---

title: "Lab 9b - Solution code"

output:

html\_document:

df\_print: paged

---

```{r setup, include=FALSE}

knitr::opts\_chunk$set(echo = TRUE)

```

```{r warning=F}

library(tidyverse)

```

# One-way ANOVA

```{r}

algorithm <- read.csv("algorithm.csv", header=T, stringsAsFactors = T)

```

## Dot plot for each algorithm

### Algorithm A

```{r}

p <- ggplot(algorithm, aes(x= Algorithm.A)) +

geom\_dotplot(binwidth=0.5, col="red", fill="red")

p <- p+ theme\_classic() + labs (x="Algorithm A")

p <- p + theme(text = element\_text(size = 20))

p

```

### Algorithm B

```{r}

p <- ggplot(algorithm, aes(x= Algorithm.B)) +

geom\_dotplot(binwidth=0.5, col="red", fill="red")

p <- p+ theme\_classic() + labs (x="Algorithm B")

p <- p + theme(text = element\_text(size = 20))

p

```

### Algorithm C

```{r}

p <- ggplot(algorithm, aes(x= Algorithm.C)) +

geom\_dotplot(binwidth=0.5, col="red", fill="red")

p <- p+ theme\_classic() + labs (x="Algorithm C")

p <- p + theme(text = element\_text(size = 20))

p

```

### Algorithm D

```{r}

p <- ggplot(algorithm, aes(x= Algorithm.D)) +

geom\_dotplot(binwidth=0.5, col="red", fill="red")

p <- p+ theme\_classic() + labs (x="Algorithm D")

p <- p + theme(text = element\_text(size = 20))

p

```

## Data preparation

Converting the dataset to the right format - 2 columns, one for the algorithm and one for the time.

```{r}

# creating mini data frames, one for each algorithm,

# each one containscone column with the name of the algorithm and one column with the time

algoA <- data.frame(time = algorithm$Algorithm.A, algorithm= "A")

algoB <- data.frame(time = algorithm$Algorithm.B, algorithm= "B")

algoC <- data.frame(time = algorithm$Algorithm.C, algorithm= "C")

algoD <- data.frame(time = algorithm$Algorithm.D, algorithm= "D")

# putting the 4 mini data frames together

dataAlgo <- rbind(algoA, algoB, algoC,algoD)

```

## One-way ANOVA test

Applying the one-way ANOVA test.

H0: There is no difference in mean values i.e. µA = µB = µC = µD.

H1: at least two means are different.

```{r}

anova <- aov( time ~ algorithm, data = dataAlgo)

summary(anova)

```

The test indicates that there is no significant difference in means.

### Checking the validity of the test

The residuals need to be checked. The test is only valid if the residuals are normally distributed.

#### Q-Q plot

```{r}

# first construct a data frame with the rediduals. Then use it in the plot

anovaFrame <- data.frame(residuals = anova$residuals )

p <- ggplot(anovaFrame, aes(sample = residuals))

p <- p + stat\_qq(size=2) + stat\_qq\_line( alpha = 0.7, color = "red", linetype = "dashed")

p <- p + theme\_classic()

p <- p + theme(text = element\_text(size = 20))

p

```

The points are close to the theoretical line. It is reasonable to assume the distribution is normal.

#### Shapiro-Wilk test of normality

```{r}

shapiro.test(anova$residuals)

```

The p-value (0.4428) is > 0.05 so it is reasonable to assume the distribution is normal.

## Sample sizes

### Sample sizes for t-test

To get the number of samples required for a power of 0.8, with a difference of 8 units, an estimated standard deviation of 10 and a significance level of 0.05.

```{r}

a <- power.t.test(power=0.8,delta=8,sd=10,sig.level=0.05,type="two.sample")

a

```

# Exercises {-}

# Exercise 1

```{r}

methods <- read.csv("methods.csv", header=T,stringsAsFactors=T)

```

```{r}

# creating mini data frames, one for each algorithm,

# each one contains cone column with the number of documents and another with the method.

methodA <- data.frame(docNumber = methods$Method.A, method= "A")

methodB <- data.frame(docNumber = methods$Method.B, method= "B")

methodC <- data.frame(docNumber = methods$Method.C, method= "C")

# putting the 4 mini data frames together

dataMethod <- rbind(methodA, methodB, methodC)

```

## One-way ANOVA test

Applying the one-way ANOVA test.

H0: There is no difference in mean values i.e. µA = µB = µC .

