CM 2062 - Statistical Computing with R
Lab Sheet 10

One sample t-Test in R

We can use the **t.test( )** function to carry out one-sample t-tests in R. (Note: There are no built-in z-test functions in R because when we work with real data, we never know the population variance!)

The syntax is,

```
t.test(x, alternative, mu, conf.level)
```

x: data on the variable of interest
alternative: what type of alternative hypothesis is specified? (options: "two.sided", "greater", "less")
mu: the value of μ under the null hypothesis
conf.level: confidence level of the test $(1-\alpha)$

Example 1
Let's consider the weight of 10 mice in gram 17.6, 20.6, 22.2, 15.3, 20.9, 21.0, 18.9, 18.9, 18.9, 18.2. Test whether the mean weight of mice is different from 25g.

Answer
Since it has not mention in the question whether the data are coming from a normal distribution and $n < 30$, we need to check whether the data follows a normal distribution. So, we can apply **Shapiro-Wilk test** in R to check for normality of the data.

Note: Shapiro-Wilk test
Null hypothesis: the data are normally distributed
Alternative hypothesis: the data are not normally distributed

```
shapiro.test(Weight)

Shapiro-Wilk normality test

data:  Weight
W = 0.9526, p-value = 0.6993
```

From the output, since the p-value (0.6993) is greater than the significance level 0.05 implying that the distribution of the data are not significantly different from normal distribution. In other words, we can assume the normality.

```
> res <- t.test(Weight, mu = 25)
> res

        One Sample t-test

data:  Weight
t = -9.0783, df = 9, p-value = 7.953e-06
alternative hypothesis: true mean is not equal to 25
95 percent confidence interval:
 17.8172 20.6828
sample estimates:
mean of x
    19.25
```

In the result above:

t is the t-test statistic value (t = -9.078)
df is the degrees of freedom (df= 9)
p-value is the significance level of the t-test (p-value = 7.95310-6).
conf.int is the confidence interval of the mean at 95% (conf.int = [17.8172, 20.6828])
sample estimates is he mean value of the sample (mean = 19.25)

Note: If you want to test whether the mean weight of mice is less than 25g (one-tailed test), type this:

```
t.test(Weight, mu = 25,
              alternative = ”less”)
```

Or, if you want to test whether the mean weight of mice is greater than 25g (one-tailed test), type this:

```
t.test(Weight, mu = 25,
              alternative = ”greater”)
```

Interpretation of the result
The p-value of the test is 7.95310-6, which is less than the significance level $\alpha = 0.05$. We can conclude that the mean weight of the mice is significantly different from 25g with a p-value = 7.95310-6.

Access to the values returned by t.test() function

The result of **t.test()** function is a list containing the following components:

statistic: the value of the t test statistics
parameter: the degrees of freedom for the t test statistics
p.value: the p-value for the test
conf.int: a confidence interval for the mean appropriate to the specified alternative hypothesis.
estimate: the mean of the sample.

```
> res$statistic
        t
-9.078319

> res$parameter
df
 9

> res$p.value
[1] 7.953383e-06

> res$conf.int
[1] 17.8172 20.6828
attr(,”conf.level”)
[1] 0.95

> res$estimate
mean of x
    19.25
```

Exercise 1

The Sugar level of a Cookie follows a normal distribution with mean 9.8 and the standard deviation 0.05. Let’s take a sample of size 30.

1. Generate 30 random numbers (sugar levels) from the above distribution.

2. Test whether the mean sugar level of the Cookies is greater than 10.

Exercise 2

Consider the mice data set in ”datarium” package in R. Test whether the mean weight of mice is different from 25g.

Two independent samples t-Test in R

Two independent samples t-Test is used to compare the mean of two independent groups. We can use the same **t.test()** function to conduct the two independent sample t-Test.

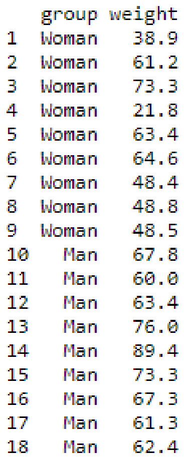
Syntax is:

```
t.test(x, y, alternative = c(”two.sided”, ”less”, ”greater”),
        , var.equal = FALSE)
```

x,y: numeric vectors
alternative: the alternative hypothesis. Allowed value is one of “two.sided” (default), “greater” or “less”.
var.equal: a logical variable indicating whether to treat the two variances as being equal. If TRUE then the pooled variance is used to estimate the variance otherwise the Welch test is used.

Note: Note that, unpaired two-samples t-test can be used only under certain conditions: when the two groups of samples (A and B), being compared, are normally distributed. This can be checked using Shapiro-Wilk test. When the variances of the two groups are equal. This can be checked using F-test.

