AMCS 210 - Applied Statistics and Data Analysis
Hernando Catequista Ombao
Applied Mathematics and Computational Sciences Program
Computer, Electrical and Mathematical Sciences & Engineering (CEMSE) Division
King Abdullah University of Science and Technology (KAUST)

HOMEWORK 2

Henrique Aparecido Laureano

Fall Semester 2017

Contents

Exercise 1

A politician running for office is interested in estimating the proportion among legal residents of the state of California who support the DREAMER's act (which we denote as π). An organization was instructed to conduct a poll with the constraint that the estimator should be within a tolerance limit of 0.05.

Suppose that the true population proportion of supporters is $\pi = 0.65$. The organization is now at a planning stage to determine the desired sample size so that there is only a slim chance of 10% that an estimator (sample proportion) falls outside of the designated tolerance limit (i.e., it falls outside of 0.60 and 0.70).

(a)

Conduct a simulation study to determine if a sample size of n = 200 respondents will be sufficient to meet the target? As a guide to approaching this problem: (i.) simulate at least B = 10.000 samples each of size n = 200 from a binomial distribution; (ii.) for each sample record the sample proportion; (iii.) determine the proportion (out B = 10.000 samples) of sample proportion that fall within the designated tolerance limit.

Solution:

```
# <code r> ============================================================================ #
## defining some variables
n <- 200 # sample size
B = 1e4 # number of replications
p <- .65 # true proportion
tol <- .05 # tolerance level
in_inter <- numeric(B) # creating empth object of size B

## simulation study code
for (i in 1:B){
  samp <- rbinom(1, n = n, p = p) # simulating samples from a binomal dist.
  prop <- mean(samp) # sample proportion
  abs_samp <- abs(prop - p) # difference to the tolerance level
  in_inter[i] <- ifelse(abs_samp <= tol, "in", "out") # if is in or out
}
# proportion of sample proportion that fall within (or out) the tol. level
prop.table(table(in_inter))
# </code r> ============================================================================ #
```

```
in_inter
    in    out
0.8622 0.1378
```

The proportion is of 14%, bigger than the desired proportion of 10%.

(b)

If the sample size of n = 200 does not satisfy the requirement above continue with procedure of finding the smallest sample size that would satisfy the tolerance limit.

Solution:

```
# <code r> ============================================================================ #
final_prop <- prop.table(table(in_inter)) ; prop_out <- final_prop[[2]]
while (final_prop[[2]] > .1){
  for (i in 1:B){
    samp <- rbinom(1, n = n + 1, p = p)
    prop <- mean(samp) ; abs_samp <- abs(prop - p)
    in_inter[i] <- ifelse(abs_samp <= tol, "in", "out")}
  final_prop <- prop.table(table(in_inter))
  prop_out[length(prop_out)+1] <- final_prop[[2]] ; n <- n + 1}
library(latticeExtra)
xyplot(prop_out ~ 200:n, type = c("h", "p"), pch = 16
       , xlab = "Sample size", ylab = "Tolerance limit"
       , main = paste0(
         "Smallest sample size: ", n, " (tolerance limit: ", min(prop_out), ")")
       , scales = list(x = list(at = c(200, n)))
       , panel = function(...){
         panel.abline(h = .1, col = 2, lty = 2)
         panel.xyplot(...)
         panel.segments(n, 0, n, min(prop_out), col = 2, lwd = 2)
         panel.points(n, min(prop_out), col = 2, pch = 16)})
# </code r> ============================================================================ #
```

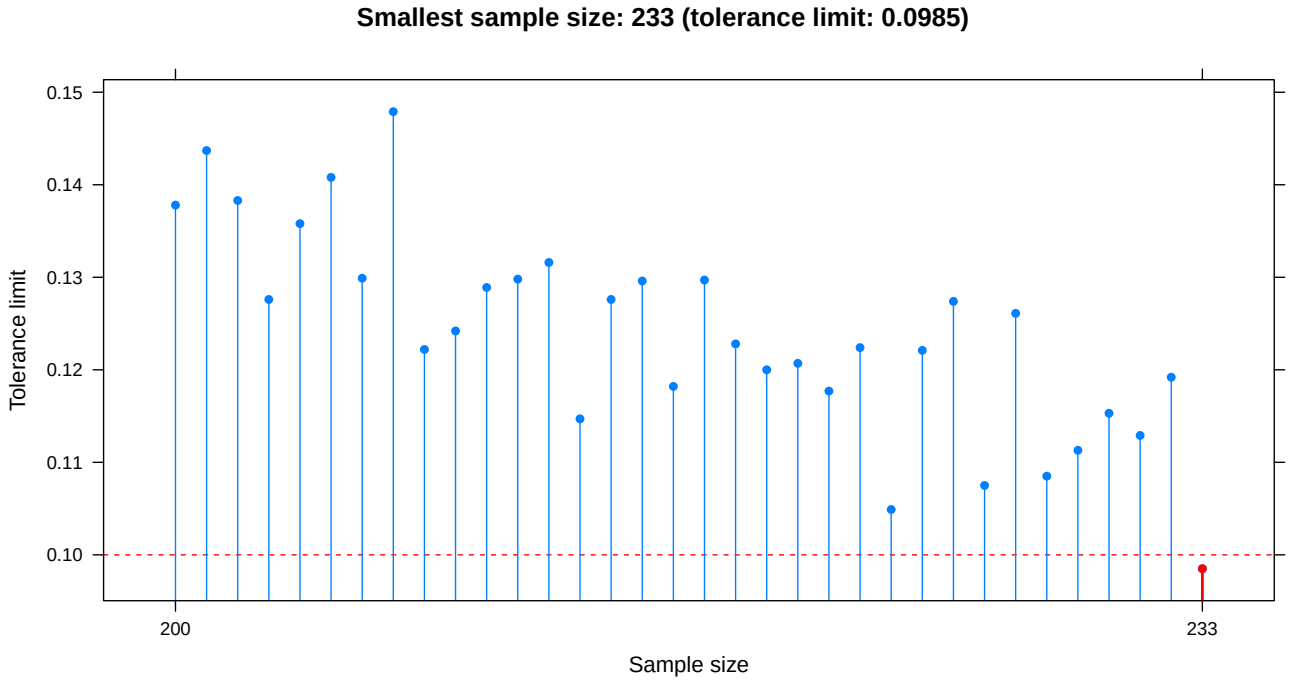

Figure 1: Sample sizes vs. tolerance limits.

Exercise 2

Exercises on calculating binomial probabilities. Let X be a binomial random variable with n = 5 and probability parameter $\pi = 0.3$. Think of this as counting the number of heads obtained by tossing n = 5 identical coins each of which has the probability of $\pi = 0.3$ that it lands in heads.

- Enumerate the elements of the sample space of X.

Solution:

- Obtain the table of probability mass function, i.e., compute the probabilities $\mathbb{P}(X = x)$ for each x in the sample space.

Solution:

- Compute $\mathbb{P}(X \geq 4)$.

Solution:

- Compute $\mathbb{P}(X < 3)$.

