Data Science 1
Statistical Inference for Two Population Means

Ann Maharaj

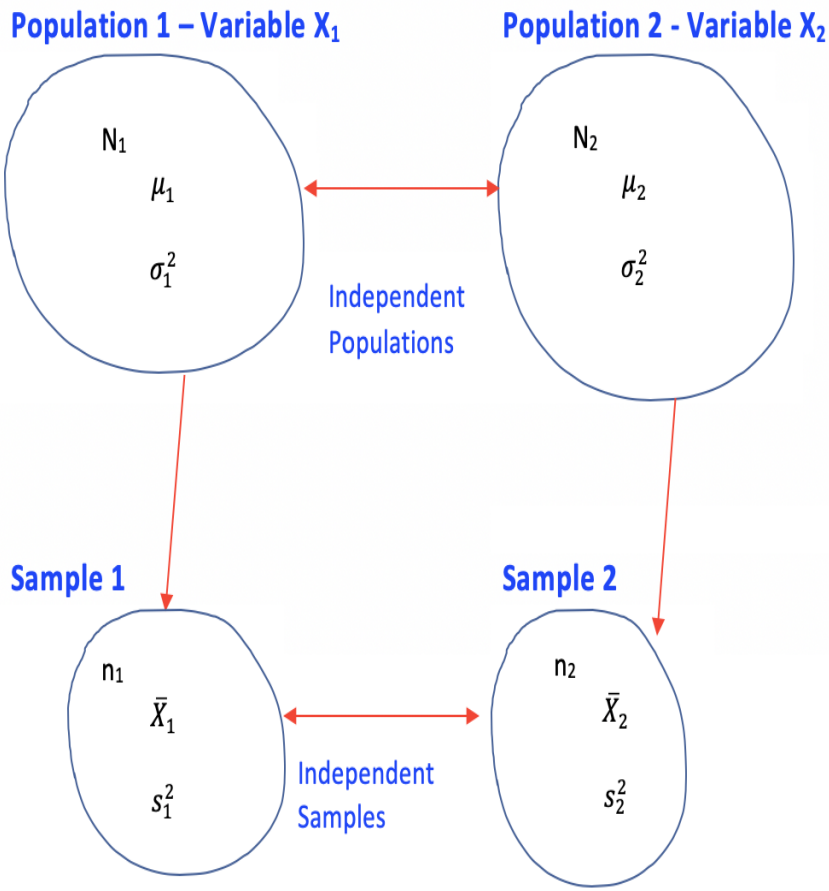
Statistical Inference for Two Populations Means

Statistical Inference for Two Populations Means

- Independent populations
  - For example, we may wish to compare mean household incomes in two suburbs, one in the inner city and the other of the outskirts of the city.
- Related populations
  - For example, we may wish to compare computer skills of employees in a particular company before and after a skills test.

Independent Populations

- We are interested in some variable of interest for each of two independent populations.
- X_1 and X_2 are variables associated with the first and second populations, respectively.
- The population means are μ_1 and μ_2, the population variances are σ_1^2 and σ_2^2.
- We take random samples of size n_1 and n_2 from the populations to estimate the difference between the means.
- We assume that X_1 and X_2 are normally distributed or the samples are sufficiently large.



Sampling Distribution - σ_1^2 and σ_2^2 are known

- A **point estimator** of $\mu_1 - \mu_2$ is the difference between sample means $\overline{X_1} - \overline{X_2}$.
- $\overline{X_1} - \overline{X_2}$ is an unbiased estimator of $\mu_1 - \mu_2$ and it is normally distributed with mean

$$E\left(\overline{X_1} - \overline{X_2}\right) = \mu_1 - \mu_2$$

and standard deviation (standard error)

$$SE\left(\overline{X_1} - \overline{X_2}\right) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

The standardised difference $\left(\overline{X_1} - \overline{X_2}\right)$ follows a standard normal distribution.

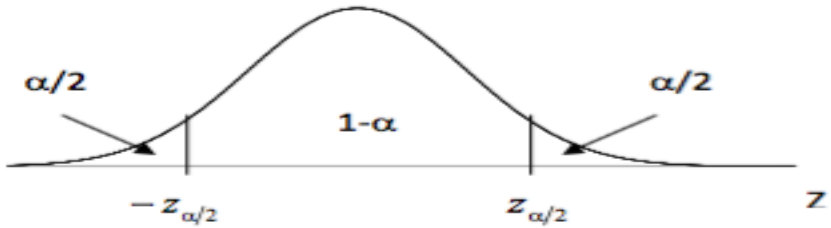
$$Z = \frac{\left(\overline{X_1} - \overline{X_2}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Confidence Interval Estimation - σ_1^2 and σ_2^2 are known

