---

title: "Lab 9 - Solutions"

output:

html\_document:

df\_print: paged

---

```{r setup, include=FALSE}

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

library(here)

setwd(here::here())

```

```{r warning=F}

library(ggplot2)

```

We want to generate sample data for the coin tossing example. We could use

```{r}

## coin tossing example. 10 throws. Count n. heads, repeat 20 times.

## Assume normal distribution with mean 5 and SD mean/3.5

set.seed(123) # ensure random sampling will give same values every time with this seed.

minVal <- 0 # set min value

maxVal <- 10 # set max value

mn <- (maxVal - minVal)/2 # average

# Generate 20 numbers (mostly) from min to max

x <- rnorm(20, mean = mn, sd = mn/3.5)

# Change the out-of-bounds generated values to min o max value

x <- pmax(minVal, x)

x <- pmin(maxVal, x)

# round all values as coin tossing produces integers (number of heads out of 10 throws)

x <- round(x)

x

```

## Creating the dataset {-}

The dataset includes and instance number and a time value. There are 10 instances.

```{r}

instance =c(1:10)

time = c(386.1, 386.1, 389.6, 406.1, 407.5, 424.0, 427.5, 429.5, 443.3, 457.8 )

swTimes <- data.frame(instance , time)

```

# Testing for normality {-}

## Q-Q Plot {-}

```{r}

p <- ggplot(swTimes, aes(sample = time))

p <- p + stat\_qq()

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

p

```

Points scatered close to ideal line so is is reasonable to assume that the distribution is normal.

## Shapiro Wilk test {-}

```{r}

shapiro.test(swTimes$time)

```

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

## Dot plot {-}

```{r}

p <- ggplot(swTimes, aes(x= time))

p <- p+ geom\_dotplot( binwidth=1 )

p <- p + labs( x="",y="")

p <- p + xlim(370,460)

p

```

The dots are scattered around an area with no pattern. It is not unreasonable to assume that the distribution is normal.

# Confidence intervals and margin of error

# Obtaining the t-value {-}

```{r}

# number of instances of data

n <- length(swTimes$time)

# Confidence and significance

confidence <- 0.95

significance <- 1- confidence

# two-tailed so half

halfSignificance <- significance/2

a <- 1 - halfSignificance

# degrees of freedom

dfTimes <- n - 1

# finding t-value

# function qt returns required value for t

t <- qt(a, dfTimes)

t

```

So the t-value is 2.262157 (at 5% significance level).

# Calculating the margin of error {-}

```{r}

#standard deviation for times data

sdTimes <- sd(swTimes$time)

# margin of error

error <- t\*sdTimes/sqrt(n)

error

```

The margin of error is 17.72528.

# Calculating the Confidence interval given margin of error and mean {-}

```{r}

# mean value

meanTimes <- mean(swTimes$time)

meanTimes

# Lower and upper limits

lowerlimit <- meanTimes - error

upperlimit <- meanTimes + error

# confidence interval

CI <- paste("Confidence interval is (", lowerlimit, ", ", upperlimit, ")")

CI

```

# One sample t-test {-}

Ho: NULL hypothesis - there is no difference between the mean and value 400 for the swTimes dataset.

H1: Alternative hypothesis - the mean differs from 400.

```{r}

t.test(swTimes$time, alternative = "two.sided", paired=F, mu= 400, conf.level=0.9)

```

The p-value > 0.05 so do not reject NULL hypothesis.

# Matched pairs tests {-}

Two search engines are used to identify relevant documents for 12 different searches.

Note that the data has been created in a csv file.

```{r}

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

```

## Testing for normality {-}

### Q-Q Plot {-}

```{r}

p <- ggplot(engines, aes(sample = Difference))

p <- p + stat\_qq()

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

p

```

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

### Shapiro-Wilk test of normality {-}

```{r}

shapiro.test(engines$Difference)

```

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

## Paired t-test {-}

Ho: NULL hypothesis - there is no difference between the two search engine means.

H1: Alternative hypothesis – there is a difference between the two search engine means.

```{r}

t.test(x=engines$Search.engine.A, y=engines$Search.engine.B , alternative = "two.sided", paired=T, mu= 0, conf.level=0.95)

```

The p-value (0.01834) is less than 0.05 so reject the NULL hypothesis and accept the alternative hypothesis. There is a significant difference in the search engines’ performances.