Solution:

```
# <code r> ========================================================================== #
(snore <- read.table("~/Dropbox/CLASS-DROPBOX/BOOK-DATA/snoreData.txt", skip = 1
                     , header = TRUE))
# </code r> ========================================================================== #
```

```
  snoring_severity heart_disease total
1                0            24  1379
2                2            35   638
3                4            21   213
4                5            30   254
```

```
# <code r> ========================================================================== #
library(latticeExtra)
print(
  barchart(total + heart_disease ~ factor(snoring_severity), snore
           , col = c("#0080ff", "gray60"), border = "transparent"
           , xlab = "Snoring severity", ylab = "Number of people"
           , scales = list(y = list(at = c(30, sort(snore$total))))
           , main = "Total numbers by level of severity", sub = "(a)"
           , key = list(corner = c(.9, .9)
                        , text = list(c("Total number", "Heart disease"))
                        , rectangle = list(border = "transparent")
                        , col = c("#0080ff", "gray60"))
```

```
           , panel = function(...){
             panel.abline(h = c(30, sort(snore$total)), col = "gray50", lty = 2)
             panel.barchart(...)
           }), position = c(0, 0, .5, 1), more = TRUE)
print(
  barchart(heart_disease/total ~ factor(snoring_severity), snore
           , col = "#0080ff", border = "transparent"
           , xlab = "Snoring severity", ylab = "Rate"
           , main = "Prop. of people with heart disease by level of severity"
           , sub = "(b)", scales = list(y = list(draw = FALSE))
           , panel = function(...){
             args <- list(...)
             panel.text(args$x, args$y, paste0(round(args$y*100, 2),"%"), pos = 3)
             panel.barchart(...)
           }), position = c(.5, 0, 1, 1))
# </code r> ========================================================================== #
```

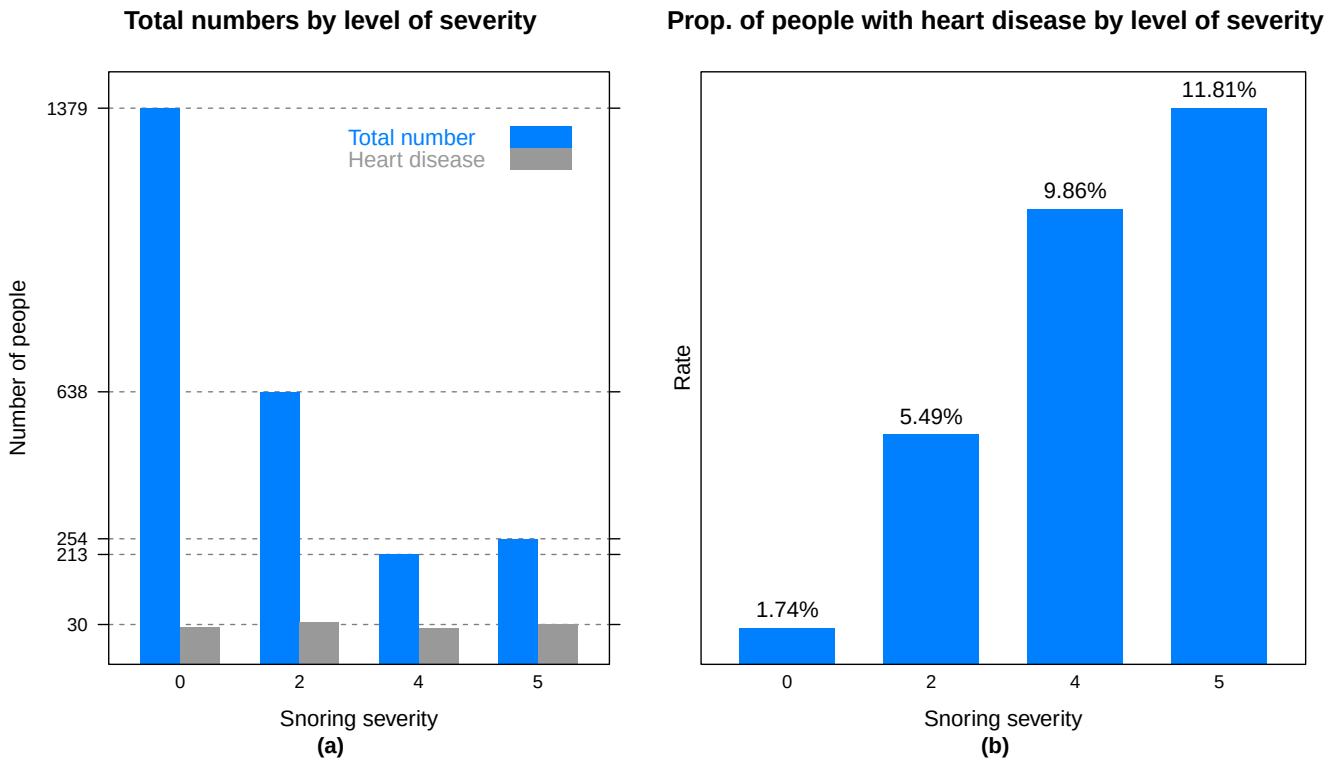

Figure 2: (a) Total number of people and number of people with heart disease by level of severity; (b) Proportion (people with heart disease/total) of people with heart disease by level of severity.

A big part, more than half, was classified by the spouses as a low degree of snoring severity. In this low levels the percentage of heart disease is very low (2, 5%). In the biggest degree, 5, the rate of heart disease is bigger, almost 12%, the double of the observed in the degree 2.

Exercise 3

Download the "calcium" data set from the course website. Provide the summary statistics (mean, standard deviation, and 5-number data summaries) and the histograms of variables "Begin" and "End". Which variable is higher on average, and which one is more dispersed? Calculate the variance of "End" for the placebo group manually; i.e., by first calculating the deviations from the mean. Provide the boxplots of variables "Begin" and "End" for different "Treatment" groups (i.e., calcium and placebo). In one paragraph, discuss your findings based on these graphs.

Solution:

```
# <code r> ============================================================================ #
(calcium <- read.table("~/Dropbox/CLASS-DROPBOX/BOOK-DATA/calcium.txt"
                       , header = TRUE, sep = ","))
# </code r> ============================================================================ #
```

```
   Treatment Begin End
1    Calcium   107 100
2    Calcium   110 114
3    Calcium   123 105
4    Calcium   129 112
5    Calcium   112 115
6    Calcium   111 116
7    Calcium   107 106
8    Calcium   112 102
9    Calcium   136 125
10   Calcium   102 104
11   Placebo   123 124
12   Placebo   109  97
13   Placebo   112 113
14   Placebo   102 105
15   Placebo    98  95
16   Placebo   114 119
17   Placebo   119 114
18   Placebo   112 114
19   Placebo   110 121
20   Placebo   117 118
21   Placebo   130 133
```

Minimum, 1st quartile, median, mean, 3rd quartile and maximum:

```
# <code r> ============================================================================ #
summary(calcium[ , -1])
# </code r> ============================================================================ #
```

```
     Begin          End
```

```
 Min.   : 98   Min.   : 95
 1st Qu.:109   1st Qu.:105
 Median :112   Median :114
 Mean   :114   Mean   :112
 3rd Qu.:119   3rd Qu.:118
 Max.   :136   Max.   :133
```

Standard deviation:

```
# <code r> ============================================================================ #
sd(calcium$Begin) ; sd(calcium$End)
# </code r> ============================================================================ #

[1] 9.708121
[1] 9.782638
```

Histograms:

```
# <code r> ============================================================================ #
par(mfrow = c(1, 2), mar = c(3, 4, 4, 2))
hist(calcium$Begin, col = "#0080ff", border = "orange", las = 1, xlab = ""
     , main = "Begin\n(a)", xlim = c(90, 140))
hist(calcium$End, col = "#0080ff", border = "orange", las = 1, xlab = ""
     , main = "End\n(b)", xlim = c(90, 140))
# </code r> ============================================================================ #
```

