Statistical test for the heart failure dataset (Ahmad et al. 2017, PlosOne)

Giuseppe Jurman

March, 1st 2019

We perform here a quick statistical exploration of the heart failure dataset originally analysed in Ahmad et al., PlosONE, 2017

To this aim we use an instance of the original data **without the time feature** and **not normalized**.

First we load data (*Davide, just set your correct path if you want to run the notebook*)

```
library(readr)
path_to_data <- "~/Google Drive/chicco_survival/data/"
filename <- paste(path_to_data,"dataset_edited_without_time.csv",sep="")
dataset_edited_without_time <- as.data.frame(read_csv(filename,))
print(head(dataset_edited_without_time))
```

```
##   gender smoking diabetes blood_pressure anaemia age ejection_fraction
## 1      0       0        0              0       1  43                50
## 2      1       1        1              0       1  73                30
## 3      1       1        0              1       0  70                20
## 4      1       0        0              0       1  65                25
## 5      1       0        0              0       0  64                60
## 6      1       0        0              0       1  75                15
##   serum_sodium serum_creatinine platelets creatinine_phosphokinase
## 1          135             1.30    237000                      358
## 2          142             1.18    160000                      231
## 3          134             1.83    263358                      582
## 4          141             1.10    298000                      305
## 5          137             1.00    242000                     1610
## 6          137             1.20    127000                      246
##   death_event
## 1           0
## 2           0
## 3           1
## 4           0
## 5           0
## 6           1
```

Now we set *death_event* as a **factor**, since the statistical tests want a factor as the second element of the "~" formula.

```
dataset_edited_without_time$death_event <-
  as.factor(dataset_edited_without_time$death_event)
```

We proceed now with the computation of the four following statistical comparative tests:

- *Wilcoxon rank test* (aka Mann–Whitney U test), applied the feature *feat* versus *death_event* to detect whether we can reject the null hypothesis that the distribution of the feature *feat* for the groups of samples tdefined by *death_event* are the same; here we cannot use the *Wilcoxon signed rank test*, which needs the samples to be paired. To apply this last method, we should use repeated subsampling.

- *Kruskal Wallis test*, applied the feature *feat* versus *death_event* to detect whether we can reject the null hypothesis that the features *feat* of the samples grouped accordingly to *death_event* were selected from populations having the same distribution (it is indeed an extension of the Wilcoxon rank test);
- *Chi squared test*, applied the feature *feat* versus *death_event* to detect whether there is a significant association between the two variables; **disclaimer**: it should be applied to contingency vectors (counts), thus applying to numerical variables may by inappropriate. Note that we need to set the parameter *simulate.p.value = TRUE* to avoid warnings due to the small sample size.

plus the additional *Shapiro* test, which is applied to a *single* feature to detect whether *feat* has been extracted from a normal distribution.

All the outputs of the tests are stored on the *alltests* lists, that we print at the end of the discussion.

```
mycols <- names(dataset_edited_without_time)
mycols <- mycols[mycols!="death_event"]

alltests <- list()

alltests[["Wilcoxon_rank"]] <-
  alltests[["Kruskal"]] <-
  alltests[["Chi"]] <-
  alltests[["Shapiro"]] <- list()

for(thecol in mycols){
  alltests[["Wilcoxon_rank"]][[thecol]] <-
    wilcox.test(as.formula(paste(thecol,"death_event",sep="~")),
                data=dataset_edited_without_time)
  alltests[["Kruskal"]][[thecol]] <-
    kruskal.test(as.formula(paste(thecol,"death_event",sep="~")),
                 data=dataset_edited_without_time)
  alltests[["Chi"]][[thecol]] <-
    chisq.test(x=as.factor(dataset_edited_without_time[,thecol]),
               y=dataset_edited_without_time$death_event,
               simulate.p.value = TRUE)
  alltests[["Shapiro"]][[thecol]] <-
    shapiro.test(dataset_edited_without_time[,thecol])
}
alltests[["Shapiro"]][["death_event"]]<-
  shapiro.test(as.numeric(dataset_edited_without_time$death_event))
```

Discussion

As a rule of thumb, the validity of the tests is assessed by looking at the resulting p-values.

