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
1

Henrique Aparecido Laureano

Fall Semester 2017

Contents

Exercise 1

The following graph (from Lindamer et. al.) shows the distributions of percentages of males and females for different age of onset of Schizophrenia. In one paragraph, discuss your findings based on this graph.
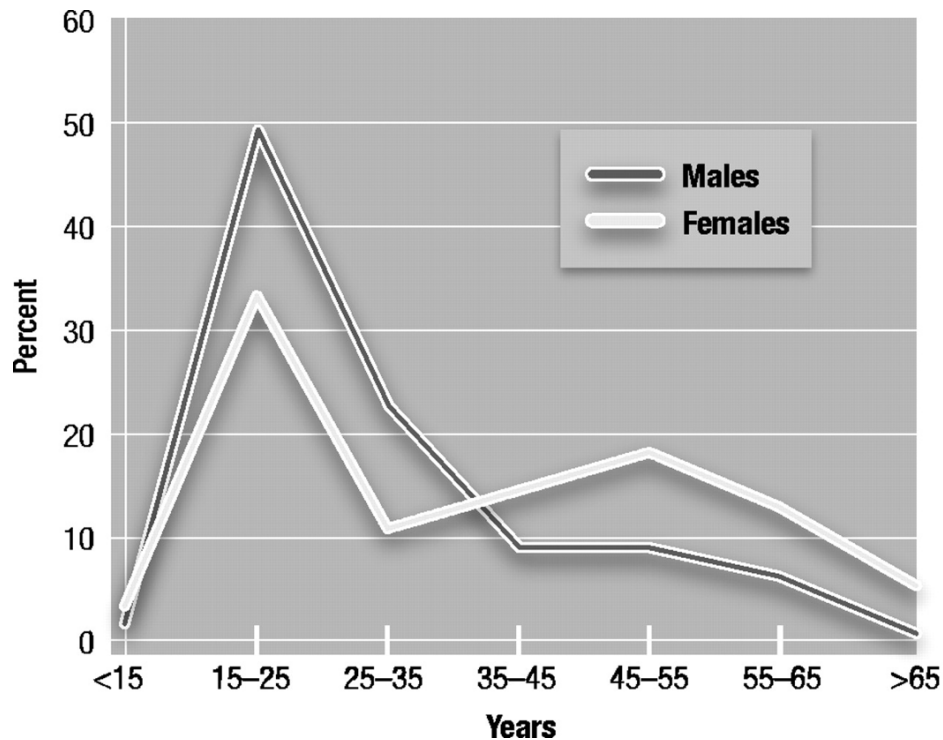

Figure 1: Percentages of males and females for different age of onset of Schizophrenia.

Solution:

For males with less than 15 years and more than 65 years old, the percentage of onset of Schizophrenia is praticaly zero, in the other side, for females, the behavior isn't the same. For females with less than 15 years the percentage is closer to 3% and for more than 65 years old the percentage is a little bigger than 5%. For males between 15 and 65 years old, we see that until 25 years the percentage grows, with its maximum in around 25 years. After this age the percentage decays fast until 45 years old and we see a constant behavior between 45 and 55 years old, in others words, for this ten years interval, the percentage in males is the same, decreasing again after this age, but with a smaller velocity. For females, until 35 years old the behavior is the same, but with a smallest intensity (the maximum is smaller. For male is around 50% and for female is around 35%). Different from the observed with males, from 25 years to 45 years old, the percentage increases in females, not with the same velocity that between 15 and 25 years old, but increases, reaching a local maximum around 20%. After this age, like for the males, the percentage decreases, in a very similar rate. In the end, with base in this graph, we can say that the percentage of onset of Schizophrenia is bigger in females than in males for the intervals < 15 and > 35-45, and bigger in males in the complementar interval (age between 15 and 35 years old, approximately).

Exercise 2

Download the "snore" data set from the class website. Choose an appropriate graph to show how the heart disease rate changes over all possible values of snoring severity. In one very brief paragraph, discuss your findings based on this graph.

Solution:

```r
# <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
```

```r
# <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> ================================================================= #
```