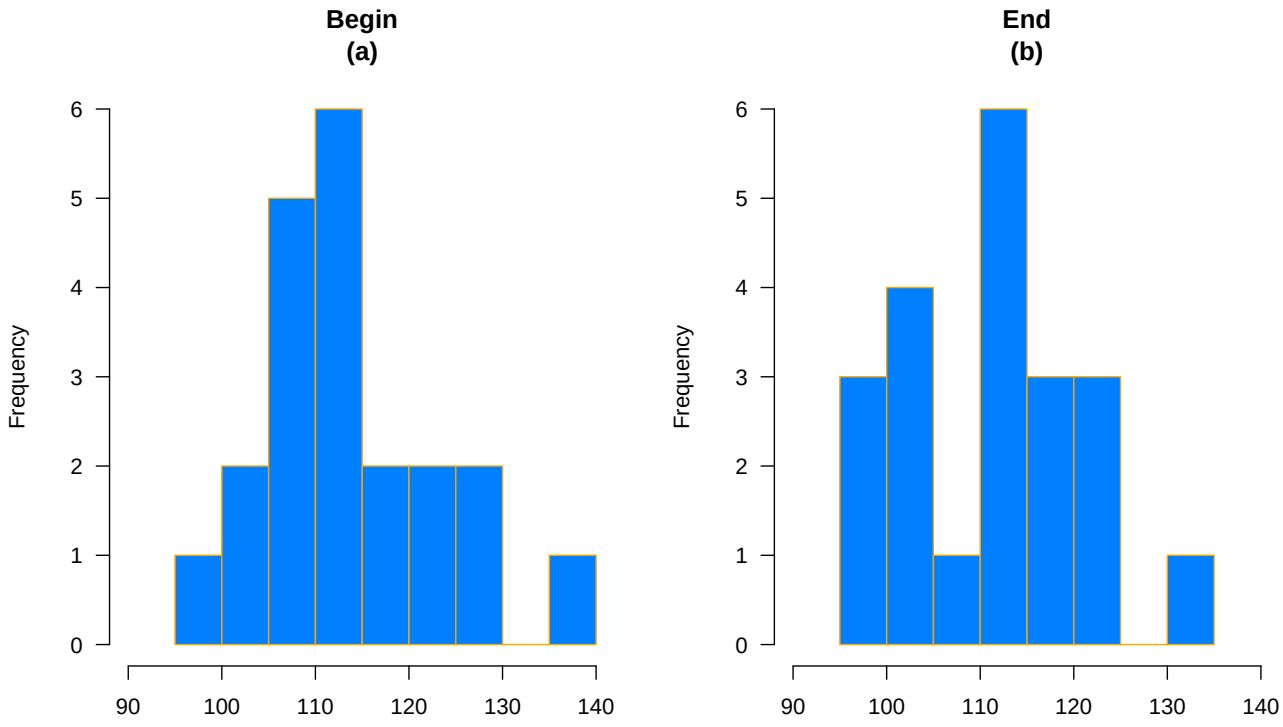

Figure 3: (a) Histogram of the "Begin" variable; (b) Histogram of the "End" variable.

In average, the variable "Begin" is higher, but both are very similar. The variable "End" is more dispersed, but with the same commentary, both the variables are very similar (standard deviations very closer to each other).

Variance of "End" for the placebo group:

```
# <code r> ============================================================================ #
(x <- calcium[calcium$Treatment == "Placebo", ]$End)

(x_bar <- mean(x))
# </code r> ============================================================================ #
```

```
 [1] 124  97 113 105  95 119 114 114 121 118 133
[1] 113.9091
```

```
# <code r> ============================================================================ #
(deviations <- x - x_bar)
# </code r> ============================================================================ #
```

```
 [1]  10.09090909 -16.90909091  -0.90909091  -8.90909091 -18.90909091
 [6]   5.09090909   0.09090909   0.09090909   7.09090909   4.09090909
[11]  19.09090909
```

```
# <code r> ============================================================================ #
(variance <- sum(deviations**2)/(length(x)-1)) ; variance == var(x)
# </code r> ============================================================================ #
```

```
[1] 128.2909
[1] TRUE
```

```
# <code r> ============================================================================ #
par(mfrow = c(1, 4), mar = c(1, 4, 4, 2))
boxplot(calcium[calcium$Treatment == "Calcium", ]$Begin, las = 1
        , main = "Calcium (a)\n(Begin)", ylim = c(90, 140))
boxplot(calcium[calcium$Treatment == "Calcium", ]$End, las = 1
        , main = "Calcium (b)\n(End)", ylim = c(90, 140))
boxplot(calcium[calcium$Treatment == "Placebo", ]$Begin, las = 1
        , main = "Placebo (c)\n(Begin)", ylim = c(90, 140))
boxplot(calcium[calcium$Treatment == "Placebo", ]$End, las = 1
        , main = "Placebo (d)\n(End)", ylim = c(90, 140))
# </code r> ============================================================================ #
```

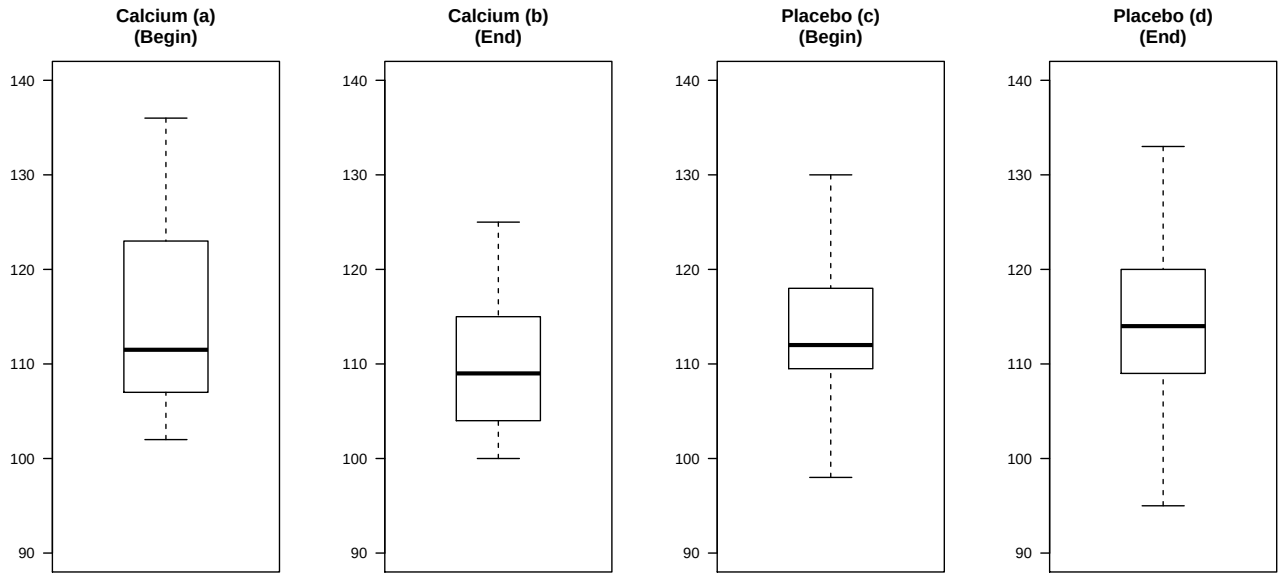

Figure 4: (a) Boxplot for the variable "Begin" in the group Calcium; (b) Boxplot for the variable "End" in the group Calcium; (c) Boxplot for the variable "Begin" in the group Placebo; (d) Boxplot for the variable "End" in the group Placebo.

We don't see much difference between "Begin" and "End" or between Calcium and Placebo. The biggest values are observed in the treatment Calcium in the "Begin", and the smallest values in the treatment Placebo in the "End". We see the smallest variance in the treatment Calcium in the "End", and the biggest variance in the treatment Placebo in the "End". Even with this findings, we reiterate that the differences are very small between the treatments and variables.

Exercise 4

Download the "wdbc" data set. Provide a bar graph for the discrete variable, and scatter plots for each pair of continuous variables. Provide the box plots of continuous variables for each level of "Diagnosis". In one paragraph, discuss your findings based on these graphs.

