**Data Analytics (CMP330)**

# **Practical 5 – Inferential Statistics and Hypothesis Testing with R**

1. **One-sample t-test**

The one-sample t-test is used to compare the mean of one sample to a known standard (or theoretical/hypothetical) mean (μ). Generally, the theoretical mean comes from a previous experiment or from an experiment where you have control and treatment conditions.

To perform one-sample t-test, the R function t.test() can be used as follow

t.test(x, mu = 0, alternative = "two.sided")

Try the following code,

# load the data

# use the R built-in dataset sleep (Data which show the effect of two soporific drugs)

**data <- sleep**

#check your data

**head(data)**

# Groups 1 and 2 refer to the same individuals, but with different drugs

# We focus on group 2 and create a vector with the extra sleep recorded in this case

**group\_2 <- sleep$extra[sleep$group==2]**

# Statistical summary of extra sleep

**summary(group\_2)**

# visualise the extra sleep by using box plot

**boxplot(group\_2, main='Extra Sleep for Group 2')**

Preliminary test to check one-sample t-test assumptions

1. Is this a large sample? - No, because n < 30.

2. Since the sample size is not large enough (less than 30, central limit theorem), we need to check whether the data follow a normal distribution.

# use Shapiro-Wilk test to check normality

**shapiro.test(group\_2)**

P-value is greater than the significance level 0.05 implying that the distribution of the data are not significantly different from a normal distribution. In other words, we can assume normality.

We want to know, if the average amount of extra sleep differs from zero, i.e. zero extra sleep is the null hypothesis (two-tailed test)?

# One-sample t-test

**x <- t.test(group\_2, mu = 0)**

**x**

**# print the p-value**

**x$p.value**

**# print the confidence interval**

**x$conf.int**

P-value is less than the significance level 0.05 (by default), we can reject the null hypothesis and accept the alternative hypothesis that the average extra sleep is significantly different from zero. The confidence interval CI (95% by default) is (0.90, 3.76) to 2dps, so the mean is 2.33 with a 95% CI of (0.90, 3.76).

We want to know, if the average amount of extra sleep is greater (one-tailed test)?

**#** One-sample t-test

**x <- t.test(group\_2, mu = 0, alternative = 'greater')**

**x**

P-value is less than the significance level 0.05, so we can reject the null hypothesis and accept the alternative hypothesis that the average extra sleep is greater than zero (i.e. the drug is effective).

We want to know, if the extra sleep is less than zero (one-tailed test)?

**#** One-sample t-test

**x <- t.test(group\_2, mu = 0, alternative = 'less')**

**x**

P-value is greater than the significance level 0.05, so we cannot reject the null hypothesis that average extra sleep is no greater than zero.

1. **Unpaired two-samples t-test**

The unpaired two-samples t-test is used to compare the mean of two independent groups. To perform the two-samples t-test comparing the means of two independent samples (x & y), the R function t.test() can be used as follow:

t.test(x, y, alternative = "two.sided", var.equal = FALSE)

# load the data

# use the R built-in dataset iris (morphologic variation of Iris flowers of three species)

# check the dataset

**head(iris)**

# Statistical summaries of the data

**summary(iris)**

# create a new dataframe

**my\_data <- iris[iris$Species!='setosa',][,c('Sepal.Length','Species')]**

# Statistical summaries of my selected data

**summary(my\_data)**

# plot Sepal length by different species

**library(ggplot2)**

**p <- ggplot(data = my\_data, aes(x=Species, y=Sepal.Length))**

**p+geom\_boxplot(aes(fill=Species))**

Preliminary test to check independent t-test assumptions

1. Are the two samples independent?

Yes, the samples from two species are independent.

2. Do the data from each of the 2 groups follow a normal distribution?

# use Shapiro-Wilk test to check normality

**versicolor\_sepal <- my\_data$Sepal.Length[my\_data$Species == "versicolor"]**

**virginica\_sepal <- my\_data$Sepal.Length[my\_data$Species == "virginica"]**

**shapiro.test(versicolor\_sepal)**

**shapiro.test(virginica\_sepal)**

From the output, the two p-values are greater than the significance level 0.05 implying that the distribution of the data are not significantly different from the normal distribution. In other words, we can assume normality.

3. Do the two populations have the same variances?

# Use F-test to test for homogeneity in variances

**var.test(versicolor\_sepal, virginica\_sepal, var.equal = TRUE)**

The p-value of F-test is p = 0.1478. It’s greater than the significance level alpha = 0.05. In conclusion, there is no significant difference between the variances of the two sets of data. Therefore, we can use the classic t-test which assumes equality of the two variances.

Compute unpaired two-samples t-test

Question : Is there any significant difference between the Sepal length of these two different Species?

# Compute t-test

**t.test(versicolor\_sepal, virginica\_sepal, var.equal = TRUE)**

From the output, the p-value is less than the significant level 0.05 implying that there is a significant difference between the Sepal length of these two different Species.