Since the standardised difference between the sample means, $\left(\overline{X_1} - \overline{X_2}\right)$ follows a N(0,1) distribution, and

$$P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = 1 - \alpha$$

$$\Rightarrow P\left(-z_{\alpha/2} < \frac{\left(\overline{X_1} - \overline{X_2}\right) - (\mu_1 - \mu_2)}{SE\left(\overline{X_1} - \overline{X_2}\right)} < z_{\alpha/2}\right) = 1 - \alpha,$$



it can be shown that on making $(\mu_1 - \mu_2)$ the subject of the formula:

- A $100(1-\alpha)\%$ Confidence Interval estimator of $\mu_1 - \mu_2$ is

$$\left(\overline{X_1} - \overline{X_2}\right) \pm z_{\alpha/2} \times SE\left(\overline{X_1} - \overline{X_2}\right)$$

where

$$SE\left(\overline{X_1} - \overline{X_2}\right) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

- If the population variances are assumed to be equal, i.e., $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$SE\left(\overline{X_1} - \overline{X_2}\right) = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

- This property of equal variance which is often assumed, is referred to as homoscedasticity.

Hypothesis Testing - σ_1^2 and σ_2^2 are known

The **population variances are known**, the random variables of interest X_1 and X_2 are normally distributed, or the sample sizes n_1 and n_2 are *sufficiently large*.

Table: 1

Null Hypothesis	$H_0 : \mu_1 - \mu_2 = D_0$	$H_0 : \mu_1 - \mu_2 = D_0$ or $H_0 : \mu_1 - \mu_2 \leq D_0$	$H_0 : \mu_1 - \mu_2 = D_0$ or $H_0 : \mu_1 - \mu_2 \geq D_0$
Alternative Hypothesis	$H_1 : \mu_1 - \mu_2 \neq D_0$	$H_1 : \mu_1 - \mu_2 > D_0$	$H_1 : \mu_1 - \mu_2 < D_0$
Level of Significance	α	α	α
Test Statistic	$Z = \frac{\left(\overline{X_1} - \overline{X_2}\right) - (\mu_1 - \mu_2)}{SE\left(\overline{X_1} - \overline{X_2}\right)}$	$Z = \frac{\left(\overline{X_1} - \overline{X_2}\right) - (\mu_1 - \mu_2)}{SE\left(\overline{X_1} - \overline{X_2}\right)}$	$Z = \frac{\left(\overline{X_1} - \overline{X_2}\right) - (\mu_1 - \mu_2)}{SE\left(\overline{X_1} - \overline{X_2}\right)}$
P-Value Approach Reject H_0 if	$p - value < \alpha$	$p - value < \alpha$	$p - value < \alpha$

$$SE\left(\overline{X_1} - \overline{X_2}\right) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

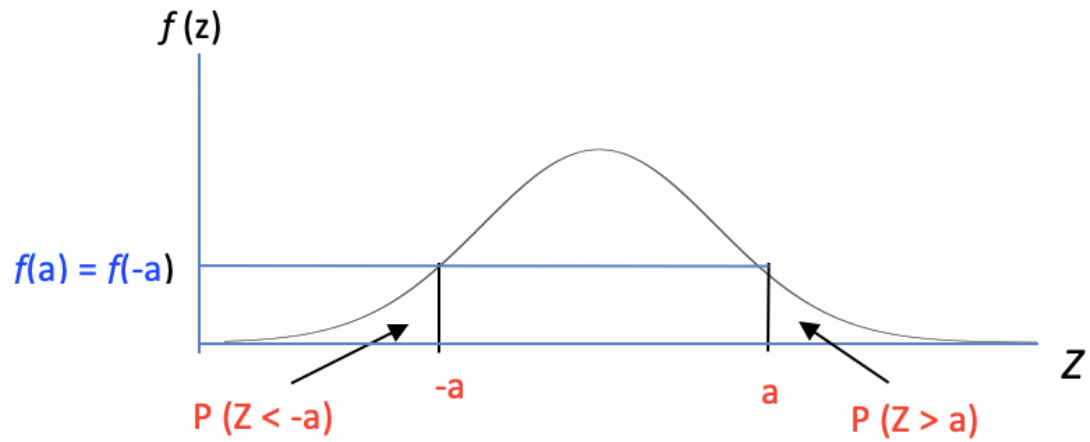
If $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$SE\left(\overline{X_1} - \overline{X_2}\right) = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

P-value approach - σ_1^2 and σ_2^2 are known

$H_0 : \mu_1 - \mu_2 = D_0$
$H_1 : \mu_1 - \mu_2 \neq D_0$

$$Z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{SE(\overline{X_1} - \overline{X_2})} \sim N(0,1)$$



a = |numerical value of the test statistic|

$f(z)$ is the density function of the standard normal distribution and $f(-a) = f(a)$