Total numbers by level of severity

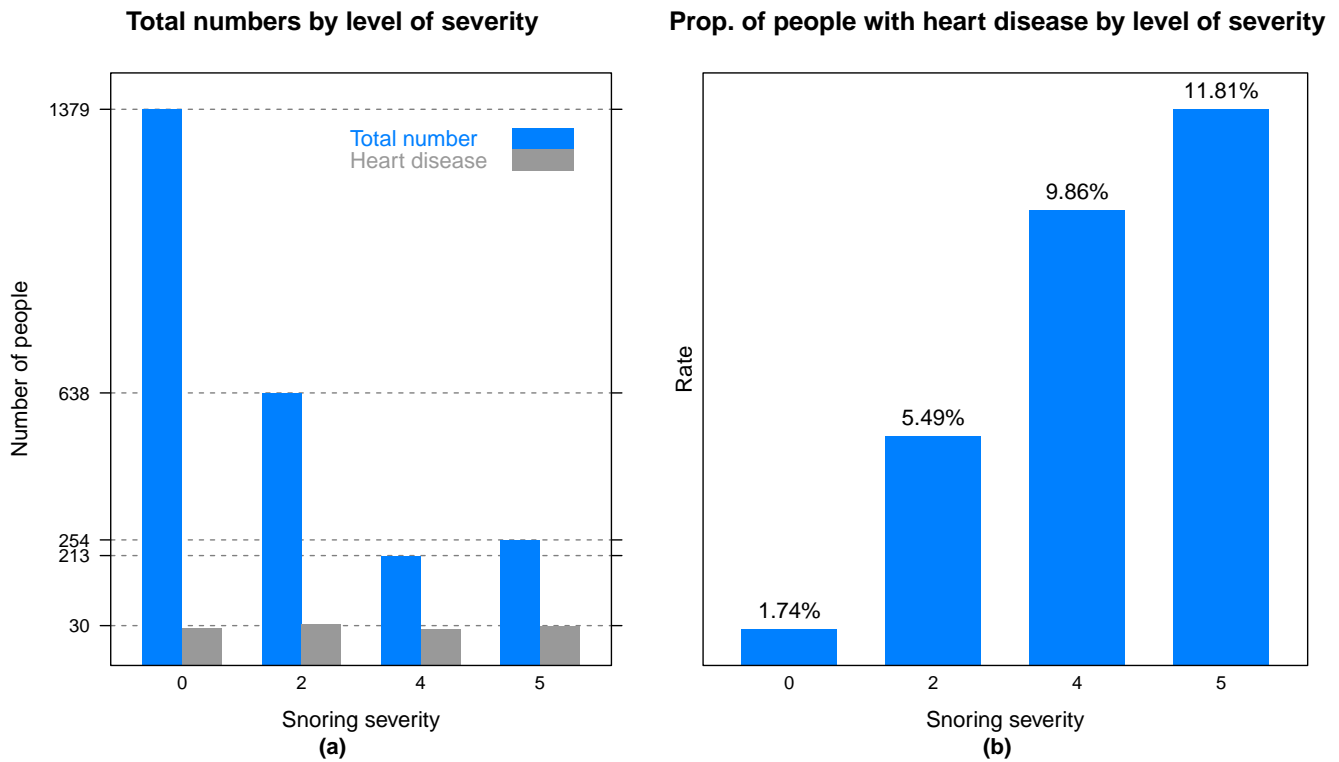**Prop. of people with heart disease by level of severity**

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:

```r
# <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:

```r
# <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> ============================================================== #
```
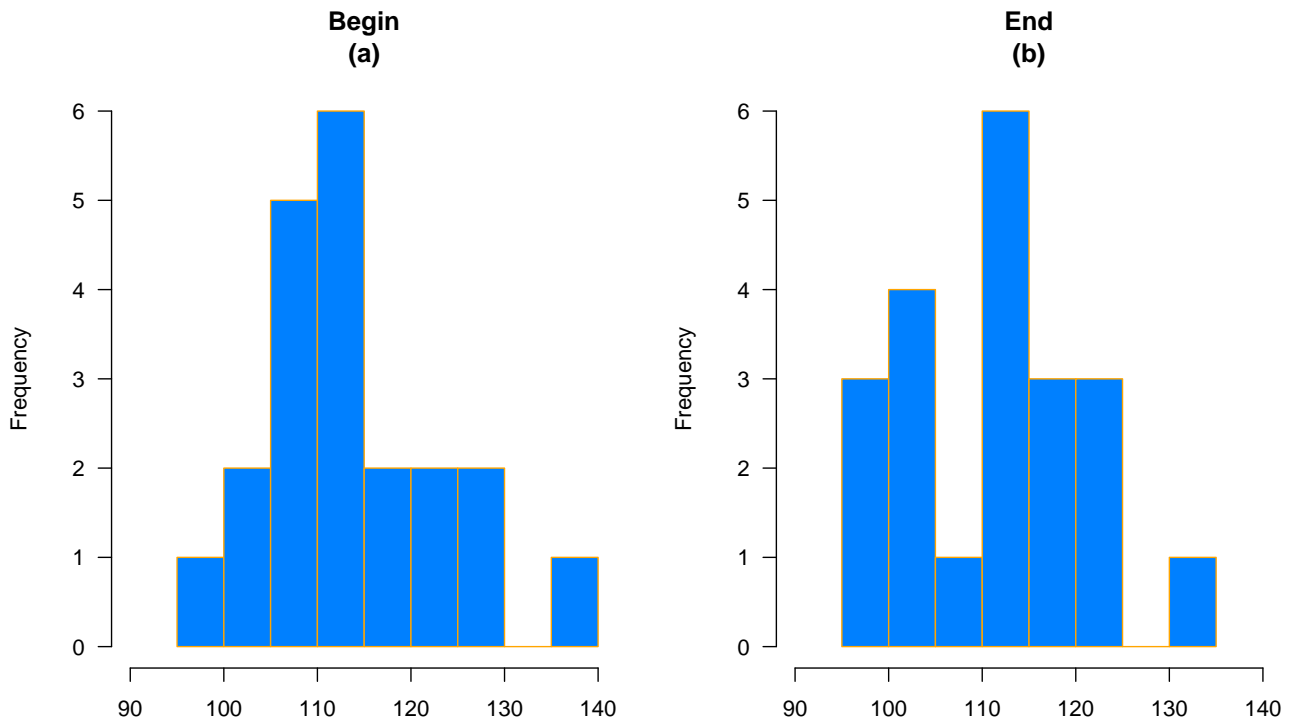

Figure 3: (a) Histogram of the "Begin" variable; (b) Histogram of the "End" variable.

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:

```r
# <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
```

```r
# <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
```

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

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

```r
# <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> =============================================================== #
```
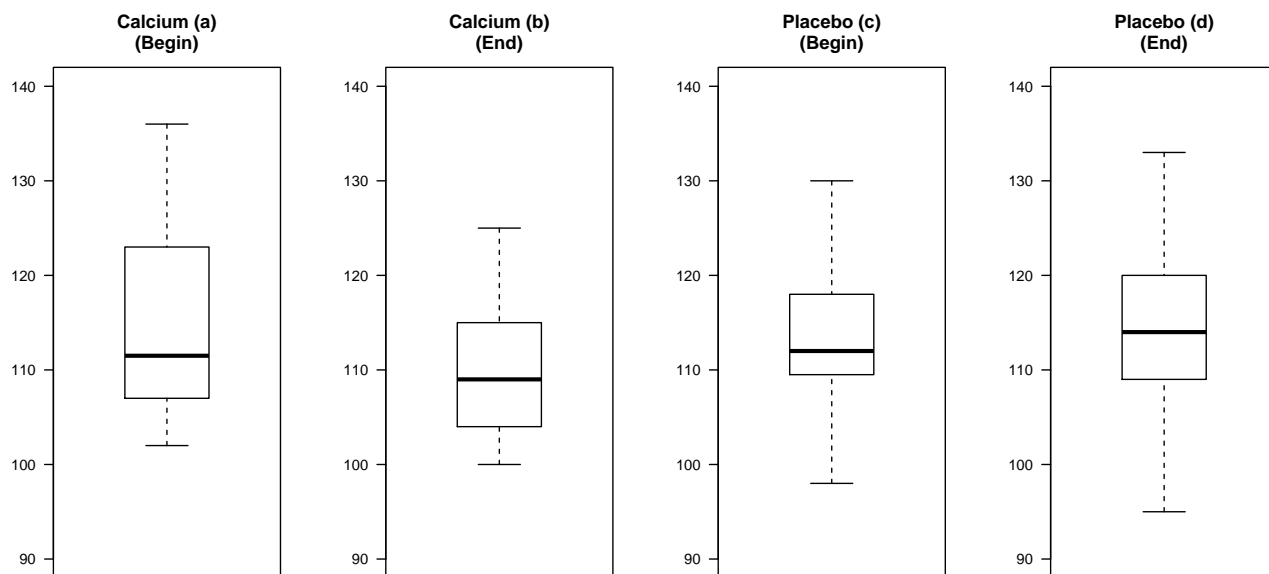

Figure 4: (a) Boxplot for the variable "Begin" in the group Calcium; (b) Boxplot for the variable "End" in the group Calcium; (c) Boxplot for the variable "Begin" in the group Placebo; (d) Boxplot for the variable "End" in the group Placebo.

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.

<u>Solution:</u>

```r
# <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> ========================================================== #
```
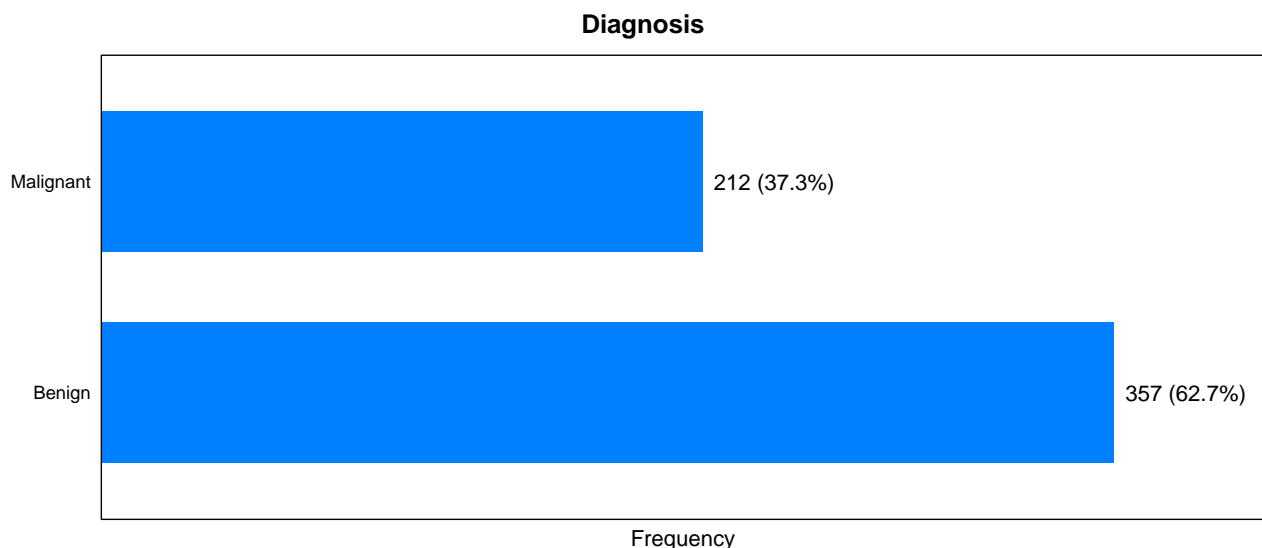
Diagnosis

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.