Example 1
Suppose that we have measured the weight of 18 individuals: 9 women and 9 men. We want to know if the mean weight of women (μ_A) is significantly different from that of men (μ_B).



```
   group weight
1  Woman   38.9
2  Woman   61.2
3  Woman   73.3
4  Woman   21.8
5  Woman   63.4
6  Woman   64.6
7  Woman   48.4
8  Woman   48.8
9  Woman   48.5
10   Man   67.8
11   Man   60.0
12   Man   63.4
13   Man   76.0
14   Man   89.4
15   Man   73.3
16   Man   67.3
17   Man   61.3
18   Man   62.4
```

Answer

First we have to create a data frame to feed the data in to R.

```
women_weight <- c(38.9, 61.2, 73.3, 21.8, 63.4, 64.6, 48.4, 48.8, 48.5)
men_weight <- c(67.8, 60, 63.4, 76, 89.4, 73.3, 67.3, 61.3, 62.4)
```

```
data <- data.frame( group = rep(c(”Woman”, ”Man”), each = 9),
  weight = c(women_weight,  men_weight))
data
```

Let’s check whether the tow sample data women_weight and men_weight are coming from normal distributions using Shapiro-Wilk test in R.

We’ll use the functions **with()** and **shapiro.test()** to compute Shapiro-Wilk test for each group of samples.

```
> with(data, shapiro.test(weight[group == ”Man”]))

        Shapiro-Wilk normality test

data:  weight[group == ”Man”]
W = 0.86425, p-value = 0.1066
```

Since p-value (0.106) > significant level (0.05), it implies that the men_weight data are coming from a normally distributed population.

```
> with(data, shapiro.test(weight[group == ”Woman”]))

        Shapiro-Wilk normality test

data:  weight[group == ”Woman”]
W = 0.94266, p-value = 0.6101
```

Since p-value (0.6101) > significant level (0.05), it implies that the women_weight data are coming from a normally distributed population.

Then we need to check whether the two population have the same variance. We’ll use **F-test** to test for homogeneity in variances. This can be performed with the function **var.test()** as given below,

```
> ftest <- var.test(weight ~ group, data = data)
> ftest

        F test to compare two variances

data:  weight by group
F = 0.36134, num df = 8, denom df = 8, p-value = 0.1714
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.08150656 1.60191315
sample estimates:
ratio of variances
         0.3613398
```

The p-value of F-test (0.1713596) > 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 two independent samples t-test witch assume equality of the two variances.

There are two different ways to specify the variables in the two independent sample t-test.

Method 1

```
> res <- t.test(women_weight,men_weight, var.equal = TRUE)
> res

        Two Sample t-test

data:  women_weight and men_weight
t = -2.7842, df = 16, p-value = 0.01327
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -29.748019  -4.029759
sample estimates:
mean of x mean of y
 52.10000  68.98889
```

Method 2

```
> res <- t.test(weight ~ group, data = data, var.equal = TRUE)
> res

        Two Sample t-test

data:  weight by group
t = 2.7842, df = 16, p-value = 0.01327
alternative hypothesis: true difference in means between group Man and group
Woman is not equal to 0
95 percent confidence interval:
  4.029759 29.748019
sample estimates:
  mean in group Man mean in group Woman
           68.98889            52.10000
```

The p-value of the test is 0.01327, which is less than the significance level 0.05. We can conclude that men's average weight is significantly different from women's average weight.

Note:

If you want to test whether the average men's weight is less than the average women's weight

```
t.test(weight ~ group, data = data,
        var.equal = TRUE, alternative = "less")
```

If you want to test whether the average men's weight is greater than the average women's weight

```
t.test(weight ~ group, data = data,
        var.equal = TRUE, alternative = "greater")
```

Exercise 1

Test the hypothesis that Clevelanders and New Yorkers spend different amounts monthly eating out. A random sample of size 50 was taken from normal distribution with mean 250 and standard deviation 75 for Clevelanders spending. A random sample of size 50 was taken from normal distribution with mean 300 and standard deviation 80 for New Yorkers spending.

(a) Generate random numbers for two spending distributions.

(b) Test the given hypothesis.

Exercise 2

Let's consider the "genderweight" data set in "datarium" package in R.

(a) Get the descriptive statistics of weight for two gender groups.

(b) Show the weight distribution of the two groups using a suitable plot.

(c) Test whether the mean weights of the two groups are different.

Paired samples t-test

The paired sample t-test is used to compare the means of two related groups of samples. Put into another words, it's used in a situation where you have two pairs of values measured for the same samples.

