2.4 Exploratory Data Analysis – R Functions

# PART 1 – THEORETICAL CONCEPTS

Box Plot Function

cohort.data$SOFA <- cut(cohort.data$sofa\_first, breaks= c(0,5, 10, 17))  
qplot(data = cohort.data, x = SOFA, y = map\_1st, color = "purple", geom = "boxplot", ylab = "MAP (mmHg)") +  
 scale\_color\_manual(values = "purple", guide = F)

Chi-Square Function

# Independence test between expiration in ICU and IAC:

chisq.test(table(cohort.data$icu\_exp\_flg, cohort.data$aline\_flg, useNA = "ifany"))

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: table(cohort.data$icu\_exp\_flg, cohort.data$aline\_flg, useNA = "ifany")  
## X-squared = 7.8895, df = 1, p-value = 0.004972

# Independence test of day 28 with IAC:  
chisq.test(table(cohort.data$day\_28\_flg, cohort.data$aline\_flg, useNA = "ifany"))

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: table(cohort.data$day\_28\_flg, cohort.data$aline\_flg, useNA = "ifany")  
## X-squared = 2.7451, df = 1, p-value = 0.09755

Density function

#plot of the histogram with density (beware, the ordinate is the density, not the count)  
**qplot**(x = x, y = ..density.., geom = **c**("histogram", "density"), binwidth = 0.2)

Distribution function

**set.seed**(2) # set the seed  
# x is a vector of normal distribution. y a uniform one.  
x <- **rnorm**(35, 0, 1)  
y <- **runif**(35, -1.5, 1.5)  
# Test of shapiro-Wilk:  
**shapiro.test**(x)

Histogram function

We provide 2 examples for producing histograms in R. The first using the ggplot2 package with a random sample of numbers in a normal distribution and the second using the qplot function in base R with an example from the arterial line case study.

**set.seed**(1) # set the seed  
**library**(ggplot2) # load the library  
# vector of 500 randomly generated values following a normal distribution.  
x <- **rnorm**(500, 0, 1)  
#plot of the histogram : qplot is more fancy than plot.  
**qplot**(x = x, y = ..count.., geom = **c**("histogram"), binwidth = 0.2 )

cohort.data$aline\_flg <- ifelse(cohort.data$aline\_flg == 1, "With IAC", "Without IAC")  
qplot(data = cohort.data, x = sofa\_first, binwidth = 1, color = "darkgreen")+  
 scale\_color\_manual(values = "darkgreen", guide = F) +  
 facet\_grid(aline\_flg~.)

Lag Plot function

**set.seed**(10) # set the seed  
# random time series: X(t) = 0.6 \* X(t-1) + epsilon(t) + epsilon(t-1) \* 0.3  
# with epsilon following a normal distribution.  
ts.sim <- **arima.sim**(**list**(order = **c**(1,0,1), ar = 0.6, ma = 0.3), n = 100)  
**plot**(ts.sim, main = "Example of 2D line plot")

#Lag plot  
**lag.plot**(ts.sim, labels = F, main = "Lag Plot", col = "darkred"  
 , lags = 3, layout = **c**(1, 3), diag = F, do.lines = F)

Autocorrelation plot Function

# autocorrelation and partial autocorrelation plot  
**par**(mfrow = **c**(1, 2))  
**acf**(ts.sim, lag.max =7, col = "darkgreen")  
**pacf**(ts.sim, lag.max =6, col = "darkgreen")

Linear Fit Function

**set.seed**(1) *# set the seed*  
n <- 50  
x <- 1:n *# generate a sequence to represent indexes, x= 1,2,…,N.*  
y <- x + **rnorm**(n, 0, 5) *# y = x, theoretical model.*

res <- **lm**(y~x) *# Regression. Fits a model of y =a\*x + c*  
res

##   
## Call:  
## lm(formula = y ~ x)  
##   
## Coefficients:  
## (Intercept) x   
## 0.5825 0.9969

**qplot**(x = x, y = y) +  
 **geom\_abline**(intercept = res$coefficients[1], slope = res$coefficients[2]  
 , color = "blue") +  
 **geom\_text**(**aes**(label = **paste0**("Y = ", **round**(res$coefficients[1], 3), " + "  
 , **round**(res$coefficients[2], 3), "x"), x = 10, y = 46), color = "black")