```r
# <code r> ==================================================================== #
panel.cor <- function(x, y, ...){
  usr <- par("usr")
  on.exit(par(usr))
  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> ==================================================================== #
```
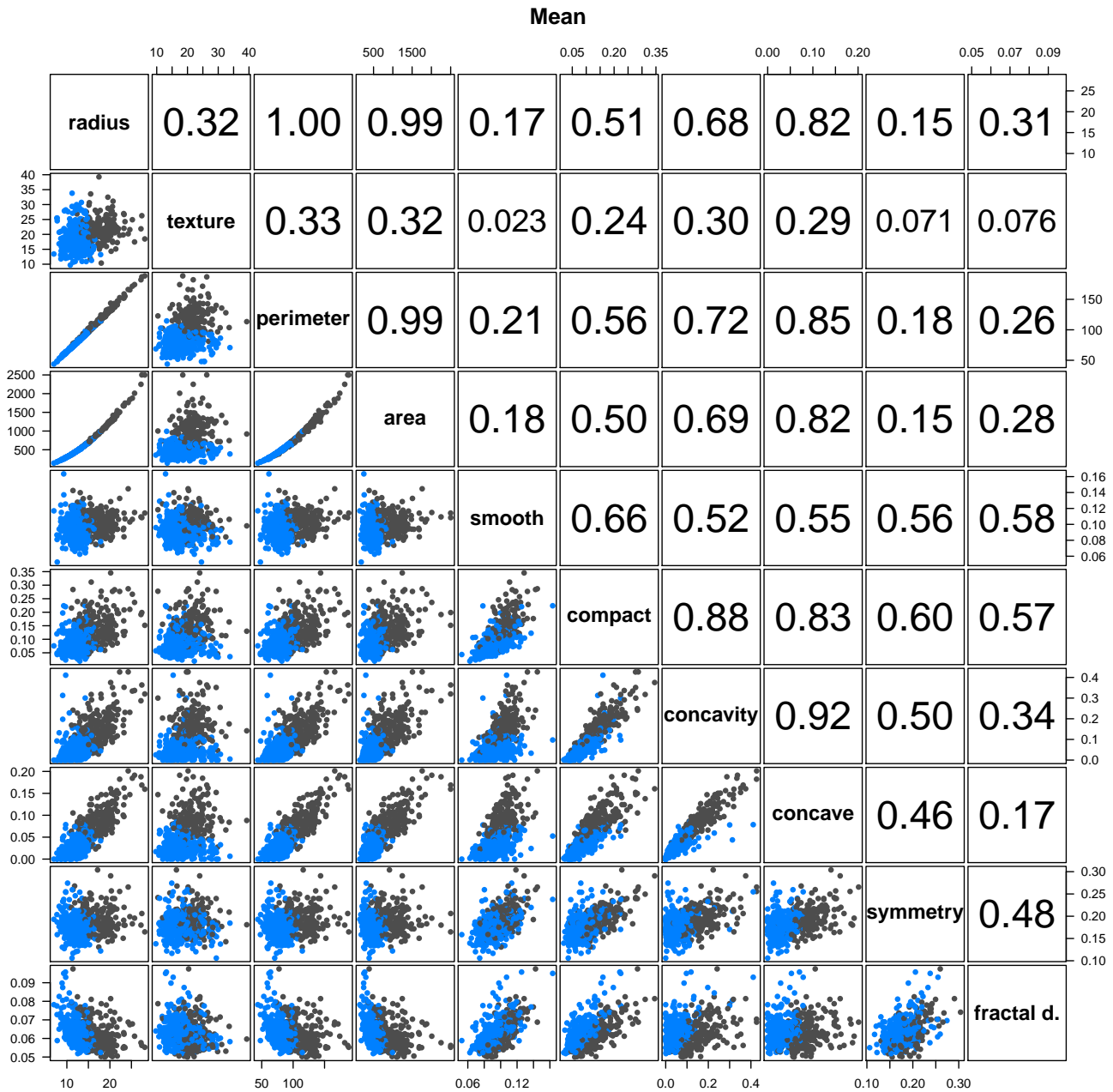
Figure 6: Scatter plots for each pair of mean variables in the lower triangular part of the matrix, in the upper triangular part we have the respectives correlations between the variables. In blue we have the corresponding observations to the benign diagnosis and in gray the corresponding observations to the malignant diagnosis.

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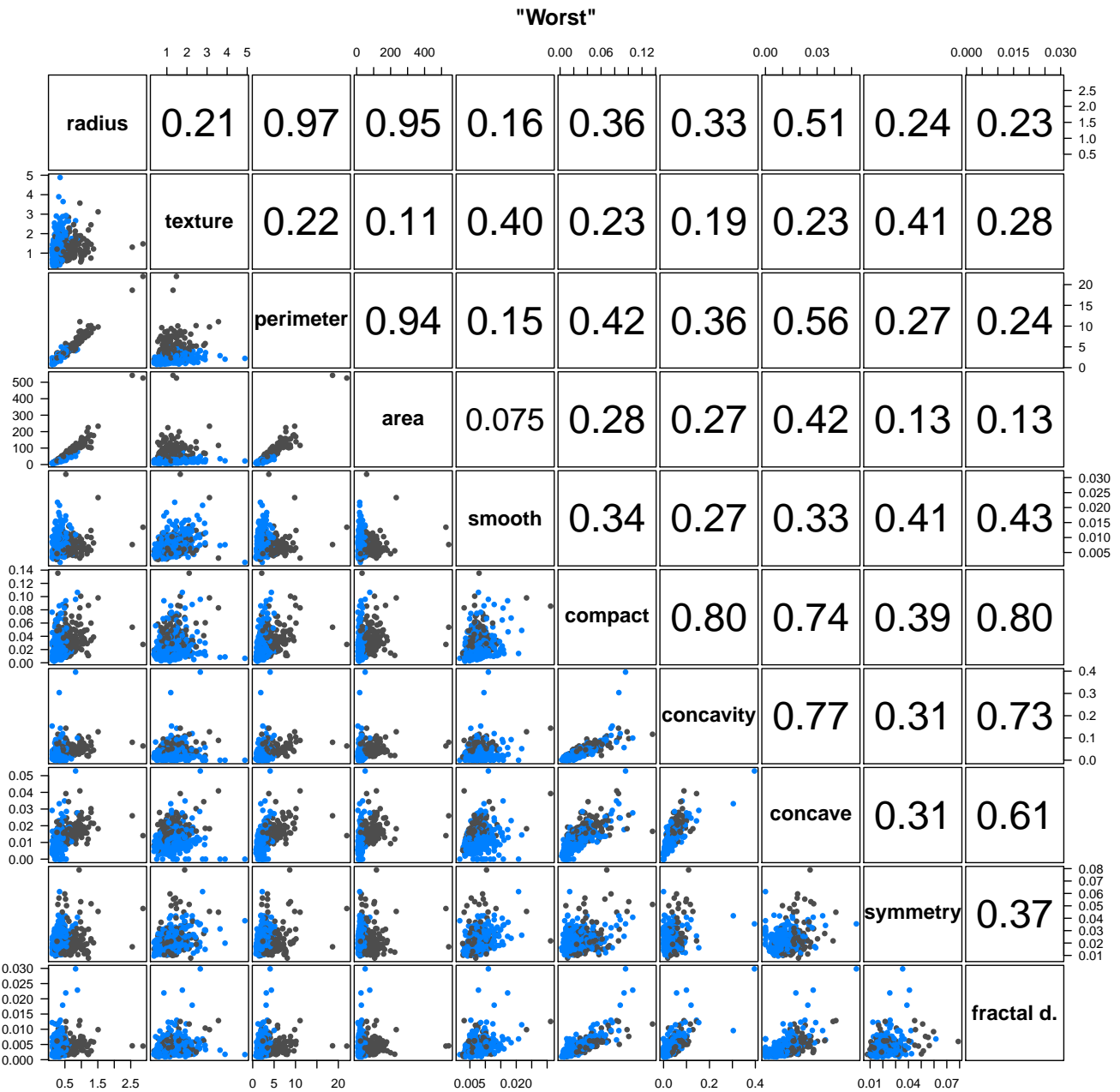
Figure 7: Scatter plots for each pair of "worst" variables in the lower triangular part of the matrix, in the upper triangular part we have the respectives correlations between the variables. In blue we have the corresponding observations of the benign diagnosis and in gray the corresponding observations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