At 95% confidence level the mean number of relevant documents retrieved by search engine B is between 1.175 and 10.323 higher than the search engine A mean.

# Tests for 2 independent samples comparison of 2 population means {-}

Data from starting salaries (in 1000 £ / year) for 2 different degree graduates is available. There is a suggestion that computing graduates (Job 2 below) earn more than business graduates (Job 1 below).

# Exercise 1

Data can be read from a file or created from scratch. Both options are available below. Comment the option not in use.

```{r}

Job.1 <- c(17.2,18.3,17.4,19.6,19,14.6,18.6,16.5,15.4,20.2,18,16.5,21.7,19.4,NA)

Job.2 <- c(17.2,19.9,21.6,21.1,19.1,18.4,22.3,19.9,21.3,18.6,19.4,16.6,20.7,23.8,20.5)

jobs <- data.frame(Job.1, Job.2)

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

```

Do computing graduates have the same starting salary than business graduates?

## Normality tests {-}

Note that the checks below are for the individual distributions.

### Q-Q Plot {-}

For business graduates:

```{r, warning=F}

p <- ggplot(jobs, aes(sample = Job.1))

p <- p + stat\_qq()

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

p <- p+ theme\_classic()

p

```

For computing graduates:

```{r, warning=F}

p <- ggplot(jobs, aes(sample = Job.2))

p <- p + stat\_qq()

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

p

```

The points look close to the ideal line. It is reasonable to assume that the distribution is normal.

### Shapiro-Wilk normality tests {-}

For business graduates:

```{r}

shapiro.test(jobs$Job.1)

```

For computing graduates:

```{r}

shapiro.test(jobs$Job.2)

```

The p-values are greater than 0.05, so the normality test is passed.

## Pooled 2-sample t-test {-}

Ho: NULL hypothesis - there is no difference between the graduate jobs starting salary means.

H1: Alternative hypothesis – there is a difference between the graduate jobs starting salary means.

Set var.equal to true to assume equal variance (pooled).

```{r}

t.test(jobs$Job.1, y=jobs$Job.2 , alternative = "two.sided", paired=F, var.equal= T, mu= 0, conf.level=0.99)

```

The p value is smaller than 0.01 so the NULL hypothesis is rejected in favour of the alternative hypothesis – there is a significant difference in means.

# Exercises {-}

# Exercise 2

## a. Dataset creation {-}

```{r}

classmateTimes <- data.frame(time = c(95.2, 103.5, 105.2, 105.2, 103.4, 102.8, 101.8, 97.9, 107.9 , 105.9, 110.5, 100.2))

```

## b. Calculation of confidence intervals {-}

The calculations appear below. Note that a t-testcould have been used to the confidence intervals.

```{r}

# number of instances of data

n <- length(classmateTimes$time)

# Confidence and significance values (at 95% and 99%)

confidence1 <- 0.95

confidence2 <- 0.99

significance1 <- 1- confidence1

significance2 <- 1- confidence2

halfSignificance1 <- significance1/2

a <- 1 - halfSignificance1

halfSignificance2 <- significance2/2

b <- 1 - halfSignificance2

# degrees of freedom

dfTimes <- n - 1

# finding t-values

t1 <- qt(a, dfTimes)

t2 <- qt(b, dfTimes)

#standard deviation

sdTimes <- sd(classmateTimes$time)

# margin of error at 95% confidence

error1 <- t1\*sdTimes/sqrt(n)

# margin of error at 99% confidence

error2 <- t2\*sdTimes/sqrt(n)

# mean value

meanTimes <- mean(classmateTimes$time)

# Lower and upper limits at 95% confidence

lowerlimit1 <- meanTimes - error1

upperlimit1 <- meanTimes + error1

# Lower and upper limits at 99% confidence

lowerlimit2 <- meanTimes - error2

upperlimit2 <- meanTimes + error2

# confidence intervals

CI95 <- paste("Confidence interval at 95% confidence is (", lowerlimit1, ", ", upperlimit1, ")")

CI99 <- paste("Confidence interval at 99% confidence is (", lowerlimit2, ", ", upperlimit2, ")")

# Outputting the confidence intervals

CI95

CI99

```

The interval at 99% confidence has a larger range. Needing more of the observations to fall within the interval requires a wider interval.