Solution:

```
# <code r> ============================================================================ #
library("kdevine") ; data(wdbc)

barchart(wdbc$diagnosis, col = "#0080ff", xlab = "Frequency", xlim = c(0, 410)
         , scales = list(x = list(draw = FALSE)
                         , y = list(labels = c("Benign", "Malignant")))
         , border = "transparent", main = "Diagnosis"
         , panel = function(...){
           args <- list(...)
```

```
            panel.text(args$x, args$y, pos = 4
                       , paste0(args$x," (", round(prop.table(args$x),3)*100,"%)"))
            panel.barchart(...)})
# </code r> ========================================================================== #
```

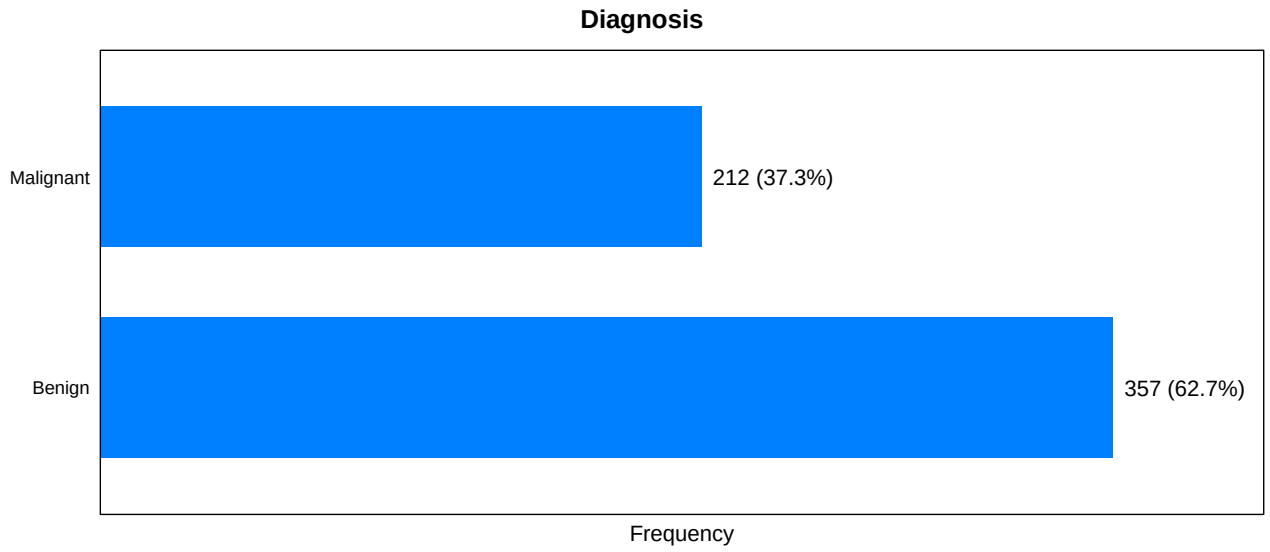

Figure 5: Frequency and percentage for each level of "Diagnosis".

This dataset contain measurements on cells in suspicious lumps in a women's breast. Variables are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. All samples are classsified as either *benign* or *malignant*.

Ten real-valued variables are computed for each cell nucleus.

The mean, standard error, and "worst" or largest (mean of the three largest values) of these variables were computed for each image, resulting in 30 variables.

This breast cancer database was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg.

Thus, we provide the scatter plots for each pair of variables divided by the mean, standard error and "worst" values, resulting in this way in three scatter plot matrices.

```
# <code r> ========================================================================== #
panel.cor <- function(x, y, ...){
  usr <- par("usr")
  on.exit(par(usr))
```

```r
  par(usr = c(0, 1, 0, 1))
  r <- abs(cor(x, y))
  txt <- format(c(r, 0.123456789), digits = 2)[1]
  text(0.5, 0.5, txt, cex = .8/strwidth(txt))
}
rotulos <- tm::removeWords(names(wdbc[2:11]), "mean ")
rotulos[c(5:6, 8, 10)] <- c("smooth", "compact", "concave", "fractal d.")

pairs(wdbc[2:11]
      , pch = 16
      , gap = .25
      , las = 1
      , upper.panel = panel.cor
      , col = c("#0080ff", "gray30")[unclass(wdbc$diagnosis)]
      , labels = rotulos
      , font.labels = 2
      , main = "Mean")
# </code r> ========================================================================== #
```

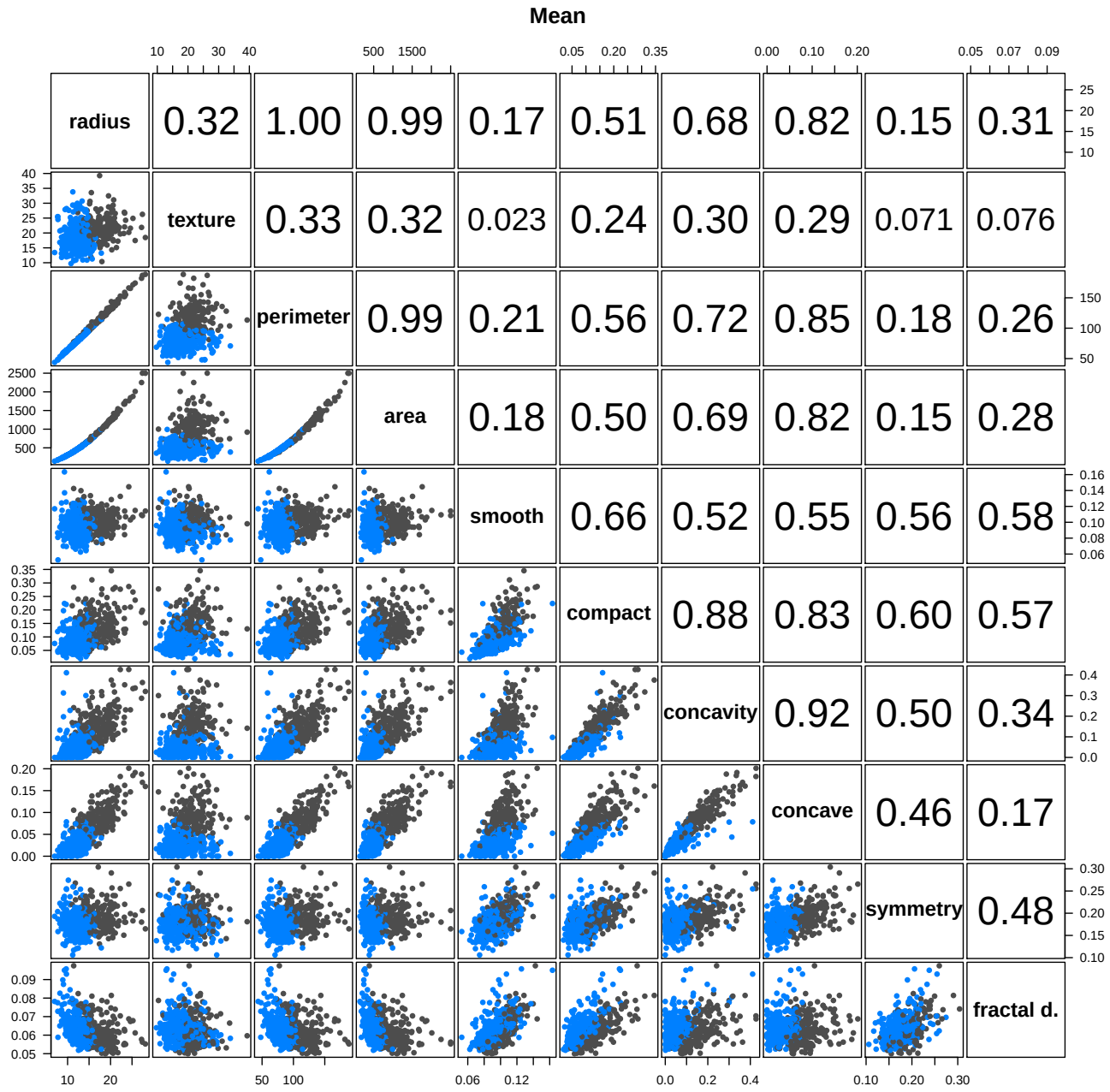

Figure 6: Scatter plots for each pair of mean variables in the lower triangular part of the matrix, in the upper triangular part we have the respectives correlations between the variables. In blue we have the corresponding observations to the benign diagnosis and in gray the corresponding observations to the malignant diagnosis.

```
# <code r> ============================================================================ #
rotulos <- tm::removeWords(names(wdbc[12:21]), "worst ")
rotulos[c(5:6, 8, 10)] <- c("smooth", "compact", "concave", "fractal d.")

pairs(wdbc[12:21], pch = 16, gap = .25, las = 1, upper.panel = panel.cor
      , col = c("#0080ff", "gray30")[unclass(wdbc$diagnosis)]
      , labels = rotulos, font.labels = 2, main = '"Worst"')
```

```
# </code r> ========================================================================== #
```