```r
# <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> ================================================================= #
```
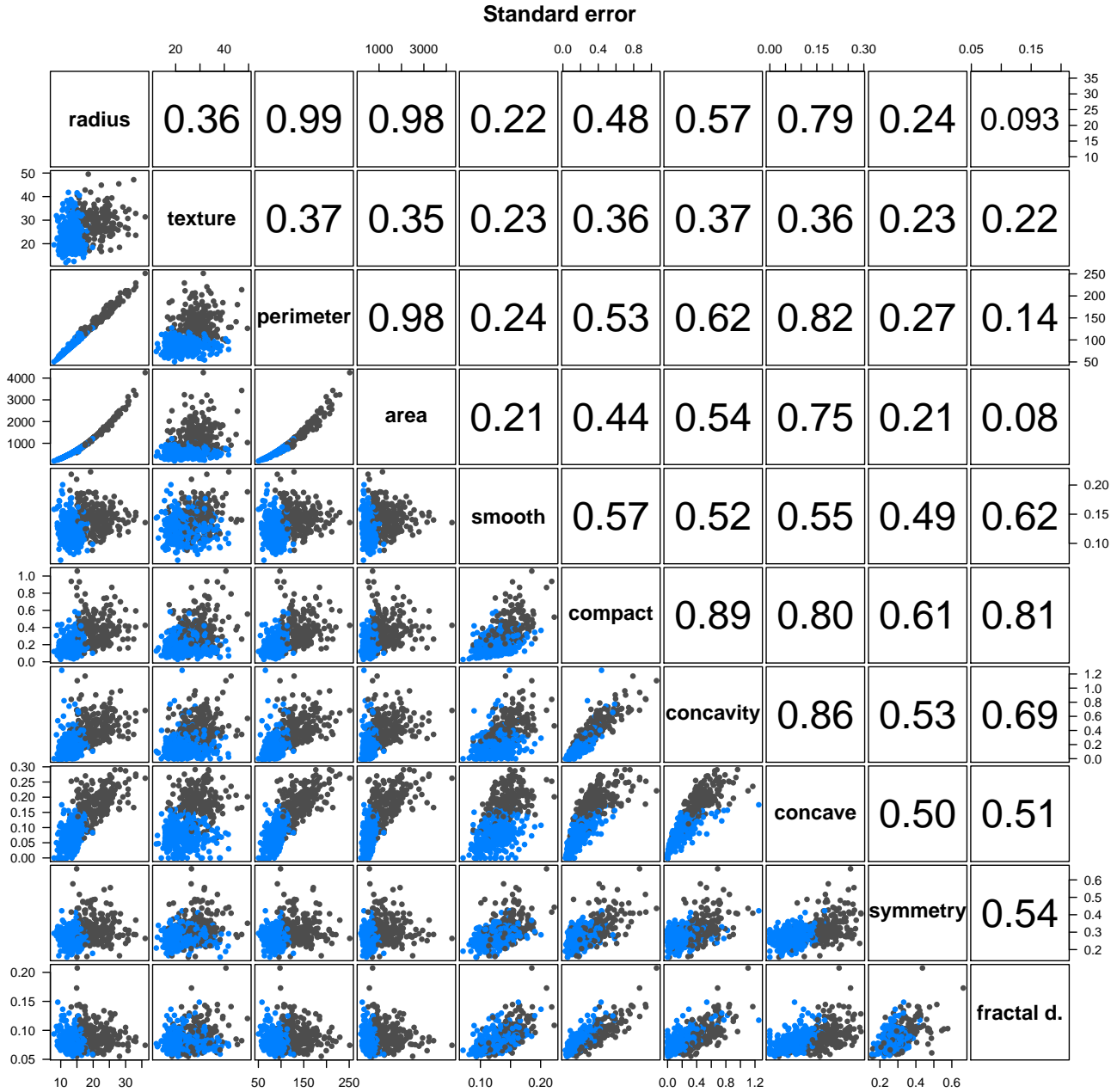

Figure 8: Scatter plots for each pair of standard error variables in the lower triangular part of the matrix, in the upper triangular part we have the respectives correlations between the variables. In blue we have the corresponding observations to the benign diagnosis and in gray the corresponding observations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

Box plots:

```r
# <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> ========================================================================= #
```
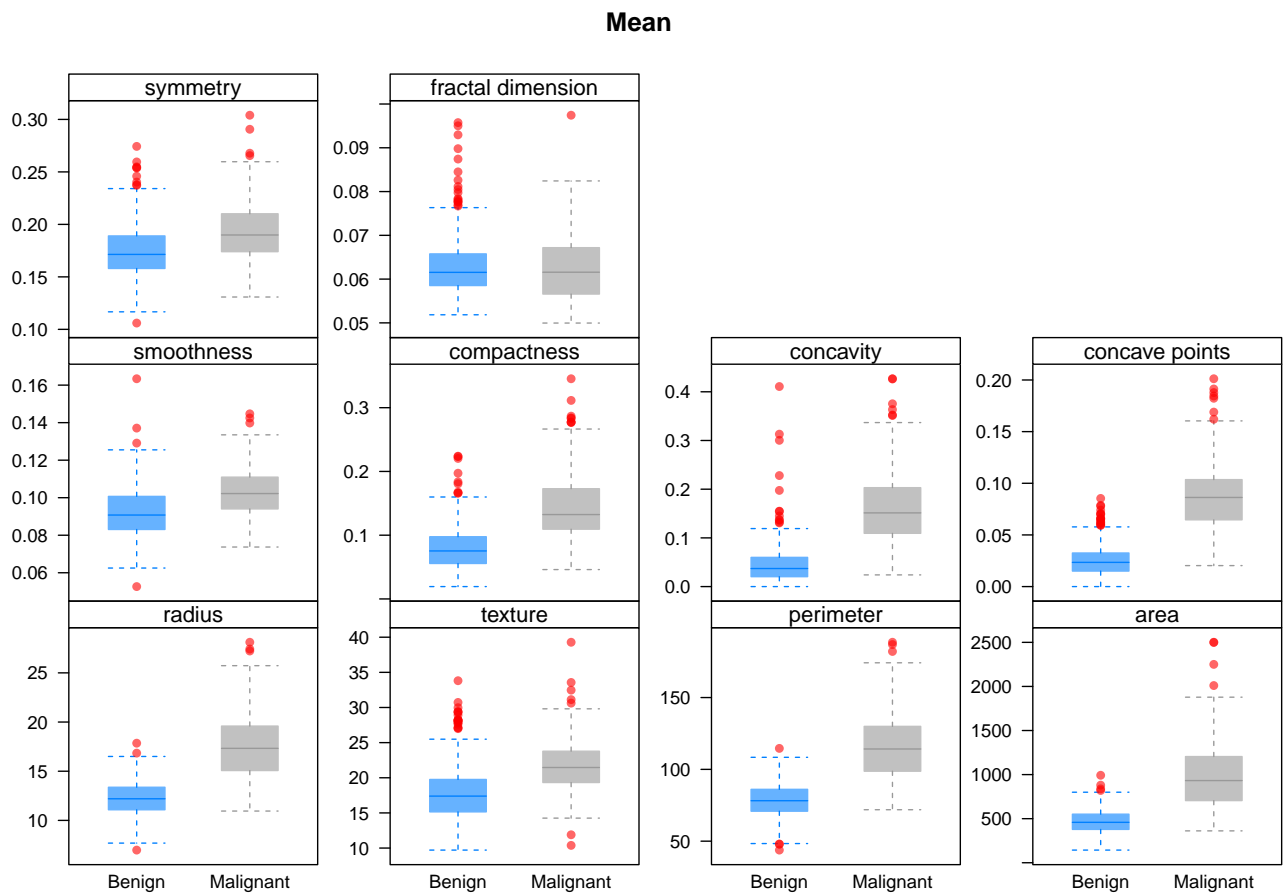

Figure 9: Box plots for each mean variable divided by "Diagnosis".