## c. Normality tests {-}

### Dot plot {-}

```{r}

p <- ggplot(classmateTimes, aes(x= time))

p <- p+ geom\_dotplot( binwidth=1 )

p <- p + labs( x="",y="")

#p <- p + xlim(370,460)

p

```

The distribution of values looks reasonable, but normality tests need to be carried out.

### Q-Q Plot {-}

```{r}

p <- ggplot(classmateTimes, aes(sample = time))

p <- p + stat\_qq()

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

p

```

The points are scatered somehow close to ideal line, although the very top and very bottom points are not quite so close.

### Shapiro Wilk test {-}

```{r}

shapiro.test(classmateTimes$time)

```

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

## d. Testing if the mean time is 100 {-}

One sample t-test used.

Ho: NULL hypothesis - there is no difference between the mean and value 100.

H1: Alternative hypothesis - the mean differs from 100 seconds.

First test at 95% confidence level.

```{r}

t.test(classmateTimes$time, alternative = "two.sided", paired=F, mu= 100, conf.level=0.95)

```

The p-value is 0.02, which is less than 0.05, so the NULL hypothesis is rejected in favour of the alternative hypothesis. There is evidence that the mean value is greater than 100, as the mean value shown is 103.2917.

Next a test at 99% confidence level is conducted. Note that there is no need to conduct the test below as the p-value will be identical to the one obtained above. The only difference will be that the confidence interval will be wider. AS this exercise does not require looking at confidence intervals, the code below is not needed.

```{r}

t.test(classmateTimes$time, alternative = "two.sided", paired=F, mu= 100, conf.level=0.99)

```

The p-value > 0.01 so the NULL hypothesis cannot be rejected. The evidence is not sufficiently strong to reject the NULL hypothesis.

Note how at 99% confidence the confidence interval includes 100.

# Exercise 3

## a. Dataset creation {-}

```{r}

fellowTimes <- data.frame(time=c(117.0, 116.3,117.9, 120.0, 125.5, 144.8, 120.6, 120.4, 137.6, 121.7, 123.2, 133.8))

```

## b. Calculation of confidence intervals {-}

The calculations appear below.

```{r}

# number of instances of data

n <- length(fellowTimes$time)

# Confidence and significance values (at 95% and 99%)

confidence1 <- 0.95

confidence2 <- 0.99

significance1 <- 1- confidence1

significance2 <- 1- confidence2

halfSignificance1 <- significance1/2

a <- 1 - halfSignificance1

halfSignificance2 <- significance2/2

b <- 1 - halfSignificance2

# degrees of freedom

dfTimes <- n - 1

# finding t-values

t1 <- qt(a, dfTimes)

t2 <- qt(b, dfTimes)

#standard deviation

sdTimes <- sd(fellowTimes$time)

# margin of error at 95% confidence

error1 <- t1\*sdTimes/sqrt(n)

# margin of error at 99% confidence

error2 <- t2\*sdTimes/sqrt(n)

# mean value

meanTimes <- mean(fellowTimes$time)

# Lower and upper limits at 95% confidence

lowerlimit1 <- meanTimes - error1

upperlimit1 <- meanTimes + error1

# Lower and upper limits at 99% confidence

lowerlimit2 <- meanTimes - error2

upperlimit2 <- meanTimes + error2

# confidence interval

CI95 <- paste("Confidence interval at 95% confidence is (", lowerlimit1, ", ", upperlimit1, ")")

CI99 <- paste("Confidence interval at 99% confidence is (", lowerlimit2, ", ", upperlimit2, ")")

# outputting values

CI95

CI99

```

Note that the above confidence intervals are for the mean.

## c. Normality tests {-}

### Dot plot {-}

```{r}

p <- ggplot(fellowTimes, aes(x= time))

p <- p+ geom\_dotplot( binwidth=1 )

p <- p + labs( x="",y="")

p

```

The distribution of values does not look normal. The distribution looks positively skewed.

### Q-Q Plot {-}

```{r}

p <- ggplot(fellowTimes, aes(sample = time))

p <- p + stat\_qq()

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

p

```

The points are not that close to ideal line.