We start with the Shapiro test of normality,

```
dummy <- c()
for(thecol in c(mycols,"death_event")) dummy[thecol] <- alltests[["Shapiro"]][[thecol]]$p.value
print(dummy)
```

```
##                  gender                 smoking                diabetes
##            1.168593e-25            4.581881e-26            5.115471e-25
##          blood_pressure                 anaemia                     age
##            1.168593e-25            6.209525e-25            5.349669e-05
##       ejection_fraction            serum_sodium       serum_creatinine
##            7.215954e-09            9.214858e-10            5.392797e-27
```

```
##                platelets creatinine_phosphokinase              death_event
##             2.883451e-12             7.050459e-28             4.581881e-26
```

The very p-values indicate that all the variables of the dataset can be considered as extracted from a normal distribution.

We move now to the comparison of the *death_event* target with the other features of the dataset.

For the Wilcoxon rank test (that is, the Mann–Whitney U test) p-values,

```
dummy <- c()
for(thecol in mycols) dummy[thecol] <-  round(alltests[["Wilcoxon_rank"]][[thecol]]$p.value,6)
print(dummy)
```

```
##                   gender                  smoking                 diabetes
##                 0.941292                 0.828190                 0.973913
##           blood_pressure                  anaemia                      age
##                 0.171016                 0.252970                 0.000167
##        ejection_fraction             serum_sodium         serum_creatinine
##                 0.000001                 0.000293                 0.000000
##                platelets creatinine_phosphokinase
##                 0.425559                 0.684040
```

we can say that the values of each of the features *age*, *ejection_fraction*, *serum_sodium* and *serum_creatinine* are extracted from different distributions between the samples in groups *death_event=0* and *death_event=1*, while null hypothesis of same distribution for the two groups of samples cannot be ruled out for *gender*, *smoking*, *diabetes*, *blood-pressure*, *anaemia*, *platelets* and *creatinine_phosphokinase*.

For the Kruskal Wallis p-values,

```
dummy <- c()
for(thecol in mycols) dummy[thecol] <- round(alltests[["Kruskal"]][[thecol]]$p.value,6)
print(dummy)
```

```
##                   gender                  smoking                 diabetes
##                 0.940603                 0.827500                 0.973244
##           blood_pressure                  anaemia                      age
##                 0.170746                 0.252624                 0.000166
##        ejection_fraction             serum_sodium         serum_creatinine
##                 0.000001                 0.000292                 0.000000
##                platelets creatinine_phosphokinase
##                 0.425142                 0.683513
```

we can say that for *age*, *ejection_fraction*, *serum_sodium*, *serum_creatinine* the null hypothesis of same distribution for the two groups of samples *death_event=0* and *death_event=1* can be ruled out, while the same cannot be stated for *gender*, *smoking*, *diabetes*, *blood-pressure*, *anaemia*, *platelets* and *creatinine_phosphokinase*. This is consistent with the result of the Wilcoxon rank test, as expected.

For the χ^2 p-values,

```
dummy <- c()
for(thecol in mycols) dummy[thecol] <- round(alltests[["Chi"]][[thecol]]$p.value,6)
print(dummy)
```

```
##                   gender                  smoking                 diabetes
##                 1.000000                 0.893053                 1.000000
##           blood_pressure                  anaemia                      age
##                 0.203398                 0.258371                 0.002999
##        ejection_fraction             serum_sodium         serum_creatinine
##                 0.000500                 0.006997                 0.000500
```

```
##                platelets creatinine_phosphokinase
##                 0.626687                 0.373313
```

again we have the same results of the Wilcoxon rank and Kruskal Wallis tests, that is, only *age*, *ejection_fraction*, *serum_sodium*, *serum_creatinine* have a significant relation with *death_event*, while for the other features the null hypothesis of independence cannot be discarded.

Overall, we can conclude that these statistical tests indicate *age*, *ejection_fraction*, *serum_sodium*, *serum_creatinine* as the four important features of the dataset for discriminating survival as expressed by the *death_event* variable.