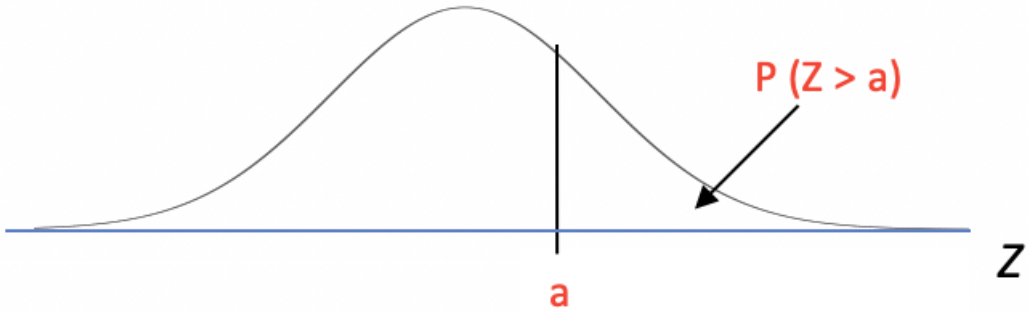
p-value = P(Z > a) + P(Z < -a)

Reject H_0 if $p - value < \alpha$

P-value approach - σ_1^2 and σ_2^2 are known

$H_0 : \mu_1 - \mu_2 = D_0$ or $H_0 : \mu_1 - \mu_2 \leq D_0$
$H_1 : \mu_1 - \mu_2 > D_0$

$$Z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{SE(\overline{X_1} - \overline{X_2})} \sim N(0, 1)$$



a = numerical value of the test statistic

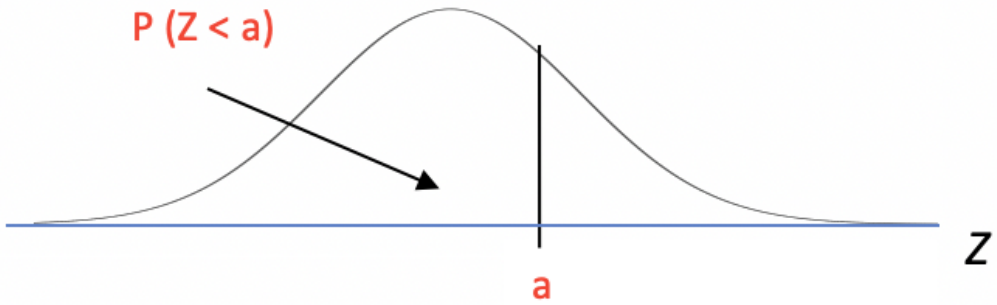
p-value = P(Z > a)

Reject H_0 if $p - value < \alpha$

P-value approach - σ_1^2 and σ_2^2 are known

$H_0 : \mu_1 - \mu_2 = D_0$ or $H_0 : \mu_1 - \mu_2 \geq D_0$
$H_1 : \mu_1 - \mu_2 < D_0$

$$Z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{SE(\overline{X_1} - \overline{X_2})} \sim N(0, 1)$$



a = numerical value of the test statistic

p-value = P(Z < a)

Reject H_0 if $p - value < \alpha$

Sampling Distribution - σ_1^2 and σ_2^2 are unknown

The random variables of interest, X_1 and X_2 are normally distributed.

- When the population variances, σ_1^2 and σ_2^2 are unknown, they are estimated by the sample variances s_1^2 and s_2^2, respectively. Then

$$SE\left(\overline{X_1} - \overline{X_2}\right) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

- Then the standardised difference between the sample means $\overline{X_1} - \overline{X_2}$ follows a t-distribution.

$$t = \frac{\left(\overline{X_1} - \overline{X_2}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_{df^*},$$

with

$$df^* = \frac{\left[s_1^2/n_1 + s_2^2/n_2\right]^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}.$$

- Fractional values of df^* are rounded down.

Sampling Distribution - σ_1^2 and σ_2^2 are unknown

The random variables of interest, X_1 and X_2 are normally distributed.

If the population variances are assumed to be equal, the common variance is estimated by pooling the estimates of the variances of both the samples. The common standard deviation is:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}.$$

Then

$$SE\left(\overline{X_1} - \overline{X_2}\right) = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

The standardised difference between the sample means $\overline{X_1} - \overline{X_2}$ is

$$t = \frac{\left(\overline{X_1} - \overline{X_2}\right) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{df},$$

with

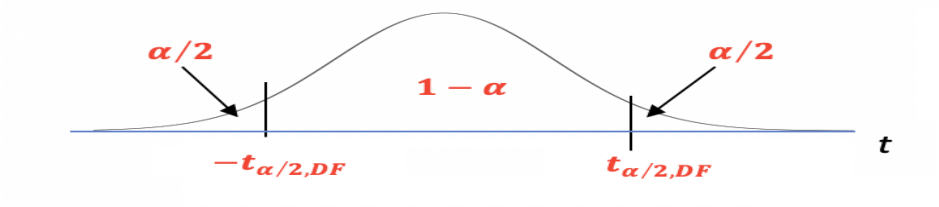
$$df = n_1 + n_2 - 2.$$

Confidence Interval Estimation - σ_1^2 and σ_2^2 are unknown

Since the standardised difference between the sample means, $\left(\overline{X_1} - \overline{X_2}\right)$ follows a t-distribution, and