### Shapiro Wilk test {-}

```{r}

shapiro.test(fellowTimes$time)

```

The p-value is < 0.05 so it is not reasonable to assume that the distribution is normal.

As it is reasonable to assume that the distribution is not normal, non-parametric tests must be used - i.e. tests based on the median value.

# Exercise 4

## Dataset creation {-}

```{r}

Week <- c(1:8)

Aberdeen <- c(7820, 6400, 8150, 7360, 7210, 8460, 6770, 7900)

Dundee <- c(6770, 6740, 7290, 7020, 7280, 7950, 6490, 7330)

profits <- data.frame(Week, Aberdeen, Dundee)

```

## b. Normality tests {-}

### Dot plots {-}

```{r}

p <- ggplot(profits, aes(x= Aberdeen))

p <- p+ geom\_dotplot( binwidth=50)

p <- p + labs( x="",y="")

p

```

```{r}

p <- ggplot(profits, aes(x= Dundee))

p <- p+ geom\_dotplot( binwidth=50)

p <- p + labs( x="",y="")

p

```

Nothing out of the ordinary observed. Values are scattered.

### Q-Q Plots {-}

```{r}

p <- ggplot(profits, aes(sample = Aberdeen))

p <- p + stat\_qq()

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

p

```

```{r}

p <- ggplot(profits, aes(sample = Dundee))

p <- p + stat\_qq()

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

p

```

The points are close to ideal line.

### Shapiro Wilk tests {-}

```{r}

shapiro.test(profits$Aberdeen)

```

```{r}

shapiro.test(profits$Dundee)

```

The p-values are > 0.05 so the it is reasonable to assume that the distributions are normal.

### c. Test difference between the profits in Aberdeen and Dundee {-}

Paired t-test

Ho: NULL hypothesis - there is no difference between the two profit means.

H1: Alternative hypothesis – there is a difference between the two profit means.

```{r}

t.test(x=profits$Aberdeen, y=profits$Dundee , alternative = "two.sided", paired=T, mu= 0, conf.level=0.95)

```

The p-value is less than 0.05 so reject the NULL hypothesis and accept the alternative hypothesis. There is a significant difference in the mean weekly profit. The mean for Aberdeen is higher (the mean of the differences is 400).

At 95% confidence level the mean weekly profits are different. Aberdeen's are higher.

So let's conduct a normality test on the difference in profits to ensure the test is valid.

```{r}

profits$difference <- profits$Aberdeen - profits$Dundee

shapiro.test(profits$difference)

```

The p-value > 0.05 so it is reasonable to assume that the difference in profits is normally distributed. The t-test can be used.

# Exercise 5

## a. Dataset creation {-}

```{r}

Recall.A <- c(71.5, 54.5, 70.2, 65.0, 83.9, 47.9, 49.1, 63.4, 64.2, 57.3, 61.7, 69.6, 75.8)

Recall.B <- c(59.9, 94.4, 67.0, 73.8, 83.3, 76.9, 70.4, 90.7, 75.8, 84.4, 74.7, 74.2, 64.5)

recallIndexing <-data.frame(Recall.A, Recall.B)

```

## b. Normality tests {-}

For Recall.A

```{r}

shapiro.test(recallIndexing$Recall.A)

```

The p-value is 0.9645 > 0.05, so it is reasonable to assume a normal distribution.

For Recall.B

```{r}

shapiro.test(recallIndexing$Recall.B)

```

The p-value is 0.9019 > 0.05, so it is reasonable to assume a normal distribution.

## c. Testing for difference in means {-}

Pooled 2-sample t-test

Ho: NULL hypothesis - there is no difference between the recall values for the 2 tools.

H1: Alternative hypothesis – there is a difference between recall values for the 2 tools.

Set var.equal to true to assume equal variance (pooled)

```{r}

t.test(x=recallIndexing$Recall.A, y=recallIndexing$Recall.B , alternative = "two.sided", paired=F, var.equal= T, mu= 0, conf.level=0.99)

```

The p- value is very small, with a value of 0.006063 < 0.05, so reject the NULL hypothesis. There is strong evidence for the mean values of the 2 tools being different, with the mean value for Tool ~B being higher.