For completeness, we print out all the tests' outputs.

```
print(alltests)
```

```
## $Shapiro
## $Shapiro$gender
##
##  Shapiro-Wilk normality test
##
## data:  dataset_edited_without_time[, thecol]
## W = 0.60343, p-value < 2.2e-16
##
##
## $Shapiro$smoking
##
##  Shapiro-Wilk normality test
##
## data:  dataset_edited_without_time[, thecol]
## W = 0.58814, p-value < 2.2e-16
##
##
## $Shapiro$diabetes
##
##  Shapiro-Wilk normality test
##
## data:  dataset_edited_without_time[, thecol]
## W = 0.62665, p-value < 2.2e-16
##
##
## $Shapiro$blood_pressure
##
##  Shapiro-Wilk normality test
##
## data:  dataset_edited_without_time[, thecol]
## W = 0.60343, p-value < 2.2e-16
##
##
## $Shapiro$anaemia
##
##  Shapiro-Wilk normality test
##
## data:  dataset_edited_without_time[, thecol]
## W = 0.62961, p-value < 2.2e-16
##
##
```

```
## $Shapiro$age
##
##  Shapiro-Wilk normality test
##
## data:  dataset_edited_without_time[, thecol]
## W = 0.97547, p-value = 5.35e-05
##
##
## $Shapiro$ejection_fraction
##
##  Shapiro-Wilk normality test
##
## data:  dataset_edited_without_time[, thecol]
## W = 0.94732, p-value = 7.216e-09
##
##
## $Shapiro$serum_sodium
##
##  Shapiro-Wilk normality test
##
## data:  dataset_edited_without_time[, thecol]
## W = 0.93903, p-value = 9.215e-10
##
##
## $Shapiro$serum_creatinine
##
##  Shapiro-Wilk normality test
##
## data:  dataset_edited_without_time[, thecol]
## W = 0.55147, p-value < 2.2e-16
##
##
## $Shapiro$platelets
##
##  Shapiro-Wilk normality test
##
## data:  dataset_edited_without_time[, thecol]
## W = 0.91151, p-value = 2.883e-12
##
##
## $Shapiro$creatinine_phosphokinase
##
##  Shapiro-Wilk normality test
##
## data:  dataset_edited_without_time[, thecol]
## W = 0.51426, p-value < 2.2e-16
##
##
## $Shapiro$death_event
##
##  Shapiro-Wilk normality test
##
## data:  as.numeric(dataset_edited_without_time$death_event)
## W = 0.58814, p-value < 2.2e-16
```

```
##
##
##
## $Chi
## $Chi$gender
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  as.factor(dataset_edited_without_time[, thecol]) and dataset_edited_without_time$death_event
## X-squared = 0.0055707, df = NA, p-value = 1
##
##
## $Chi$smoking
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  as.factor(dataset_edited_without_time[, thecol]) and dataset_edited_without_time$death_event
## X-squared = 0.047644, df = NA, p-value = 0.8931
##
##
## $Chi$diabetes
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  as.factor(dataset_edited_without_time[, thecol]) and dataset_edited_without_time$death_event
## X-squared = 0.0011287, df = NA, p-value = 1
##
##
## $Chi$blood_pressure
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  as.factor(dataset_edited_without_time[, thecol]) and dataset_edited_without_time$death_event
## X-squared = 1.8827, df = NA, p-value = 0.2034
##
##
## $Chi$anaemia
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  as.factor(dataset_edited_without_time[, thecol]) and dataset_edited_without_time$death_event
## X-squared = 1.3131, df = NA, p-value = 0.2584
##
##
## $Chi$age
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
```

```
## data:  as.factor(dataset_edited_without_time[, thecol]) and dataset_edited_without_time$death_event
## X-squared = 69.147, df = NA, p-value = 0.002999
##
##
## $Chi$ejection_fraction
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  as.factor(dataset_edited_without_time[, thecol]) and dataset_edited_without_time$death_event
## X-squared = 65.332, df = NA, p-value = 0.0004998
##
##
## $Chi$serum_sodium
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  as.factor(dataset_edited_without_time[, thecol]) and dataset_edited_without_time$death_event
## X-squared = 45.801, df = NA, p-value = 0.006997
##
##
## $Chi$serum_creatinine
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  as.factor(dataset_edited_without_time[, thecol]) and dataset_edited_without_time$death_event
## X-squared = 92.428, df = NA, p-value = 0.0004998
##
##
## $Chi$platelets
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  as.factor(dataset_edited_without_time[, thecol]) and dataset_edited_without_time$death_event
## X-squared = 172.08, df = NA, p-value = 0.6267
##
##
## $Chi$creatinine_phosphokinase
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  as.factor(dataset_edited_without_time[, thecol]) and dataset_edited_without_time$death_event
