COMP1814 – Coursework – Data Generation, Hypothesis Testing and Fitting Distribution

Kamelia Ikonomova
ID: 001085165
ki4866p

Hypothesis testing in R.

Comparing means and fitting distributions

200 mice received a treatment "Nutritional Supplement" during 6 months. We want to know whether the treatment has an impact on the weight of the mice.

Another 200 rats received the same treatment but results seem to differ.

To answer to this question, the weight of the mice has been measured before and after the treatment. This gives us 200 sets of values before treatment and 200 sets of values after treatment from measuring twice the weight of the **same mice**, and another 200 sets of values before treatment and 200 sets of values after treatment from measuring twice the weight of the **same rats.**

For this test you are required to create two datasets for each set of mice and rats. The data will be created from artificial data.

Create the following:

Task 1: Data Generation

    a. The weights of **mice** as "before" and "after" the treatment, coming from a normal distribution with mean = 20, and variance = 2.
For the "after" treatment add to the mean 1 unit and to the variance 0.5 units, that is mean = 21 and variance = 2.5.

```
# a. randomly generate data for weights of mice as "before" and "after"

# ND of before: m=20, variance=2 - after: m=21, variance=2.5

mice.before <- c(

  weight = round(rnorm(200,20, sd = sqrt(2)),1))

mice.after <- c(

  weight = round(rnorm(200,21,sd = sqrt(2.5)),1))

set.seed(1234)

mice.data <- data.frame(

  name = c(paste0(rep("mouse ", 200),1:200),
```

```
        paste0(rep("mouse ", 200),1:200)),
  groups = rep(c("before", "after"),each=200),
  weight = c(mice.before, mice.after))
View(mice.data)
write.csv(mice.data)
```

b. The weights of **rats** as "before" and "after" the treatment, coming from a Weibull distribution with shape = 10, and scale = 20.
For the "after" treatment remove from the shape 1 unit and add to the scale 1 unit, that is shape = 9 and scale = 21.   [5 marks]

```
# b. randomly generate data for weights of rats as "before" and "after"
# WD of before: shape = 10, scale = 20 - after: shape = 9, scale = 21
rats.before <- c(
  weight = round(rweibull(200,10,20),1))
rats.after <- c(
  weight = round(rweibull(200,9,21),1))
# data frame
set.seed(1234)
rats.data <- data.frame(
  name = c(paste0(rep("rat ", 200),1:200),
        paste0(rep("rat ", 200),1:200)),
  groups = rep(c("before", "after"),each=200),
  weight = c(rats.before, rats.after))
View(rats.data)
write.csv(rats.data)
```

c. Using the function 'qplot' with 'geom = density' (from your lecture notes) compare for each of your data sets mice(before, after) and rats(before, after)   [10 marks]

```
# c. graphs with qplot() and geom = "density"
qplot(data = mice.data, geom = "density",
    xlab = "Groups of Mice", ylab = "Weights",
    weight, color = groups, linetype = groups)
qplot(data = rats.data, geom = "density",
```

xlab = "Groups of Rats", ylab = "Weights",

weight, color = groups, linetype = groups)

d. Perform the same operation using 'geom = boxplot'. [5 marks]

```
# d. graphs with qplot() and geom = "boxplot"
library(tidyverse)
qplot(data = mice.data, geom = "boxplot",
    x = groups, y = weight,
    xlab = "Groups of Mice", ylab = "Weights",
    fill = groups,
    # order = c("before", "after"),
    # palette = c("fadadd", "faedfa")
)
qplot(data = rats.data, geom = "boxplot",
    x = groups, y = weight,
    xlab = "Groups of Rats", ylab = "Weights",
    fill = groups
)
```

