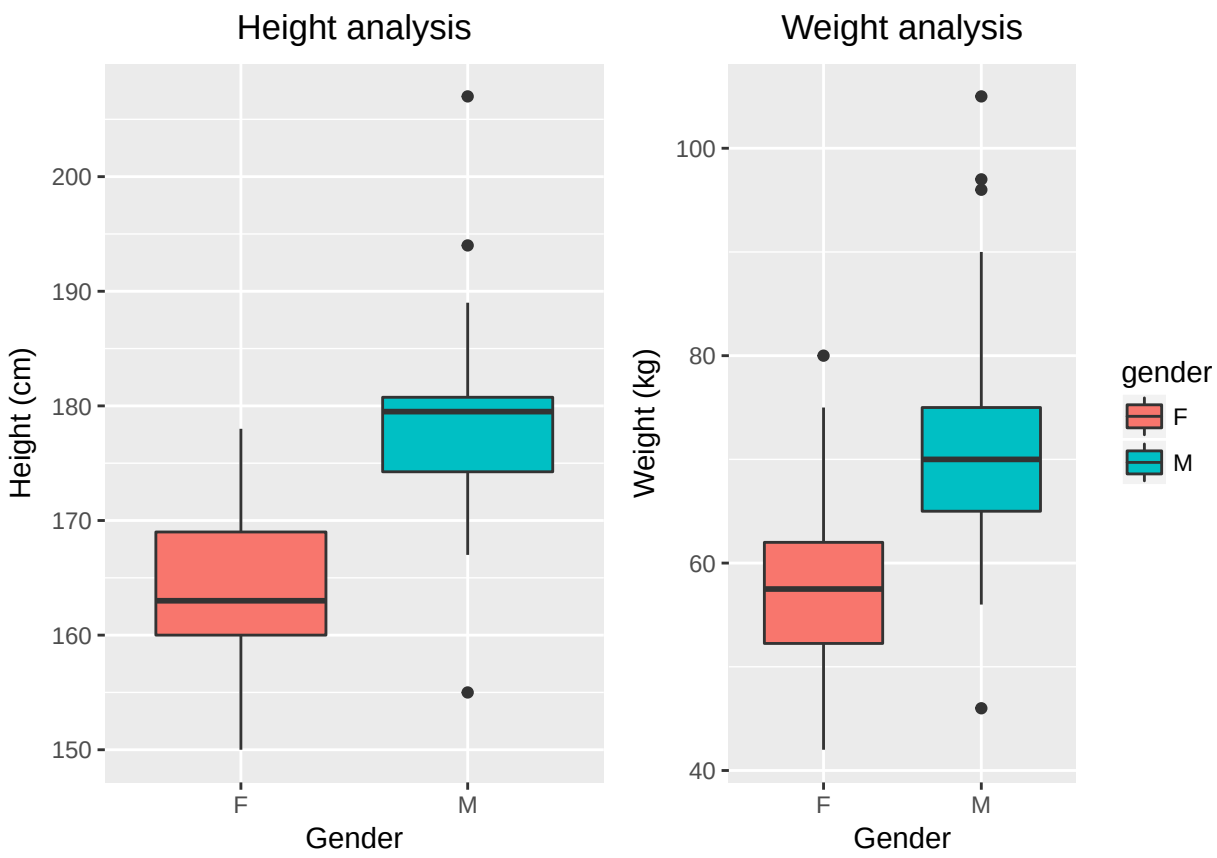
```
# Data cleaning
# Degree
levels(classTraits$degree)[grep('^[b,B]iochem*', levels(classTraits$degree) )] <- 'Biochemistry'
levels(classTraits$degree)[grep('^[b,B]iolo*', levels(classTraits$degree) )] <- 'Biology'
levels(classTraits$degree)[grep('^[b,B]iotech*', levels(classTraits$degree) )] <- 'Biotechnology' # Gro
levels(classTraits$degree)[grep('^[b,B]iomed*', levels(classTraits$degree) )] <- 'Biomedical_Sciences'
levels(classTraits$degree)[grep('^[m,M]icro*', levels(classTraits$degree) )] <- 'Microbiology'
levels(classTraits$degree)[grep('Physics*', levels(classTraits$degree) )] <- 'Biophysics'
levels(classTraits$degree)[c(grep('[c,C]omputer', levels(classTraits$degree)),grep('Informatician', lev
'Computer_Sciences'  # Groups computer science, computer vision, informatician
levels(classTraits$degree)[grep('mol', levels(classTraits$degree) )] <- 'Molecular_Biology'

# ABO
  levels(classTraits$ABO)[grep('[O,o,O]', levels(classTraits$ABO) )] <- 'O'

# Hair color
  levels(classTraits$hair_color)[grep('[B,b]lack', levels(classTraits$hair_color) )] <- 'black'
 levels(classTraits$hair_color)[grep('[B,b]lond', levels(classTraits$hair_color) )] <- 'blonde'
  levels(classTraits$hair_color)[grep('redhead', levels(classTraits$hair_color) )] <- 'red'
  levels(classTraits$hair_color)[grep('brown', levels(classTraits$hair_color) )] <- 'brown'

# Eye color
```

```
   levels(classTraits$eye_color)[c(grep('[b,B]rown', levels(classTraits$eye_color) ),grep('[b,B]lack', l
# There  is one individual that has marked their eyes as 'blonde', so it will be assumed that there was
  classTraits[which(classTraits$eye_color == 'blonde'),c('hair_color', 'eye_color')]<-