$$P\left(-t_{\alpha/2,DF} < t < t_{\alpha/2,DF}\right) = 1 - \alpha$$

$$\Rightarrow P\left(-t_{\alpha/2,DF} < \frac{\left(\overline{X_1} - \overline{X_2}\right) - (\mu_1 - \mu_2)}{SE\left(\overline{X_1} - \overline{X_2}\right)} < t_{\alpha/2,DF}\right) = 1 - \alpha,$$



it can be shown that on making $(\mu_1 - \mu_2)$ the subject of the formula:

- A $100(1-\alpha)\%$ Confidence Interval estimator of $\mu_1 - \mu_2$ is

$$\left(\overline{X_1} - \overline{X_2}\right) \pm t_{\alpha/2,DF} \times SE\left(\overline{X_1} - \overline{X_2}\right)$$

where

$$SE\left(\overline{X_1} - \overline{X_2}\right) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

and DF=df^*

- If the population variances are assumed to be equal,

$$SE\left(\overline{X_1} - \overline{X_2}\right) = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}.$$

and DF = df

Hypothesis Testing - σ_1^2 and σ_2^2 are unknown

The random variables of interest X_1 and X_2 are normally distributed.

Table: 1

Null Hypothesis	$H_0 : \mu_1 - \mu_2 = D_0$	$H_0 : \mu_1 - \mu_2 = D_0$ or $H_0 : \mu_1 - \mu_2 \leq D_0$	$H_0 : \mu_1 - \mu_2 = D_0$ or $H_0 : \mu_1 - \mu_2 \geq D_0$
Alternative Hypothesis	$H_1 : \mu_1 - \mu_2 \neq D_0$	$H_1 : \mu_1 - \mu_2 > D_0$	$H_1 : \mu_1 - \mu_2 < D_0$
Level of Significance	α	α	α
Test Statistic	$t = \frac{\left(\overline{X_1} - \overline{X_2}\right) - (\mu_1 - \mu_2)}{SE\left(\overline{X_1} - \overline{X_2}\right)}$	$t = \frac{\left(\overline{X_1} - \overline{X_2}\right) - (\mu_1 - \mu_2)}{SE\left(\overline{X_1} - \overline{X_2}\right)}$	$t = \frac{\left(\overline{X_1} - \overline{X_2}\right) - (\mu_1 - \mu_2)}{SE\left(\overline{X_1} - \overline{X_2}\right)}$
P-Value Approach Reject H_0 if	$p - value < \alpha$	$p - value < \alpha$	$p - value < \alpha$

If we assume equal population variances:

$$SE\left(\overline{X_1} - \overline{X_2}\right) = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

If we assume unequal population variances:

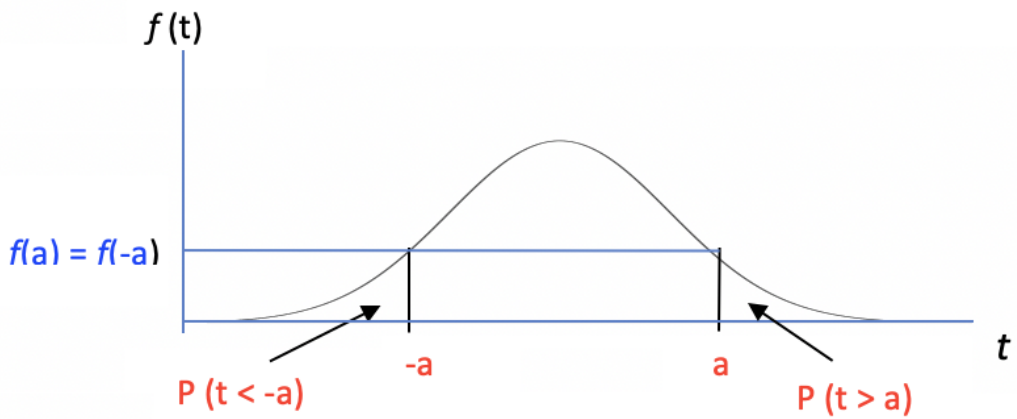
$$SE\left(\overline{X_1} - \overline{X_2}\right) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

P-value approach - σ_1^2 and σ_2^2 are unknown

$H_0 : \mu_1 - \mu_2 = D_0$
$H_1 : \mu_1 - \mu_2 \neq D_0$

$$t = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{SE\left(\overline{X_1} - \overline{X_2}\right)}$$