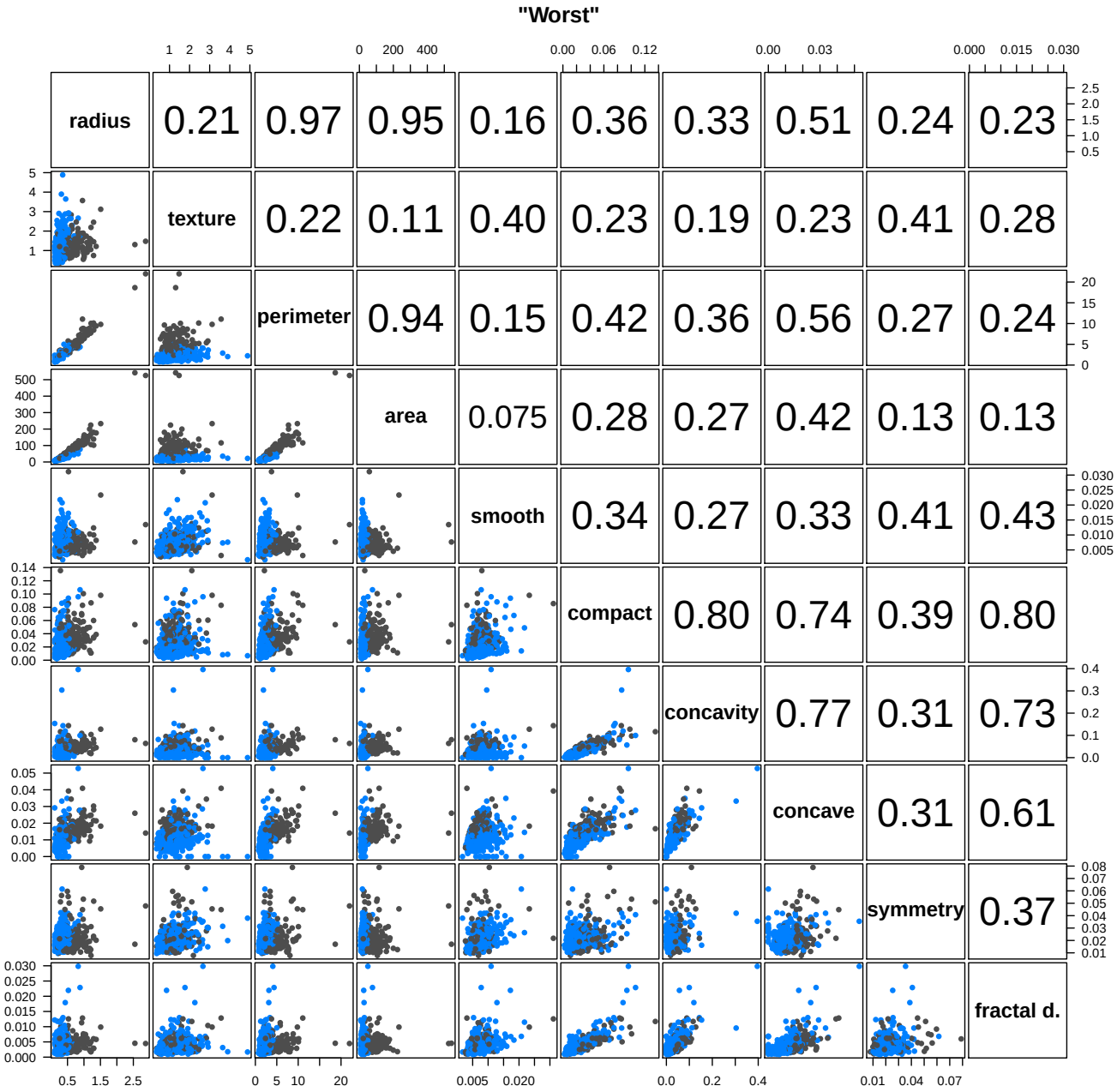

Figure 7: Scatter plots for each pair of "worst" variables in the lower triangular part of the matrix, in the upper triangular part we have the respectives correlations between the variables. In blue we have the corresponding observations of the benign diagnosis and in gray the corresponding observations of the malignant diagnosis.

```
# <code r> ========================================================================== #
rotulos <- tm::removeWords(names(wdbc[22:31]), "sd ")
rotulos[c(5:6, 8, 10)] <- c("smooth", "compact", "concave", "fractal d.")
```

```
pairs(wdbc[22:31], pch = 16, gap = .25, las = 1, upper.panel = panel.cor
      , col = c("#0080ff", "gray30")[unclass(wdbc$diagnosis)]
      , labels = rotulos, font.labels = 2, main = "Standard error")
# </code r> ========================================================================== #
```