H1: at least two means are different.

```{r}

anova <- aov( docNumber ~ method, data = dataMethod)

summary(anova)

```

p-value > 0.05. Do not reject . There is not enough evidence of any differences between the mean numbers of relevant documents retrieved by the three methods.

The valididy of the test needs to be checked.

### Checking whether the residuals are normally distributed

Checking normal distribution of residuals using a Q-Q plot.

```{r}

anovaFrame <- data.frame(residuals = anova$residuals )

p <- ggplot(anovaFrame, aes(sample = residuals))

p <- p + stat\_qq()

p <- p + stat\_qq\_line( )

p

```

The points are close to the line suggesting it is reasonable to assume that the distribution is normal.

Checking normal distribution of residuals using the Shapiro-Wilk test

```{r}

shapiro.test(anovaFrame$residuals)

```

A p-value > 0.05 indicates that it is reasonable to assume that the residuals are normally distributed.

# Exercise 2

```{r}

tool <- read.csv("tool.csv", header=T,stringsAsFactors=T)

```

```{r}

# creating mini data frames, one for each algorithm,

# each one contains cone column with the number of documents and another with the method.

toolA <- data.frame(percentage = tool$Tool.A, tool = "A")

toolB <- data.frame(percentage = tool$Tool.B, tool= "B")

toolC <- data.frame(percentage = tool$Tool.C, tool= "C")

toolD <- data.frame(percentage = tool$Tool.D, tool= "D")

# putting the 4 mini data frames together

dataTool <- rbind(toolA, toolB, toolC, toolD)

```

H0: There is no difference in mean values i.e. µA = µB = µC = µD.

H1: at least two means are different.

```{r}

anova <- aov( percentage ~ tool, data = dataTool)

summary(anova)

```

The p-value (5.78e-06) indicates that there are differences in means.

The valididy of the test needs to be checked.

### Checking whether the residuals are normally distributed

Checking normal distribution of residuals using a Q-Q plot.

```{r}

anovaFrame <- data.frame(residuals = anova$residuals )

p <- ggplot(anovaFrame, aes(sample = residuals))

p <- p + stat\_qq()

p <- p + stat\_qq\_line( )

p

```

The points are close to the line, although one of the points (lower left) may be a cause for concern.

Checking normal distribution of residuals using the Shapiro-Wilk test

```{r}

shapiro.test(anovaFrame$residuals)

```

A p-value > 0.05 indicates that it is reasonable to assume that the residuals are normally distributed.

# Exercise 3

Sample sizes for different powers with 5% significance level and a standard deviation of 10 and a difference of 8.

```{r}

a <- power.t.test(power=0.8,delta=8,sd=10,sig.level=0.05,type="two.sample")

b<- power.t.test(power=0.85,delta=8,sd=10,sig.level=0.05,type="two.sample")

c <- power.t.test(power=0.9,delta=8,sd=10,sig.level=0.05,type="two.sample")

d <- power.t.test(power=0.95,delta=8,sd=10,sig.level=0.05,type="two.sample")

e <- power.t.test(power=0.99,delta=8,sd=10,sig.level=0.05,type="two.sample")

f <- power.t.test(power=0.999,delta=8,sd=10,sig.level=0.05,type="two.sample")

## Displaying the results

cat( "",

"n = ", ceiling(a$n), "units = ", a$delta, "power = ", a$power, "significance level = ", a$sig.level, "\n",

"n = ", ceiling(b$n), "units = ", b$delta, "power = ", b$power, "significance level = ", b$sig.level, "\n",

"n = ", ceiling(c$n), "units = ", c$delta, "power = ", c$power, "significance level = ", c$sig.level, "\n",

"n = ", ceiling(d$n), "units = ", d$delta, "power = ", d$power, "significance level = ", d$sig.level, "\n",

"n = ", ceiling(e$n), "units = ", e$delta, "power = ", e$power, "significance level = ", e$sig.level, "\n",

"n = ", ceiling(f$n), "units = ", f$delta, "power = ", f$power, "significance level = ", f$sig.level, sep = "\t")

```

Displaying the results graphically.

```{r}

powerVal <- c(a$power,b$power, c$power, d$power, e$power, f$power)

sampleSize <- c(a$n, b$n, c$n, d$n, e$n, f$n)

data <- data.frame(power = powerVal, sampleSeize = sampleSize)

p <- ggplot(data, aes(x=power, y=sampleSize))

p <- p+ geom\_point(colour="red", size=3) + theme\_classic() +labs(x= "Power", y= "Sample size (n)")

p <- p + theme(text = element\_text(size = 20))

p