$t \sim t_{df}$ or $t \sim t_{df*}$



a = |numerical value of the test statistic|

$f(t)$ is the density function of the t-distribution distribution and $f(a) = f(-a)$

p-value = P(t > a) + P(t < -a)

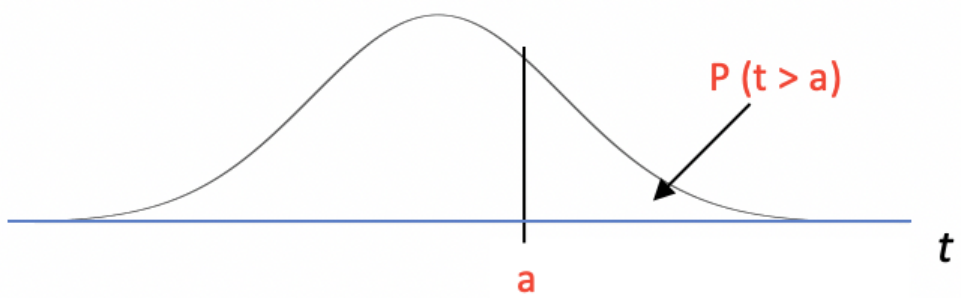
Reject H_0 if $p - value < \alpha$

P-value approach - σ_1^2 and σ_2^2 are unknown

$H_0 : \mu_1 - \mu_2 = D_0$ or $H_0 : \mu_1 - \mu_2 \leq D_0$
$H_1 : \mu_1 - \mu_2 > D_0$

$$t = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{SE\left(\overline{X_1} - \overline{X_2}\right)}$$

$t \sim t_{df}$ or $t \sim t_{df*}$



a = numerical value of the test statistic

p-value = P(t > a)

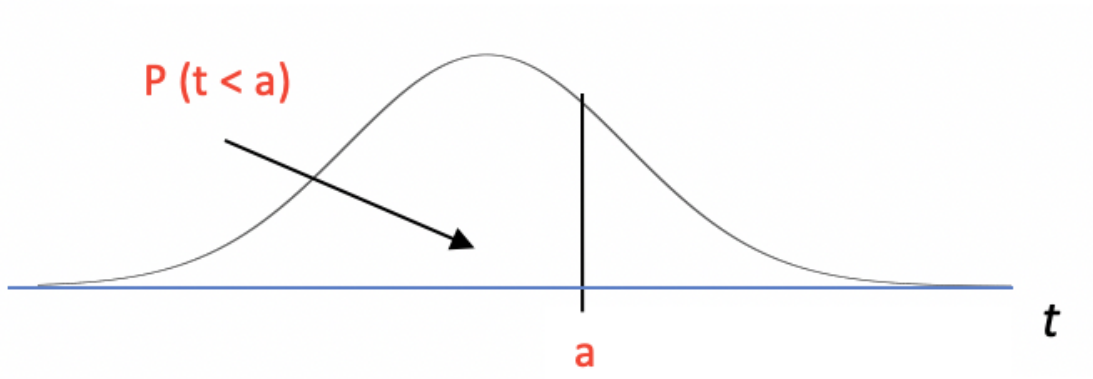
Reject H_0 if $p - value < \alpha$

P-value approach - σ_1^2 and σ_2^2 are unknown

$H_0 : \mu_1 - \mu_2 = D_0$ or $H_0 : \mu_1 - \mu_2 \geq D_0$
$H_1 : \mu_1 - \mu_2 < D_0$

$$t = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{SE\,(\overline{X_1} - \overline{X_2})}$$

$t \sim t_{df}$ or $t \sim t_{df*}$



a = numerical value of the test statistic

p-value = P(t < a)

Reject H_0 if $p - value < \alpha$

Example 1

A primary school has anticipated that a new instructional method will more effectively improve the reading ability of Grade 2 students than the standard method currently in use. To test this hypothesis, 20 children were randomly divided into two groups of 10 each. One group was instructed using the standard method while the other group was instructed using the new method. The students scores on reading tests are in the file reading.csv.

1. Test the hypothesis that the new instructional method has improved reading ability. Use the 5% level of significance.
2. Obtain a 95% confidence interval estimate of the difference in the reading ability between the new and current instructional methods.
3. Perform diagnostic checks to check for outliers and to assess the validity of the normality assumptions.

Example 1 - Part 1

```
> #t-test
> t.test(x,y, alternative="less", var.equal = FALSE)

        Welch Two Sample t-test

data:  x and y
t = -3.8021, df = 17.688, p-value = 0.0006702
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -3.586971
sample estimates:
mean of x mean of y
     69.2      75.8
```

```
> t.test(x,y, alternative="less", var.equal = TRUE)

        Two Sample t-test

data:  x and y
t = -3.8021, df = 18, p-value = 0.0006527
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -3.589847
sample estimates:
mean of x mean of y
     69.2      75.8
```

Example 1: Part 1

μ_1 and μ_2: population means associated with the standard and new reading instructional methods, respectively.
σ_1^2 and σ_2^2: population variances associated with the standard and new reading instructional methods, respectively.