We want to know whether the average Sepal length of versicolor is less than virginica, type the following code:

**t.test(versicolor\_sepal, virginica\_sepal, var.equal = TRUE, alternative = "less")**

We want to know whether the average Sepal length of versicolor is greater than virginica, type the following code:

**t.test(versicolor\_sepal, virginica\_sepal, var.equal = TRUE, alternative = "greater")**

1. **Paired samples t-test**

The paired samples t-test is used to compare the means between two related groups of samples. In this case, you have two values (i.e., pair of values) for the same samples. For example, paired t-test can be used to compare the results before and after treatment. To perform paired samples t-test comparing the means of two paired samples (x & y), the R function t.test() can be used as follow:

t.test(x, y, paired = TRUE, alternative = "two.sided")

# load the data

# use the R built-in dataset sleep (Data which show the effect of two soporific drugs)

# check the details of the dataset

**help(sleep)**

# check the dataset

**head(sleep)**

# Statistical summaries of the data

**summary(sleep)**

# plot the increase in hours of sleep by different group

**library(ggplot2)**

**p <- ggplot(data = sleep, aes(x=group, y=extra))**

**p+geom\_boxplot(aes(fill=group))**

Preliminary test to check paired t-test assumptions

1: Are the two samples paired?

Yes, since the data have been collected from measuring the extra hours of sleep for the same group of patients (10 patients) for two different drugs.

2: Is this a large sample?

No, because n < 30. Since the sample size is not large enough (less than 30), we need to check whether the differences of the pairs follow a normal distribution.

3. How to check the normality?

# use Shapiro-Wilk test to check normality

**group\_1 <- sleep$extra[sleep$group==1]**

**group\_2 <- sleep$extra[sleep$group==2]**

**shapiro.test(group\_1)**

**shapiro.test(group\_2)**

From the output, the two p-values are greater than the significance level 0.05 implying that the distribution of the data are not significantly different from the normal distribution. In other words, we can assume normality.

# Compute t-test

**t.test(group\_1, group\_2, paired = TRUE)**

From the output, the p-value is less than the significant level 0.05 implying that there is a significant difference between the extra sleep hours of the two groups (i.e. treatments).

We want to know whether the extra hours of sleep of group 1 is less than group 2, type the following code:

**t.test(group\_1, group\_2, paired = TRUE, alternative = "less")**

We want to know whether the extra hours of sleep of group 1 is less than group 2, type the following code:

**t.test(group\_1, group\_2, paired = TRUE, alternative = "greater")**

1. **One-Way ANOVA Test**

The one-way analysis of variance (ANOVA), also known as one-factor ANOVA, is an extension of independent two-samples t-test for comparing means in a situation where there are more than two groups. In one-way ANOVA, the data is organized into several groups base on one single grouping variable (also called a factor variable).

# load the data

# use the R built-in dataset PlantGrowth (the weight of plants obtained under a control and two different treatment conditions)

# check the details of the dataset

**help(PlantGrowth)**

# load the dataset

**my\_data <- PlantGrowth**

# Show the levels

**levels(my\_data$group)**

# Statistical summaries of the data

**summary(my\_data)**

# plot weight by group and color by group

**library(ggplot2)**

**p <- ggplot(data = my\_data, aes(x=group, y=weight))**

**p+geom\_boxplot(aes(fill=group))**

Compute one-way ANOVA test

We want to know if there is any significant difference between the average weights of plants in the 3 experimental conditions. The R function aov() can be used to compute one-way ANOVA test. The function summary.aov() is used to summarize the analysis of variance model.

# Compute the analysis of variance

**res.aov <- aov(weight ~ group, data = my\_data)**

# Summary of the analysis

**summary(res.aov)**

As the p-value is less than the significance level 0.05, we can conclude that there are significant differences between the groups highlighted with “\*" in the model summary.

**Exercise 1. For an extra check on normality, produce Q-Q plots for the variables from the Iris dataset considered in section 3 above.**

**Exercise 2. Consider the sample of heights given in the lecture slides:**

**1.35 1.65 1.34 1.46 2.15 1.92 1.90 1.75 2.00 1.80 1.63 1.86 1.82 1.81 1.57**

**Determine the mean, standard deviation and standard error of the mean.**

**Exercise 3. Use the prop.test function (see lecture slides) to determine whether there is a significant difference between the proportions of males and females on the Titanic who survived.**

**Hints:**

**Read in the train.csv file we have used in previous weeks.**

**Find the relevant numbers you need for the proportions, e.g. if the data frame is df**

**num\_female\_survive <- sum(df$Sex=="female" & df$Survived==1)**

**You could also obtain these numbers from a contingency table:**

**tbl1 <- table(df$Survived, df$Sex)**

**You can then use the relevant numbers to call the prop.test function.**