```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> ================================================================= #
```
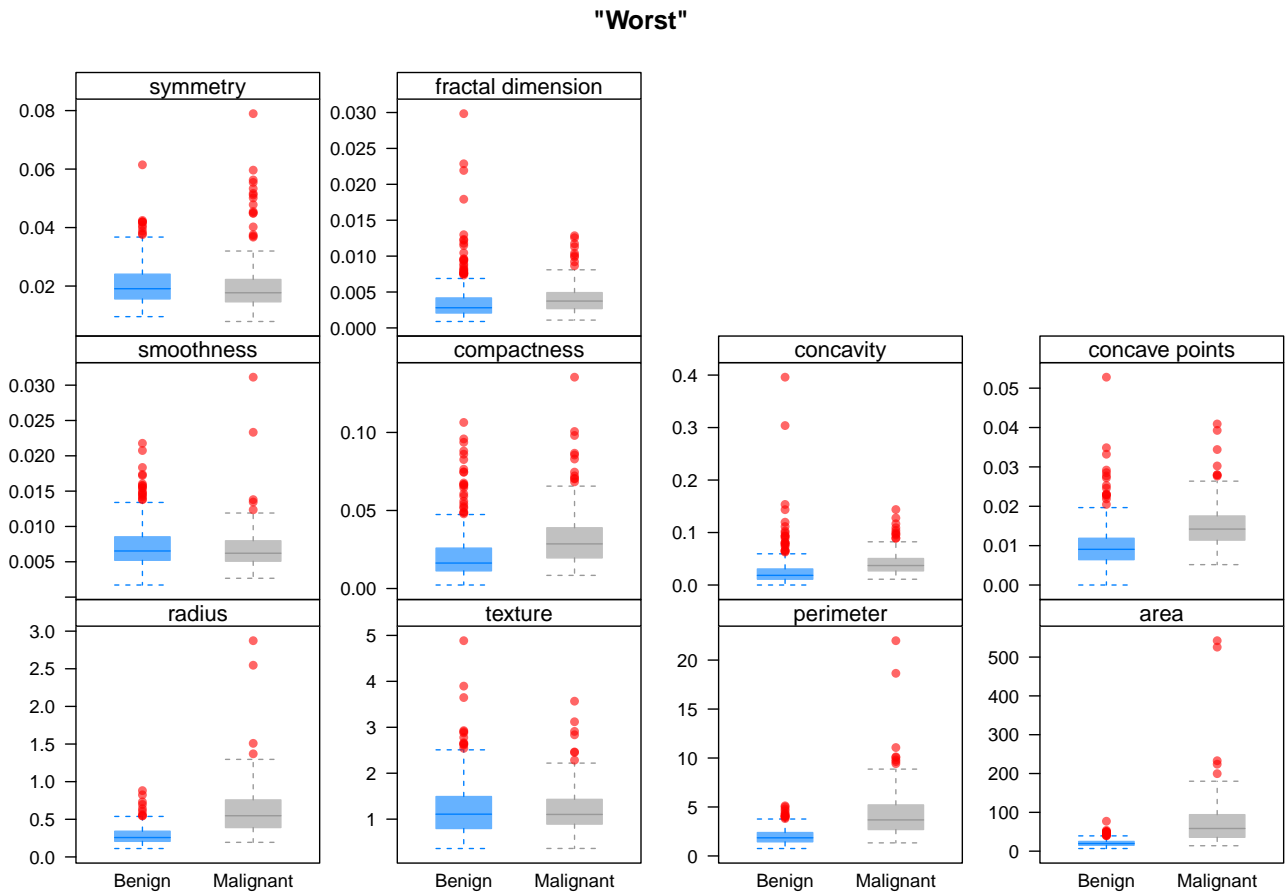

Figure 10: Box plots for each "worst" variable divided by "Diagnosis".