Task 2: Appropriateness for Hypothesis t-testing

a. For your **mice** data set (combined "before" + "after") examine whether the data passes normality qualitatively (QQ plot) and quantitatively (Shapiro-Wilk test). [5 marks]

```
# a. normality test for Mice Treatment data set
library("ggpubr")
mice.qqplot <- ggqqplot(mice.data, x = "weight", color = "groups")
mice.qqplot + facet_grid(. ~ groups)
mice.qqplot +  stat_qq() + stat_qq_line()
shapiro.test(with(mice.data,
        weight[groups == "before"] - weight[groups == "after"]))
```

b. For your **rats** data set (combined "before" + "after") examine whether the data passes normality qualitatively (QQ plot) and quantitatively (Shapiro-Wilk test). [5 marks]

b. normality test for Rats Treatment data set

rats.qqplot <- ggqqplot(rats.data, x = "weight", color = "groups")

rats.qqplot + facet_grid(. ~ groups)

rats.qqplot + stat_qq() + stat_qq_line()

shapiro.test(with(rats.data,

       weight[groups == "before"] - weight[groups == "after"]))

c.   Explain the output of your analysis for each of the dataset and discuss the appropriate test
to test your hypothesis.   [15 marks]

Task 3: Hypothesis testing

a.   For the normal data set (**mice**) examine perform a paired t-test and explain your
findings. You need to extract and comment on all output of the t-test:
     a.   T-test statistic
     b.   Degrees of freedom
     c.   P-value
     d.   Confidence Interval
     e.   Sample estimates

```
# shapiro-wilk normality test for mice.data

mice.stest <- with(mice.data,

        weight[groups == "before"] - weight[groups == "after"])

shapiro.test(mice.stest)

# CMT: p-value = 0.1314 > 0.05 --> Normal


# paired t-test for mice.data

pd <- PairedData::paired(mice.before, mice.after)

pd

# method 1

mice.ttest.1 <- t.test(mice.before, mice.after, paired = TRUE)

mice.ttest.1

# method 2

mice.ttest.2 <- t.test(weight ~ groups, data = mice.data, paired = TRUE)

mice.ttest.2
```

b. For the **rats** dataset perform a non-parametric -test and comment on your finding.

[10 marks]

```
# non-parametric test for rats.data

# QQ-plot

rats.qqplot <- ggplot(data = rats.data,

        aes(sample = weight))

rats.qqplot + stat_qq() + stat_qq_line()

rats.qqplot + stat_qq() + stat_qq_line() + facet_grid(. ~ groups)

# Data: NORMAL, sample size=200: LARGE

# CMT: The points are quite close to lie on a straight line


# Wilcoxon rank-sum test

rats.wtest <- wilcox.test(weight ~ groups, data = rats.data, conf.int = TRUE)
```

rats.wtest

with(rats.data, wilcox.test(x = weight[groups == "before"],

                 y = weight[groups == "after"]))

Task 4: Fitting distributions

a. Lastly, for the **rats** datasets use the function 'fitdist' (from the 'fitdistrplus' R package) and examine the best-fit distribution (even if we know the TRUE distribution). Fit a Weibull, a lognormal and a Gamma distributions and discuss your findings making use of the package comparison tool for a Density, CDF, QQ, and PP.   [20 marks]

```
PairedData::paired(rats.before, rats.after)

library(fitdistrplus)

rats.result <- rats.after - rats.before

descdist(rats.before, discrete = FALSE, boot = 20)

descdist(rats.after, discrete = FALSE, boot = 20)

descdist(rats.result, discrete = FALSE, boot = 20)

plotdist(rats.result, histo = TRUE, demp = TRUE)


# best-fit distribution for rats.data

rats.result <- c(rats.before, rats.after)

# weibull

w.result <- fitdist(rats.result, "weibull")

summary(w.result)

# gamma

g.rats <- fitdist(rats.result, "gamma")

summary(g.rats)

# log-normal

ln.rats <- fitdist(rats.result, "lnorm")

summary(ln.rats)


# packages comparison tools

par(mfrow=c(2, 2))

plot.legend <- c("Weibull", "Gamma", "Log-normal")

# density. CMT: Weibull

denscomp(list(w.rats, g.rats, ln.rats), legendtext = plot.legend)
```

```r
# cdf. CMT: Weibull
cdfcomp(list(w.rats, g.rats, ln.rats), legendtext = plot.legend)
# qq. CMT: Weibull
qqcomp(list(w.rats, g.rats, ln.rats), legendtext = plot.legend)


# plot each candidate to find the best fit model
# CMT: CDF & QQ?
denscomp(w.rats)
cdfcomp(w.rats)
qqcomp(w.rats)
```

--------- End of Coursework --