    classTraits[which(classTraits$eye_color == 'blonde'),c('eye_color', 'hair_color')]

classTraits$eye_color<-factor(classTraits$eye_color)

# Tongue_rolling
  levels(classTraits$tongue_rolling)[c(grep('[n,N]', levels(classTraits$tongue_rolling) ),grep('[b,B]la
  levels(classTraits$tongue_rolling)[c(grep('[y, Y]', levels(classTraits$tongue_rolling) ),grep('[b,B]l

# Univariate analysis in continuous variables (height and weight)

sapply(classTraits[,c("height", "weight")], summary)
```

```
## $height
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   150.0   168.0   175.0   173.3   180.0   207.0       1
##
## $weight
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   42.00   60.00   67.00   67.18   74.00  105.00
```

```
library(ggplot2)
library(gridExtra)
p1<-ggplot(classTraits)+geom_boxplot(aes(x = gender , y = height , fill = gender))+labs(title  ="Height
p2<-ggplot(classTraits)+geom_boxplot(aes(x = gender , y = weight,fill = gender))+labs(title = "Weight a
grid.arrange(p1,p2, nrow = 1)
```

```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```




```
# Distribution of categorical variables, The NAs are not counted in the proportions
sapply(classTraits[,c("gender", "degree", "ABO", "Rh", "hair_color", "eye_color", "tongue_rolling")], FU
```

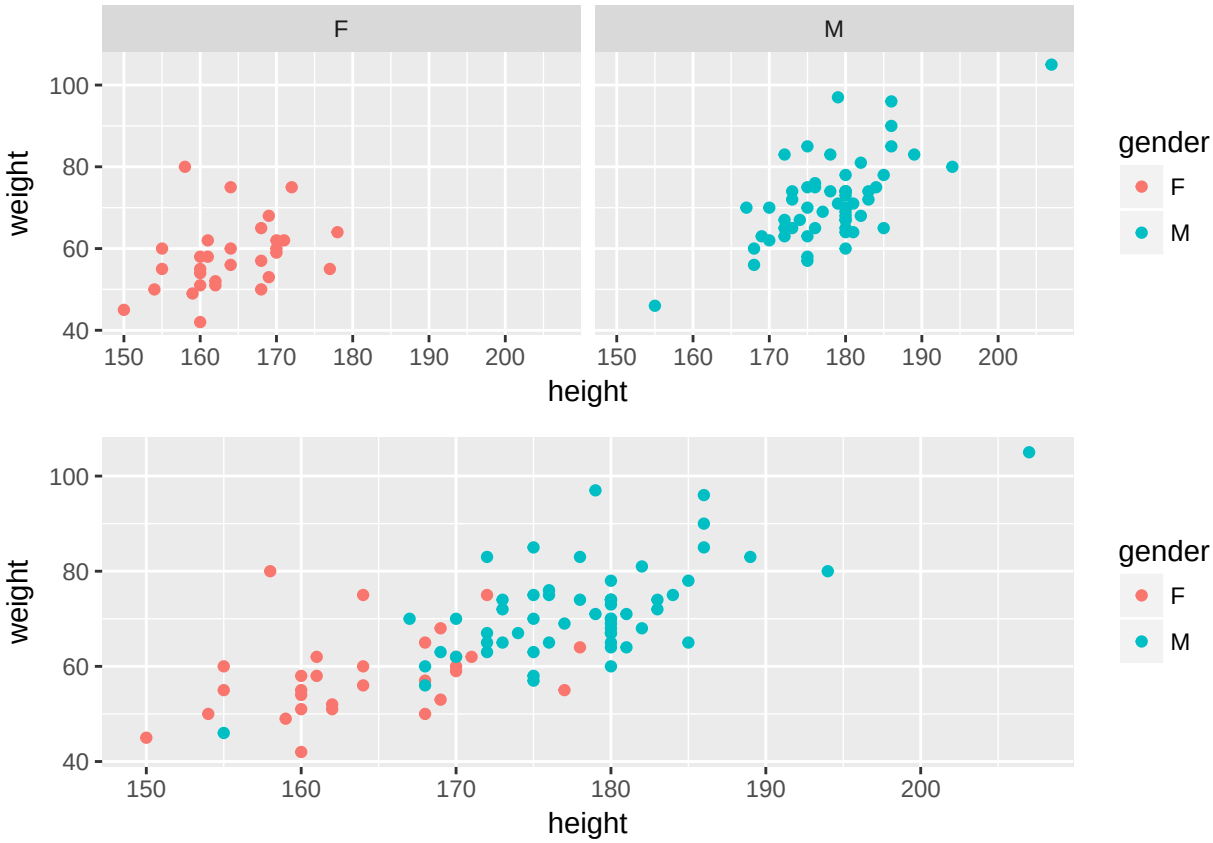
```
## $gender
## x
##         F         M
## 0.3370787 0.6629213
##
## $degree
## x
##        Biochemistry             Biology Biomedical_Sciences
##           0.20224719          0.13483146          0.07865169
##       Biotechnology            Genetics         Microbiology
##           0.23595506          0.14606742          0.06741573
##      Nanotechnology            Agronomy           Biophysics
##           0.01123596          0.01123596          