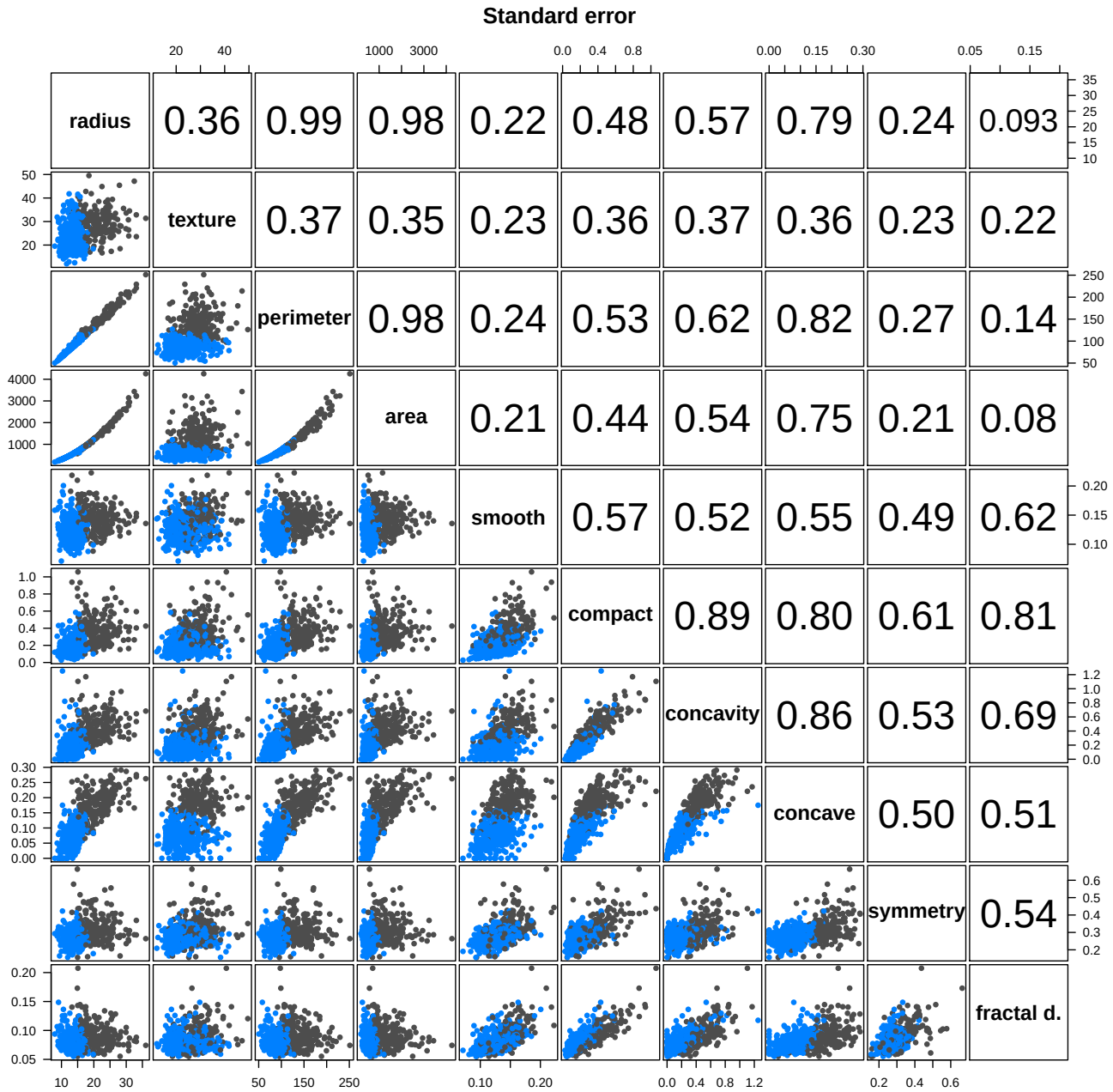

Figure 8: Scatter plots for each pair of standard error variables in the lower triangular part of the matrix, in the upper triangular part we have the respectives correlations between the variables. In blue we have the corresponding observations to the benign diagnosis and in gray the corresponding observations to the malignant diagnosis.

Box plots:

```
# <code r> ============================================================================ #
wdbc_mean <- reshape2::melt(wdbc[1:11], id.vars = "diagnosis")
levels(wdbc_mean$diagnosis) <- c("Benign", "Malignant")
levels(wdbc_mean$variable) <- tm::removeWords(levels(wdbc_mean$variable), "mean ")

bwplot(value ~ diagnosis | variable, wdbc_mean
       , layout = c(4, 3), scales = list(y = list(relation = "free", rot = 0))
       , ylab = NULL, strip = strip.custom(bg = "white"), pch = "|"
       , main = "Mean"
       , par.settings = list(
         box.rectangle = list(
           col = c("#0080ff", "gray60"), fill = c("#0080ff", "gray60")
           , alpha = .6, border = "transparent")
         , box.umbrella = list(col = c("#0080ff", "gray60"))
         , plot.symbol = list(col = "red", pch = 16, alpha = .6)))
# </code r> ============================================================================ #
```

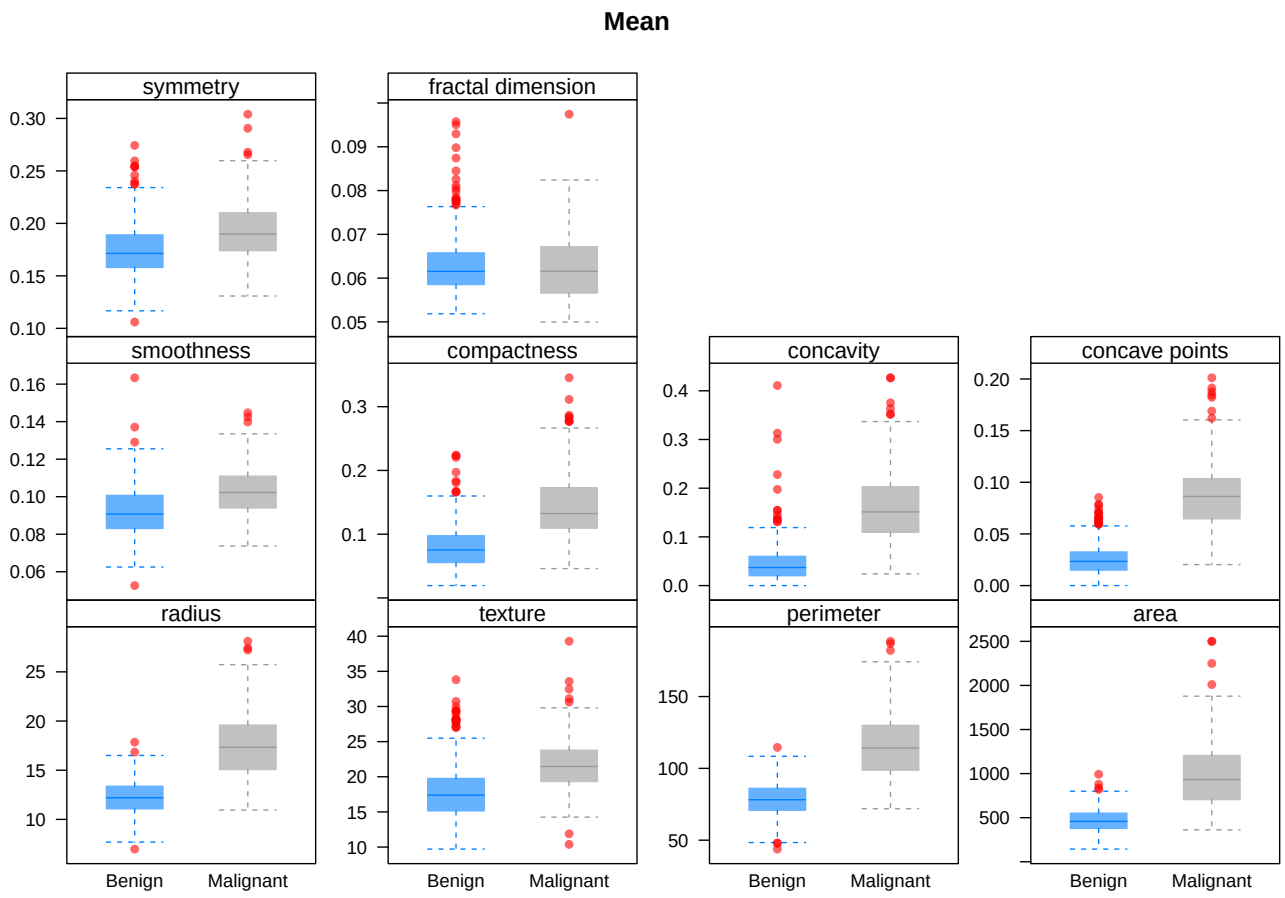

Figure 9: Box plots for each mean variable divided by "Diagnosis".

```r
# <code r> ============================================================================ #
wdbc_worst <- reshape2::melt(wdbc[c(1, 12:21)], id.vars = "diagnosis")
levels(wdbc_worst$diagnosis) <- c("Benign", "Malignant")
levels(wdbc_worst$variable) <-
  tm::removeWords(levels(wdbc_worst$variable), "worst ")

bwplot(value ~ diagnosis | variable, wdbc_worst
       , layout = c(4, 3), scales = list(y = list(relation = "free", rot = 0))
       , ylab = NULL, strip = strip.custom(bg = "white"), pch = "|"
       , main = '"Worst"'
       , par.settings = list(
         box.rectangle = list(
           col = c("#0080ff", "gray60"), fill = c("#0080ff", "gray60")
           , alpha = .6, border = "transparent")
         , box.umbrella = list(col = c("#0080ff", "gray60"))
         , plot.symbol = list(col = "red", pch = 16, alpha = .6)))
# </code r> ============================================================================ #
```