- Hypotheses:
  $H_0 : \mu_1 - \mu_2 = 0$
  $H_1 : \mu_1 - \mu_2 < 0$
- Level of significance: $\alpha = 0.05$
- Test statistic and degrees of freedom:
  If we assume unequal population variances
  $t = -3.8021$ with $df = 17.688 \approx 17$
  If we assume equal population variances
  $t = -3.8021$ with $df = 18$
- P-value
  If we assume unequal population variances:
  $p - value = 0.00067$
  If we assume equal population variances:
  $p - value = 0.00065$.
- Decision Rule: Reject H_0 if p-value < 0.05
- Conclusion: Since p-value < 0.05, regardless of whether we assume equal population variances or not, reject H_0 at the 5% level or indeed any reasonable level of significance.
  There is sufficient evidence to conclude that the new instructional method has improved reading performance.

Example 1 - Part 2

```
> # this will enable us to get a 95% two-sided confidence iterval
> t.test(x,y, alternative="two.sided", var.equal = FALSE)

        Welch Two Sample t-test

data:  x and y
t = -3.8021, df = 17.688, p-value = 0.00134
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.251593  -2.948407
sample estimates:
mean of x mean of y
     69.2      75.8
```

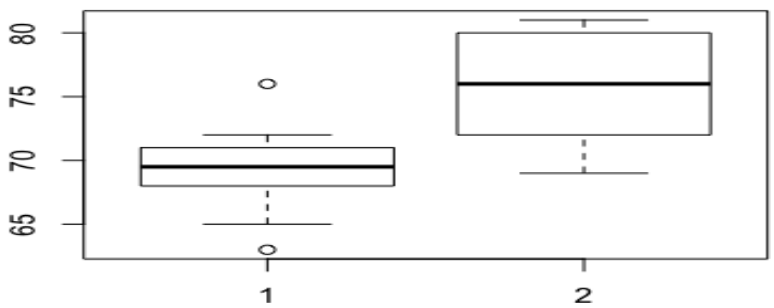
```
> t.test(x,y, alternative="two.sided", var.equal = TRUE)

        Two Sample t-test

data:  x and y
t = -3.8021, df = 18, p-value = 0.001305
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.246981  -2.953019
sample estimates:
mean of x mean of y
     69.2      75.8
```

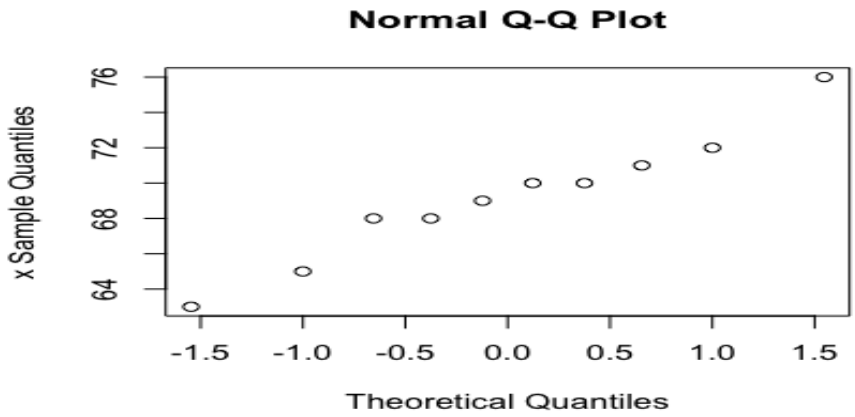
Example 1 - Part 2

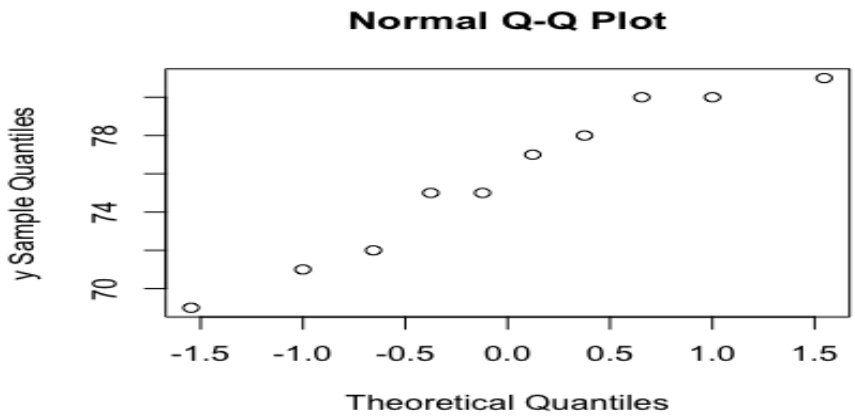
- Note that the 95% confidence interval derives from the fact that approximately 95% of intervals, computed in this manner for repeated samples of sizes n_1 and n_2 will cover the true mean difference $\mu_1 - \mu_2$.
- If we assume unequal population variances:
  The 95% confidence interval estimate of the difference in population means $\mu_1 - \mu_2$ is
  $[-10.251, -2.948]$
- If we assume equal population variances:
  The 95% confidence interval estimate of the difference in population means $\mu_1 - \mu_2$ is
  $[-10.246, -2.953]$
- We conclude with 95% confidence that the mean of the reading scores for the new instructional method is at least 2.95 units higher and can be as much as 10.25 units higher than the mean reading score of the standard instructional method.
- Because the confidence interval does not contain zero, the null hypothesis
  $H_0 : \mu_1 - \mu_2 = 0$
  is rejected in favour of the alternative
  $H_1 : \mu_1 - \mu_2 \neq 0$.