## X-squared = 209.84, df = NA, p-value = 0.3733
##
##
##
## $Kruskal
## $Kruskal$gender
##
##  Kruskal-Wallis rank sum test
```

```
##
## data:  gender by death_event
## Kruskal-Wallis chi-squared = 0.0055521, df = 1, p-value = 0.9406
##
##
## $Kruskal$smoking
##
##  Kruskal-Wallis rank sum test
##
## data:  smoking by death_event
## Kruskal-Wallis chi-squared = 0.047485, df = 1, p-value = 0.8275
##
##
## $Kruskal$diabetes
##
##  Kruskal-Wallis rank sum test
##
## data:  diabetes by death_event
## Kruskal-Wallis chi-squared = 0.0011249, df = 1, p-value = 0.9732
##
##
## $Kruskal$blood_pressure
##
##  Kruskal-Wallis rank sum test
##
## data:  blood_pressure by death_event
## Kruskal-Wallis chi-squared = 1.8764, df = 1, p-value = 0.1707
##
##
## $Kruskal$anaemia
##
##  Kruskal-Wallis rank sum test
##
## data:  anaemia by death_event
## Kruskal-Wallis chi-squared = 1.3087, df = 1, p-value = 0.2526
##
##
## $Kruskal$age
##
##  Kruskal-Wallis rank sum test
##
## data:  age by death_event
## Kruskal-Wallis chi-squared = 14.178, df = 1, p-value = 0.0001663
##
##
## $Kruskal$ejection_fraction
##
##  Kruskal-Wallis rank sum test
##
## data:  ejection_fraction by death_event
## Kruskal-Wallis chi-squared = 24.523, df = 1, p-value = 7.341e-07
##
##
## $Kruskal$serum_sodium
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  serum_sodium by death_event
## Kruskal-Wallis chi-squared = 13.121, df = 1, p-value = 0.0002919
##
##
## $Kruskal$serum_creatinine
##
##  Kruskal-Wallis rank sum test
##
## data:  serum_creatinine by death_event
## Kruskal-Wallis chi-squared = 40.935, df = 1, p-value = 1.574e-10
##
##
## $Kruskal$platelets
##
##  Kruskal-Wallis rank sum test
##
## data:  platelets by death_event
## Kruskal-Wallis chi-squared = 0.63606, df = 1, p-value = 0.4251
##
##
## $Kruskal$creatinine_phosphokinase
##
##  Kruskal-Wallis rank sum test
##
## data:  creatinine_phosphokinase by death_event
## Kruskal-Wallis chi-squared = 0.1662, df = 1, p-value = 0.6835
##
##
##
## $Wilcoxon_rank
## $Wilcoxon_rank$gender
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  gender by death_event
## W = 9787, p-value = 0.9413
## alternative hypothesis: true location shift is not equal to 0
##
##
## $Wilcoxon_rank$smoking
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  smoking by death_event
## W = 9867, p-value = 0.8282
## alternative hypothesis: true location shift is not equal to 0
##
##
## $Wilcoxon_rank$diabetes
##
##  Wilcoxon rank sum test with continuity correction
```

```
##
## data:  diabetes by death_event
## W = 9764, p-value = 0.9739
## alternative hypothesis: true location shift is not equal to 0
##
##
## $Wilcoxon_rank$blood_pressure
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  blood_pressure by death_event
## W = 8953.5, p-value = 0.171
## alternative hypothesis: true location shift is not equal to 0
##
##
## $Wilcoxon_rank$anaemia
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  anaemia by death_event
## W = 9059, p-value = 0.253
## alternative hypothesis: true location shift is not equal to 0
##
##
## $Wilcoxon_rank$age
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  age by death_event
## W = 7121, p-value = 0.0001668
## alternative hypothesis: true location shift is not equal to 0
##
##
## $Wilcoxon_rank$ejection_fraction
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  ejection_fraction by death_event
## W = 13176, p-value = 7.368e-07
## alternative hypothesis: true location shift is not equal to 0
##
##
## $Wilcoxon_rank$serum_sodium
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  serum_sodium by death_event
## W = 12262, p-value = 0.0002928
## alternative hypothesis: true location shift is not equal to 0
##
##
## $Wilcoxon_rank$serum_creatinine
##
##  Wilcoxon rank sum test with continuity correction
```

```
##
## data:  serum_creatinine by death_event
## W = 5298, p-value = 1.581e-10
## alternative hypothesis: true location shift is not equal to 0
##
##
## $Wilcoxon_rank$platelets
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  platelets by death_event
## W = 10300, p-value = 0.4256
## alternative hypothesis: true location shift is not equal to 0
##
##
## $Wilcoxon_rank$creatinine_phosphokinase
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  creatinine_phosphokinase by death_event
## W = 9460, p-value = 0.684
## alternative hypothesis: true location shift is not equal to 0
```