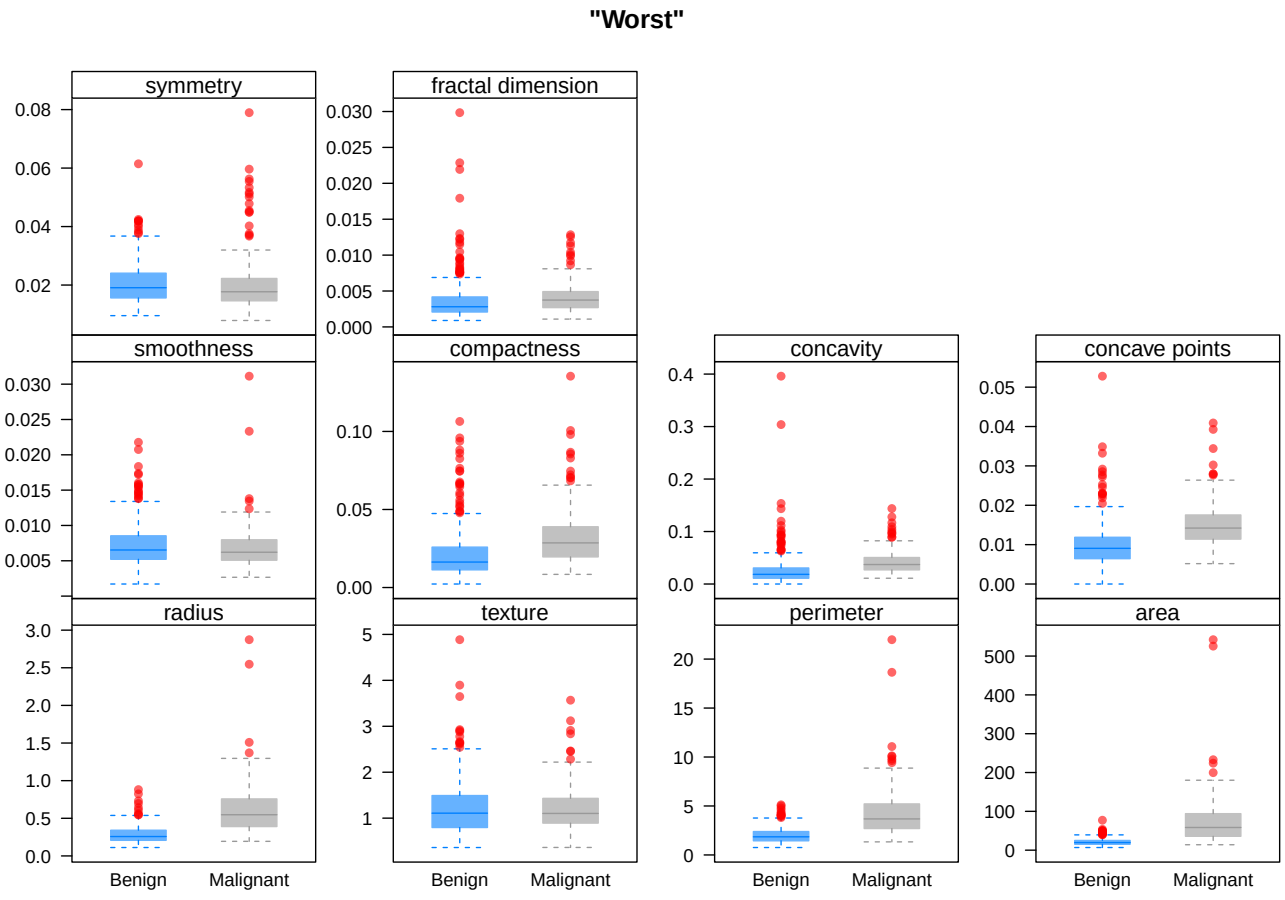

Figure 10: Box plots for each "worst" variable divided by "Diagnosis".

```
# <code r> ============================================================================ #
wdbc_sd <- reshape2::melt(wdbc[c(1, 22:31)], id.vars = "diagnosis")
levels(wdbc_sd$diagnosis) <- c("Benign", "Malignant")
levels(wdbc_sd$variable) <- tm::removeWords(levels(wdbc_sd$variable), "sd ")

bwplot(value ~ diagnosis | variable, wdbc_sd
       , layout = c(4, 3), scales = list(y = list(relation = "free", rot = 0))
       , ylab = NULL, strip = strip.custom(bg = "white"), pch = "|"
       , main = "Standard error"
       , par.settings = list(
         box.rectangle = list(
           col = c("#0080ff", "gray60"), fill = c("#0080ff", "gray60")
           , alpha = .6, border = "transparent")
         , box.umbrella = list(col = c("#0080ff", "gray60"))
         , plot.symbol = list(col = "red", pch = 16, alpha = .6)))
# </code r> ============================================================================ #
```

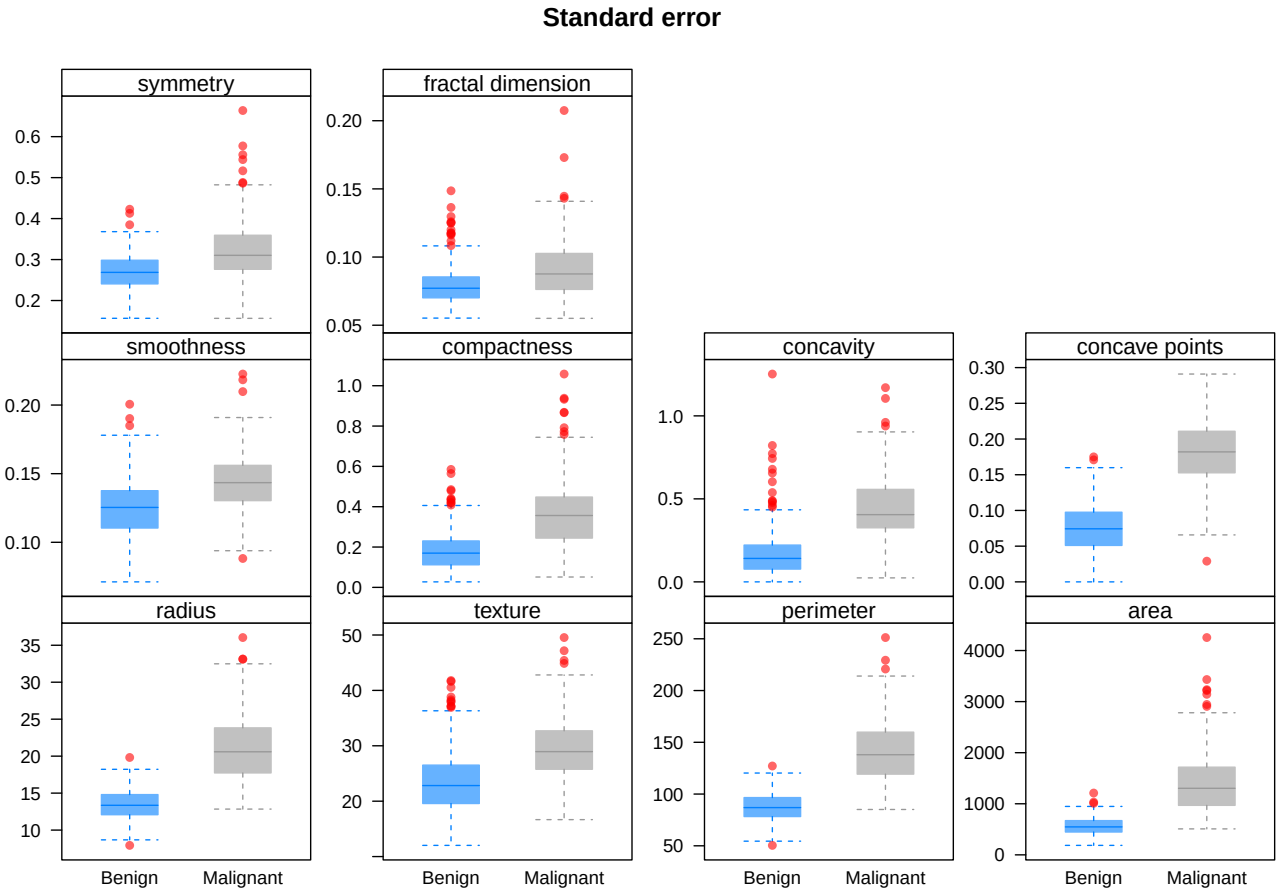

Figure 11: Box plots for each standard error variable divided by "Diagnosis".