Example 1 - Part 3

```
> #check for outliers
> boxplot(x,y)
>
> #assess normality assumptions of x and y
> qqnorm(x, ylab = "x Sample Quantiles")
> shapiro.test(x)

        Shapiro-Wilk normality test

data:  x
W = 0.9717, p-value = 0.9061

> qqnorm(y, ylab = "y Sample Quantiles")
> shapiro.test(y)

        Shapiro-Wilk normality test

data:  y
W = 0.93885, p-value = 0.5403
```





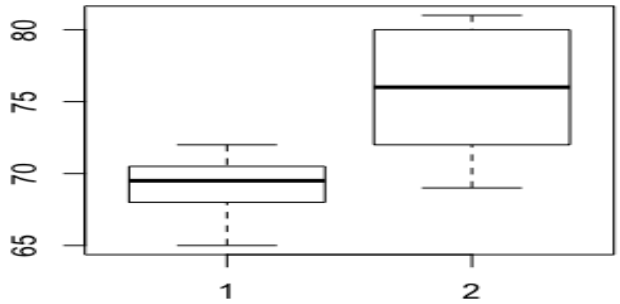


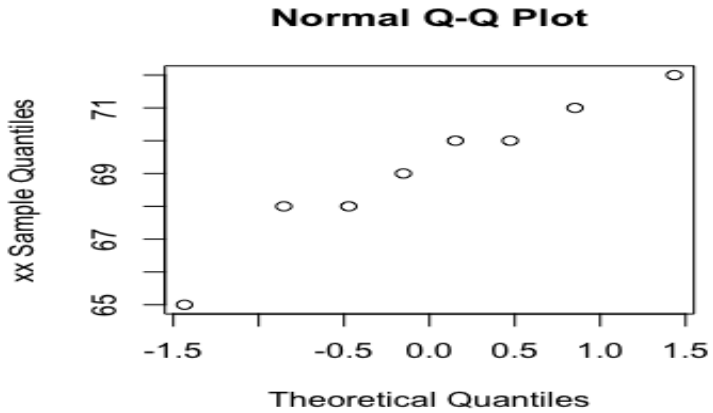
Example 1 - Part 3

- From the boxplot, we observe that the reading scores from the standard instructional method display two outliers, while those of the new instructional method do not display any outliers.
- From the Normal QQ-plot as well as from the results of the Shapiro-Wilk normality test for each variable, we conclude that the assumption of normality of the reading scores from both methods is valid.
- We remove the two outliers from the standard instructional method reading scores, and compute the results for the test in Part 1 and the confidence intervals in Part 2 again.

Example 1 - After removal of outliers

```
> boxplot(xx,y)
> qqnorm(xx, ylab = "xx Sample Quantiles")
> shapiro.test(xx)

        Shapiro-Wilk normality test

data:  xx
W = 0.95087, p-value = 0.72
```





- xx refers to the reading scores for the standard instructional method after the two outliers have been removed.
- The assumption of normality of reading scores for the standard instructional method is still valid.

Example 1-After removal of outliers

```
> t.test(xx,y, alternative="less", var.equal = FALSE)

        Welch Two Sample t-test

data:  xx and y
t = -4.4074, df = 14.11, p-value = 0.0002928
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -4.008958
sample estimates:
mean of x mean of y
   69.125    75.800

> t.test(xx,y, alternative="less", var.equal = TRUE)

        Two Sample t-test

data:  xx and y
t = -4.122, df = 16, p-value = 0.0003994
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -3.847798
sample estimates:
mean of x mean of y
   69.125    75.800
```

```
> t.test(xx,y, alternative="two.sided", var.equal = FALSE)

        Welch Two Sample t-test

data:  xx and y
t = -4.4074, df = 14.11, p-value = 0.0005856
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -9.920913 -3.429087
sample estimates:
mean of x mean of y
   69.125    75.800

> t.test(xx,y, alternative="two.sided", var.equal = TRUE)

        Two Sample t-test

data:  xx and y
t = -4.122, df = 16, p-value = 0.0007988
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.107875  -3.242125
sample estimates:
mean of x mean of y
   69.125    75.800
```