```r
# <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> ==================================================================== #
```
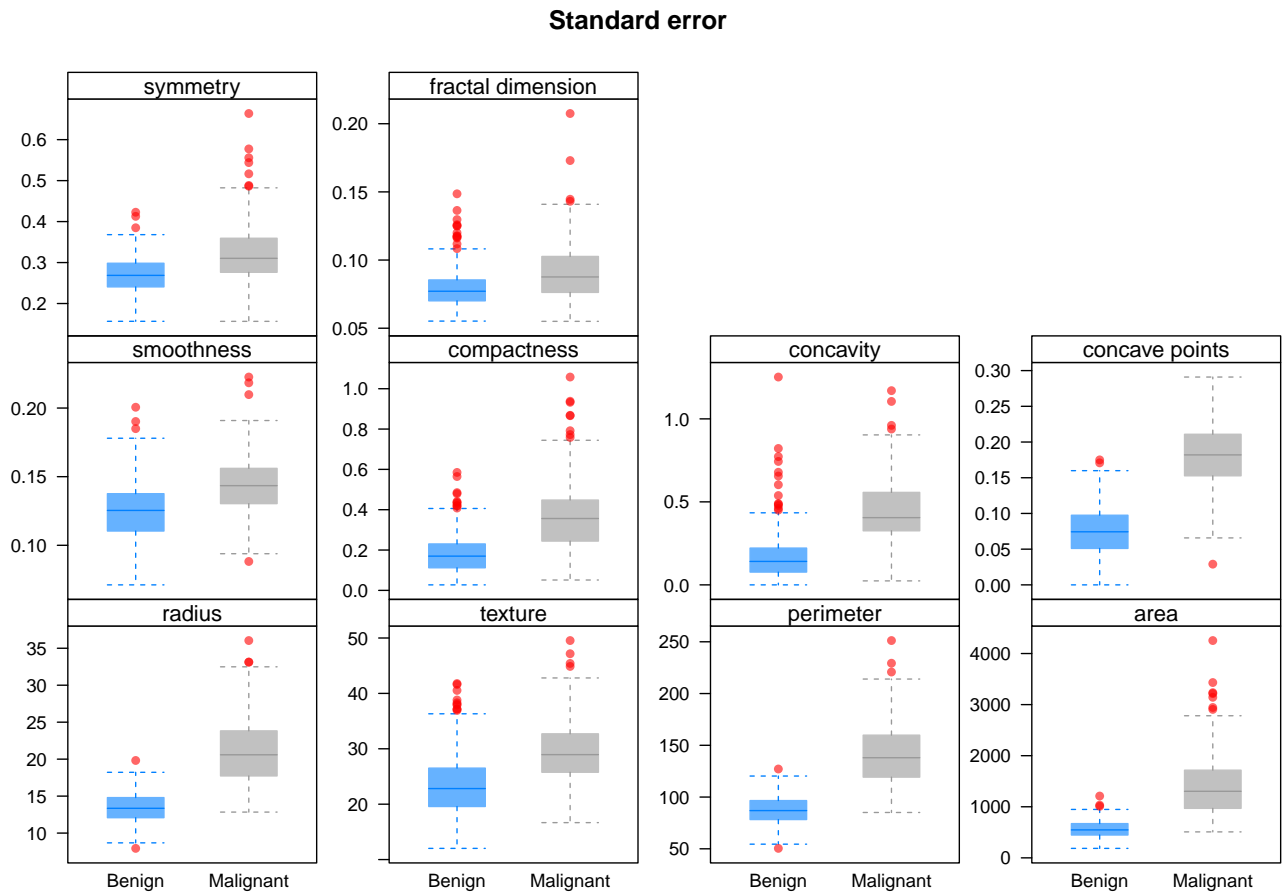

Figure 11: Box plots for each standard error variable divided by "Diagnosis".

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,

15

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 <u>causes</u> 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:

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

```r
# <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:

```r
# <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:

```r
# <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> ================================================================= #
```
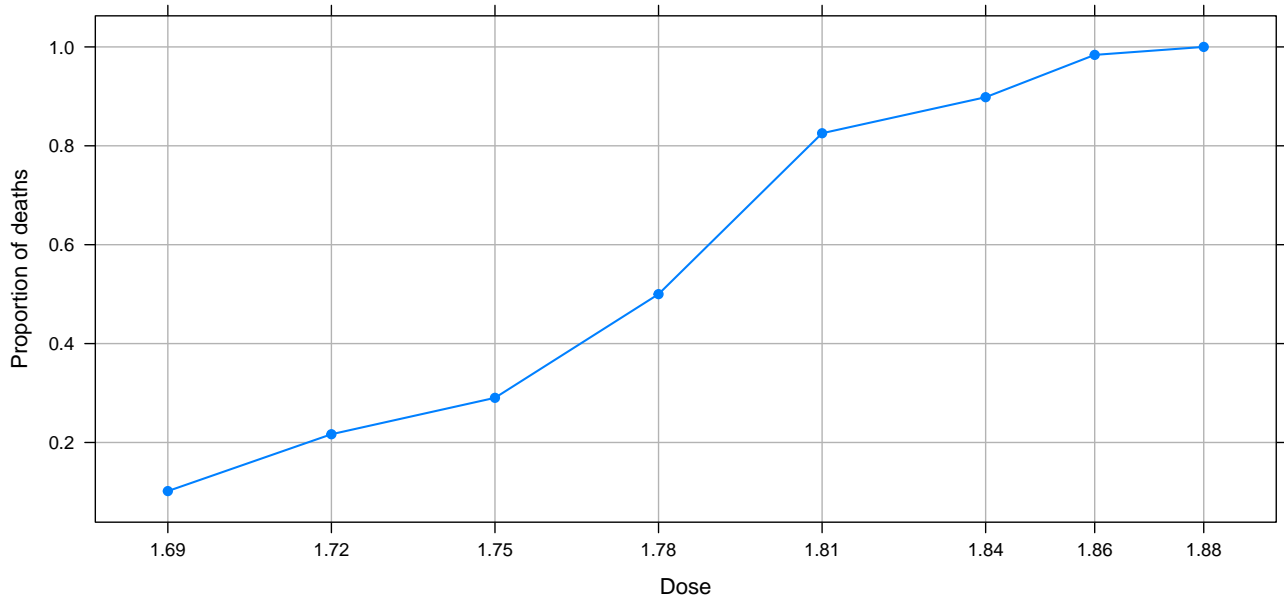
Figure 12: Proportion of deaths against dose.

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.

∎