Outliers function

**library**(outliers) # loading the library  
**set.seed**(1) # set the seed  
**grubbs.test**(**rnorm**(30, 0, 1), two.sided = F, type = 10)

Stem and Leaf Function

# The function: (code Matti Pastell)  
stem.func <- function(x, y){  
 plot(x, y, pch = 16)  
 for (i in 1:length(x)){  
 lines(c(x[i], x[i]), c(0, y[i]), col = 4)  
 }  
 lines(c(x[1] - 2, x[length(x)] + 2), c(0, 0), col = 2)  
}  
# An example:  
x <- seq(0, 2\*pi, by = 0.2)  
  
stem.func(x,sin(x))

# PART 2 - CASE STUDY

Tabulate Function

# IAC frequency:

table(cohort.data$aline\_flg, useNA = "ifany")

##   
## 0 1   
## 792 984

# 28 days’ state:

table(cohort.data$day\_28\_flg, cohort.data$aline\_flg, useNA = "ifany")

##   
## 0 1  
## 0 679 814  
## 1 113 170

# IAC vs expire in ICU:

table(cohort.data$icu\_exp\_flg, cohort.data$aline\_flg, useNA = "ifany")

##   
## 0 1  
## 0 734 872  
## 1 58 112

# proportion matrix between expiration in hospital and IAC:  
prop.table(table(cohort.data$hosp\_exp\_flg, cohort.data$aline\_flg))

##   
## 0 1  
## 0 0.39527027 0.46734234  
## 1 0.05067568 0.08671171

# three way ftable: ARDS \* pneumonia \* IAC  
ftable(cohort.data$ards\_flg, cohort.data$pneumonia\_flg, cohort.data$aline\_flg, useNA = "ifany")

## 0 1  
##   
## 0 0 619 779  
## 1 140 135  
## 1 0 26 53  
## 1 7 17

# with margins: ARDS \* IAC, resp \* IAC  
addmargins(table(cohort.data$ards\_flg, cohort.data$aline\_flg, useNA = "ifany"), 1)

##   
## 0 1  
## 0 759 914  
## 1 33 70  
## Sum 792 984

addmargins(table(cohort.data$resp\_flg, cohort.data$aline\_flg, useNA = "ifany"), 2)

##   
## 0 1 Sum  
## 0 514 697 1211  
## 1 278 287 565

Summary Function

summary(cohort.data[, c("hour\_icu\_intime", "icu\_los\_day", "gender")])

## hour\_icu\_intime icu\_los\_day gender   
## Min. : 0.00 Min. : 0.500 : 1   
## 1st Qu.: 3.00 1st Qu.: 1.370 F: 750   
## Median : 9.00 Median : 2.185 M:1025   
## Mean :10.59 Mean : 3.346   
## 3rd Qu.:19.00 3rd Qu.: 4.003   
## Max. :23.00 Max. :28.240

Side-by-side histograms Function

cohort.data$aline\_flg <- ifelse(cohort.data$aline\_flg == 1, "With IAC", "Without IAC")  
qplot(data = cohort.data, x = sofa\_first, binwidth = 1, color = "darkgreen")+  
 scale\_color\_manual(values = "darkgreen", guide = F) +  
 facet\_grid(aline\_flg~.)

Case-study Boxplot Function

cohort.data$SOFA <- cut(cohort.data$sofa\_first, breaks= c(0,5, 10, 17))  
qplot(data = cohort.data, x = SOFA, y = map\_1st, color = "purple", geom = "boxplot", ylab = "MAP (mmHg)") +  
 scale\_color\_manual(values = "purple", guide = F)