0.01123596
##                 BML    Computer_Sciences               Maths
##           0.01123596          0.04494382          0.02247191
##   Molecular_Biology
##           0.02247191
##
## $ABO
## x
##         O         A        AB         B
## 0.3200000 0.4533333 0.0800000 0.1466667
##
## $Rh
```

```
## x
##      rh-      rh+
## 0.369863 0.630137
##
## $hair_color
## x
##      black     blonde      brown        red
## 0.22471910 0.11235955 0.65168539 0.01123596
##
## $eye_color
## x
##      blue     brown     green
## 0.1235955 0.7191011 0.1573034
##
## $tongue_rolling
## x
##        no       yes
## 0.3483146 0.6516854
```

```
# Bi variate analysis on continuous variables, with both height and weight recorded
p<-ggplot(classTraits, aes(height, weight, color = gender)) + geom_point()
p3 <-p+ facet_grid(.~gender)
grid.arrange(p3, p, ncol = 1)
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



```r
var(classTraits[,c("height", "weight")], use= 'complete.obs')
```

```
##          height   weight
## height 95.77377  83.6395
## weight 83.63950 137.3939
```

```r
cor(classTraits[,c("height", "weight")], use= 'complete.obs')
```

```
##           height    weight
## height 1.0000000 0.7291289
## weight 0.7291289 1.0000000
```

```r
# Bi variate analysis deleting outliers

subset<-classTraits[which ((classTraits$gender == "M" & classTraits$height %in% c(160:190) & classTraits

var(subset)
```

```
##          height   weight
## height 73.90440 61.19512
## weight 61.19512 97.31165
```

```r
cor(subset)
```

```
##           height    weight
## height 1.0000000 0.7216045
## weight 0.7216045 1.0000000
```

```r
p5<-ggplot(classTraits, aes(x = eye_color))+geom_bar( aes(fill = hair_color ), position = 'stack')
p5
```