Approximatelly 2/3 of the patients were with the exam classified as benign. When we look to the mean variables, three pairs present a correlation (linear correspondence) extremaly close to 1. Between this ten variables we see every behavior types. Relations with a very high correlation,

with a medium correlation and with a very small correlation. In general, the same behavior is seen with the same variables when we looked to the "worst" and standard error variables. We can highlight as very high correlations the relations between the variables area and perimeter, radius and perimeter, and radius and area. As a very small correlation we can highlight the relations between the variables texture and symmetry, texture and fractal dimension, area and symmetry, and area and fractal dimension. When we looked to the box plots, in general, for all the ten variables in the three different measure types, the biggest values are present in the patients classified as malignant.

Exercise 5

Consider this dataset derived from the Framingham Heart Study where one of the goals was to study possible links between high systolic blood pressure (SBP) and coronary heart disease (CHD). Participants with SBP $\geq$ 165 mm Hg were put in the "high" SBP category.

	CHD	No CHD
High SBP	144	62
Normal SBP	120	419

(a)

Marginal proportions. **From the table, what is the proportion of participants who had CHD? What is the proportion of participants who had high SBP?**

Solution:

CHD = 144 + 120 = 264, NoCHD = 62 + 419 = 481, n = CHD + NoCHD = 745

Proportion of participants who had CHD: CHD/n = 264/745 = 0.3543624 (35%).

HiSBP = 144 + 62 = 206, NorSBP = 120 + 419 = 539,
n = CHD + NoCHD = HiSBP + NorSBP = 745

Proportion of participants who had high SBP: HiSBP/n = 206/745 = 0.2765101 (28%).

(b)

Conditional proportions. **Among the participants with high SBP, what is the proportion of those who also have CHD? Among the participants with normal SBP, what is the proportion of those who also have CHD? Does this provide some evidence of a link between elevated systolic blood pressure and coronary heart disease? Can one now claim that elevated systolic blood pressure causes coronary heart disease?**

Solution:

Among the participants with high SBP, the proportion of those who also have CHD:
CHD/HiSBP = 144/206 = 0.6990291 (70%).

Among the participants with normal SBP, the proportion of those who also have CHD:
CHD/NorSBP = 120/539 = 0.2226345 (22%).

Does this provide some evidence of a link between elevated systolic blood pressure and coronary heart disease? Yes. Because in the patients with normal SBP the proportion with CHD is 22 percent, almost 1/5, while in the patients with high SBP the proportion with CHD is 70 percent, much higher.

Can one now claim that elevated systolic blood pressure causes coronary heart disease? No. With this descriptive analysis we can see a possible correlation between high SBP and CHD, but this not implies in causality. we don't have enough information to claim this type of result. Correlation is different form causality.

(c)

Odds and odds ratios. **Compute the odds of having CHD among the high SBP group. Compute the odds of having CHD among the normal SBP group. Compute the odds ratio of having CHD for the high SBP vs normal SBP groups.**

Solution:

Odds of having CHD among the high SBP group:

$$\frac{144/206}{1-144/206} = 2.322581.$$

Odds of having CHD among the normal SBP group:

$$\frac{120/539}{1-120/539} = 0.2863962.$$

Odds ratio of having CHD for the high SBP vs normal SBP groups:

$$\frac{2.322581}{0.286396} = 8.109678.$$

The odds of having CHD for the high SBP groups is 8.11 times of the odds of CHD for the normal SBP group.

Exercise 6

The beetle mortality data. Groups of beetles were exposed to varying doses of toxins and the number of deaths (`ytotal`) out of the total exposure (`ntotal`) were recorded. Enter the variables in R:

$$\begin{aligned} \texttt{dose} &= \mathrm{c}(1.69, 1.72, 1.75, 1.78, 1.81, 1.84, 1.86, 1.88) \\ \texttt{ntotal} &= \mathrm{c}(59, 60, 62, 56, 63, 59, 62, 60) \\ \texttt{ytotal} &= \mathrm{c}(6, 13, 18, 28, 52, 53, 61, 60) \end{aligned}$$

```
# <code r> ============================================================================ #
dose <- c(1.69, 1.72, 1.75, 1.78, 1.81, 1.84, 1.86, 1.88)
ntotal <- c(59, 60, 62, 56, 63, 59, 62, 60)
ytotal <- c(6, 13, 18, 28, 52, 53, 61, 60)
# </code r> ============================================================================ #
```

(a)

Calculate the proportion of deaths for each dose-group.

Solution:

```
# <code r> ============================================================================ #
rbind(dose, "death proportion" = round(ytotal/ntotal, 2))
# </code r> ============================================================================ #
```

```
                [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
dose            1.69 1.72 1.75 1.78 1.81 1.84 1.86 1.88
death proportion 0.10 0.22 0.29 0.50 0.83 0.90 0.98 1.00
```

(b)

Plot the proportion of deaths against dose. Describe the trend: is it increasing/ decreasing, is there an asymptote feature?

Solution:

```
# <code r> ============================================================================ #
xyplot(ytotal/ntotal ~ dose, type = c("p", "l")
       , pch = 19, lwd = 1.5, xlab = "Dose", ylab = "Proportion of deaths"
       , scales = list(x = list(at = dose))
       , panel = function(...){
         panel.abline(v = dose, h = seq(.2, 1, .2), col = "gray70")
         panel.xyplot(...)})
# </code r> ============================================================================ #
```

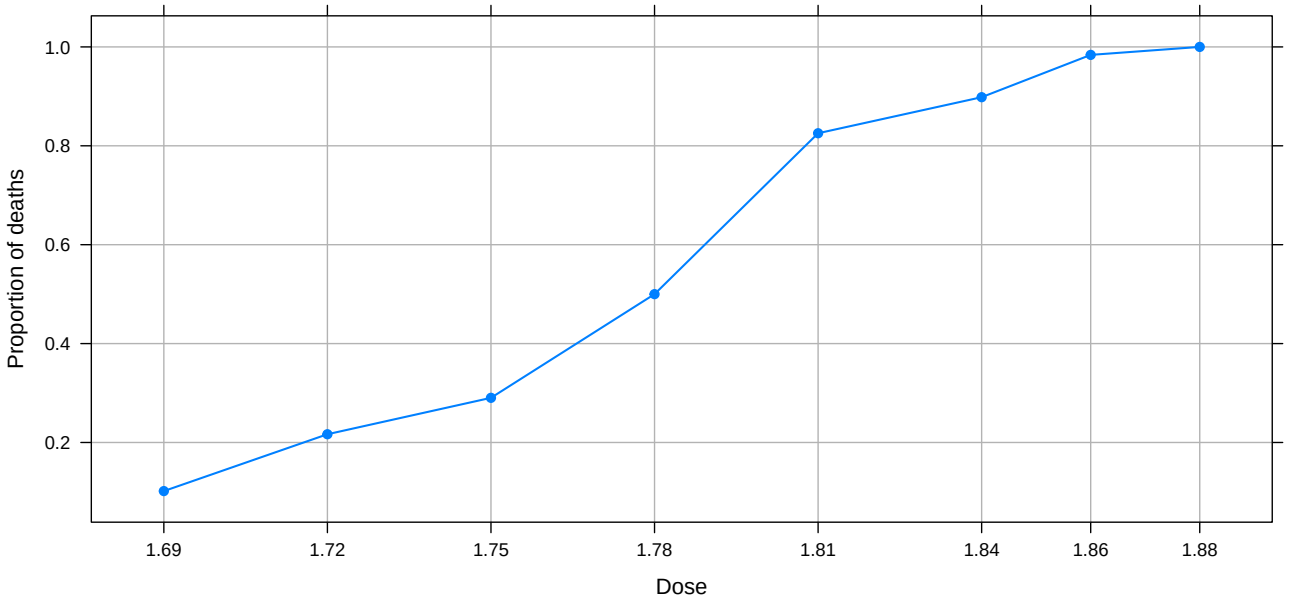

Figure 12: Proportion of deaths against dose.

The trend is increasing. For small doses the proportions are very close to zero, with the biggest doses the proportions are very close to one. With the last dose we have a proportion equal to one, the maximum possible. Whats means that with the biggest dose all the beetles are dead.

■