Syntax:

```
t.test(x, y,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)
```

Example 1

let's say that we work at a large health clinic and we are testing a new drug, Procardia, whose work is to reduce hypertension. We find 2000 individuals with high systolic blood pressure normally distributed with mean=150 mmHg and SD=10 mmHg and we provide them Procardia for a month, and then measure their blood pressure again. We find that the mean systolic blood pressure has decreased to 144 mmHg with a standard deviation of 9 mmHg. Test whether the mean systolic blood pressure of the patients before giving the treatment is higher than the mean systolic blood pressure of the patients after giving the treatment.

Answer

Before conduct the paired sample t-test let's check the variances of the two distributions are equal or not using f-test.

```
> pre_Treatment <- c(rnorm(2000, mean = 150, sd = 10))
> post_Treatment <- c(rnorm(2000, mean = 144, sd = 9))

> ftest <- var.test(pre_Treatment, post_Treatment)
> ftest

        F test to compare two variances

data:  pre_Treatment and post_Treatment
F = 1.1801, num df = 1999, denom df =
1999, p-value = 0.0002155
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.08105 1.28831
sample estimates:
ratio of variances
          1.180139
```

Since the p-value is less than 0.05 it suggest that the variances of the two groups are different.

Note

If the data are not coming from the normal distributions, before apply the paired sample t-test you have to check whether the difference distribution follows a normal distribution.

```
> t.test(pre_Treatment, post_Treatment, paired = TRUE,
var.equal = FALSE, alternative= ”greater”)

        Paired t-test

data:  pre_Treatment and post_Treatment
t = 20.12, df = 1999, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 5.569336       Inf
sample estimates:
mean of the differences
               6.065421
```

Since p-value is less than 0.05, there is a significant evidence to suggest that the mean systolic blood pressure of the patients before giving the treatment is higher than the mean systolic blood pressure of the patients after giving the treatment.

Example 2

Here, we'll use an example data set, which contains the weight of 10 mice before and after the treatment. Test whether the mean weight of the mice has been increased after giving the treatment.

```
before <- c(200.1, 190.9, 192.7, 213, 241.4, 196.9, 172.2, 185.5, 205.2, 193.7)
after <- c(392.9, 393.2, 345.1, 393, 434, 427.9, 422, 383.9, 392.3, 352.2)
```

Since it has not mentioned that the data are coming from normal distributions, let's use the Shapiro-Wilk test to check whether the difference distribution follows a normal distribution.

```
> diff <- before - after
> shapiro.test(diff)

        Shapiro-Wilk normality test

data:  diff
W = 0.94536, p-value = 0.6141
```

Since the p value is greater than 0.05, there is significant evidence to suggest that the difference distribution follows a normal distribution.

Then, let's check whether the variances of the two distributions are equal or not using the f-test.

```
> ftest <- var.test(before, after)
> ftest

        F test to compare two variances

data:  before and after
F = 0.39488, num df = 9, denom df = 9, p-value
= 0.1825
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.09808242 1.58978163
sample estimates:
ratio of variances
         0.3948793
```

Since the p value is greater than 0.05, it indicate that the variances of the two distributions are equal.

Let’s apply the paired samples t-test,

```
> t.test(before, after, paired = TRUE, var.equal = TRUE, alternative= ”less”)

        Paired t−test

data:  before and after
t = −20.883, df = 9, p−value = 3.1e−09
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     −Inf −177.4177
sample estimates:
mean of the differences
                −194.49
```

Since the p value is less than 0.05, there is significant evidence to suggest that the mean weight of the mice has been increased after giving the treatment.

Exercise 1

Let’s consider the ”mice2” data set in ”datarium” package which contains the weight of 10 mice before and after the treatment.

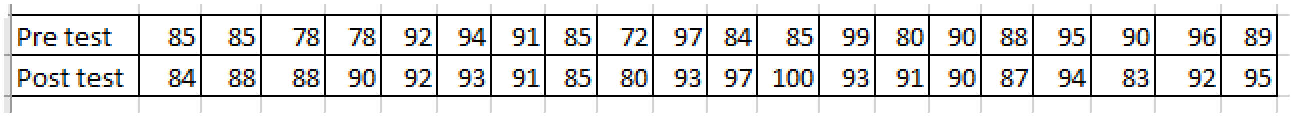
(a) Get the descriptive statistics of weight for before and after the treatment.

(b) Show the weight distribution of the before and after the treatment using a suitable plot.

(c) Test whether there is any significant difference in the mean weights before and after the treatment.

Exercise 2

The marks of 20 students for a pre test and post test are as given below.



Pre test	85	85	78	78	92	94	91	85	72	97	84	85	99	80	90	88	95	90	96	89
Post test	84	88	88	90	92	93	91	85	80	93	97	100	93	91	90	87	94	83	92	95

Find out if there is a significant difference in the mean scores between a pre-test and a post-test.