```

## Exercise 4

### Section a

```{r}

power.t.test(power=0.7,delta=10,sd=25,sig.level=0.05,type="two.sample", alternative ="two.sided")

power.t.test(power=0.8,delta=10,sd=25,sig.level=0.05,type="two.sample", alternative ="two.sided")

power.t.test(power=0.9,delta=10,sd=25,sig.level=0.05,type="two.sample", alternative ="two.sided")

power.t.test(power=0.95,delta=10,sd=25,sig.level=0.05,type="two.sample", alternative ="two.sided")

power.t.test(power=0.7,delta=15,sd=25,sig.level=0.05,type="two.sample", alternative ="two.sided")

power.t.test(power=0.8,delta=15,sd=25,sig.level=0.05,type="two.sample", alternative ="two.sided")

power.t.test(power=0.9,delta=15,sd=25,sig.level=0.05,type="two.sample", alternative ="two.sided")

power.t.test(power=0.95,delta=15,sd=25,sig.level=0.05,type="two.sample", alternative ="two.sided")

power.t.test(power=0.7,delta=20,sd=25,sig.level=0.05,type="two.sample", alternative ="two.sided")

power.t.test(power=0.8,delta=20,sd=25,sig.level=0.05,type="two.sample", alternative ="two.sided")

power.t.test(power=0.9,delta=20,sd=25,sig.level=0.05,type="two.sample", alternative ="two.sided")

power.t.test(power=0.95,delta=20,sd=25,sig.level=0.05,type="two.sample", alternative ="two.sided")

```

### Section b

```{r}

power.t.test(power=0.7,delta=10,sd=25,sig.level=0.01,type="two.sample", alternative ="two.sided")

power.t.test(power=0.8,delta=10,sd=25,sig.level=0.01,type="two.sample", alternative ="two.sided")

power.t.test(power=0.9,delta=10,sd=25,sig.level=0.01,type="two.sample", alternative ="two.sided")

power.t.test(power=0.95,delta=10,sd=25,sig.level=0.01,type="two.sample", alternative ="two.sided")

power.t.test(power=0.7,delta=15,sd=25,sig.level=0.01,type="two.sample", alternative ="two.sided")

power.t.test(power=0.8,delta=15,sd=25,sig.level=0.01,type="two.sample", alternative ="two.sided")

power.t.test(power=0.9,delta=15,sd=25,sig.level=0.01,type="two.sample", alternative ="two.sided")

power.t.test(power=0.95,delta=15,sd=25,sig.level=0.01,type="two.sample", alternative ="two.sided")

power.t.test(power=0.7,delta=20,sd=25,sig.level=0.01,type="two.sample", alternative ="two.sided")

power.t.test(power=0.8,delta=20,sd=25,sig.level=0.01,type="two.sample", alternative ="two.sided")

power.t.test(power=0.9,delta=20,sd=25,sig.level=0.01,type="two.sample", alternative ="two.sided")

power.t.test(power=0.95,delta=20,sd=25,sig.level=0.01,type="two.sample", alternative ="two.sided")

```

### Section c

```{r}

power.t.test(power=0.7,delta=10,sd=25,sig.level=0.01,type="two.sample", alternative ="one.sided")

power.t.test(power=0.8,delta=10,sd=25,sig.level=0.01,type="two.sample", alternative ="one.sided")

power.t.test(power=0.9,delta=10,sd=25,sig.level=0.01,type="two.sample", alternative ="one.sided")

power.t.test(power=0.95,delta=10,sd=25,sig.level=0.01,type="two.sample", alternative ="one.sided")

power.t.test(power=0.7,delta=15,sd=25,sig.level=0.01,type="two.sample", alternative ="one.sided")

power.t.test(power=0.8,delta=15,sd=25,sig.level=0.01,type="two.sample", alternative ="one.sided")

power.t.test(power=0.9,delta=15,sd=25,sig.level=0.01,type="two.sample", alternative ="one.sided")

power.t.test(power=0.95,delta=15,sd=25,sig.level=0.01,type="two.sample", alternative ="one.sided")

power.t.test(power=0.7,delta=20,sd=25,sig.level=0.01,type="two.sample", alternative ="one.sided")

power.t.test(power=0.8,delta=20,sd=25,sig.level=0.01,type="two.sample", alternative ="one.sided")

power.t.test(power=0.9,delta=20,sd=25,sig.level=0.01,type="two.sample", alternative ="one.sided")

power.t.test(power=0.95,delta=20,sd=25,sig.level=0.01,type="two.sample", alternative ="one.sided")

```