Example 1 - After removal of outliers

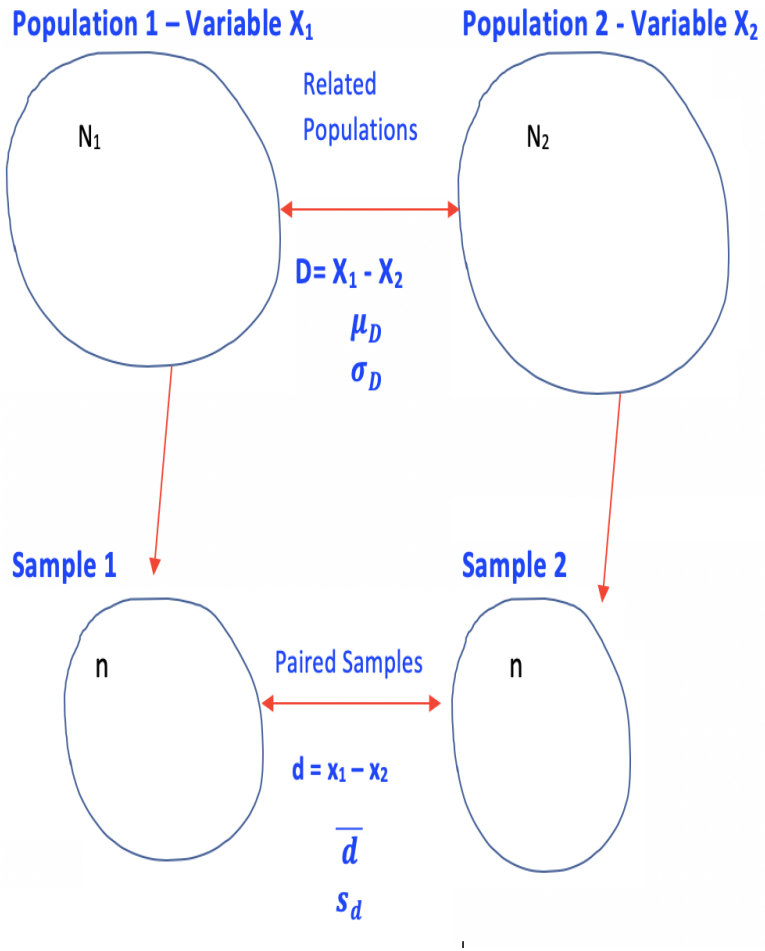
- As before that the null hypothesis
  $H_0 : \mu_1 - \mu_2 = 0$
  is rejected in favour of the alternative
  $H_1 : \mu_1 - \mu_2 < 0$
  albeit will slightly smaller p-values.

- The confidence intervals are slightly narrower.

- Hence, we conclude that removing the outliers,
  - does not affect of the outcome of the test of hypothesis.
  - leads to slightly narrower but more accurate confidence interval estimates.

Paired Samples

- When two samples are selected in such a way as to be dependent or related, each item or person in one sample has a corresponding match or related item in the other sample.
- Such samples also called paired or related samples because the population from which they are drawn are related.
- The procedure for paired samples does not directly analyse two separate variables X_1 and X_2 but analyses the difference between these two variables, i.e., $D = X_1 - X_2$.
- The samples are paired in some natural way.
  - For example, tests results for a sample of employees before and after a program of training.
  - Hence we have a common sample size $n_1 = n_2 = n$.



Paired Samples

- We take the differences in the sample values and then proceed as in the one sample case.
- $D = X_1 - X_2$: then for the paired samples, we calculate the difference d, the mean difference $\overline{d}$ and the standard deviation of the differences s_d.
- Note the D is the difference between the population values and d is the difference between the sample values.

Before X_1	After X_2	Difference $D = X_1 - X_2$
x_{11}	x_{21}	$d_1 = x_{11} - x_{21}$
x_{12}	x_{22}	$d_2 = x_{12} - x_{22}$
.	.	.
.	.	.
.	.	.
x_{1n}	x_{2n}	$d_n = x_{1n} - x_{2n}$

Mean of d

$$\overline{d} = \frac{\sum_{i=1}^{n} d_i}{n}$$

Standard deviation of d

$$s_d = \sqrt{\frac{\sum_{i=1}^{n} \left(d_i - \overline{d}\right)^2}{n-1}}$$

Sampling Distribution: Paired Samples

- $\overline{d}$ is an unbiased estimator of μ_D, which is the mean of $D = X_1 - X_2$.
- If it is assumed that $D = X_1 - X_2$ is normally distributed, the sampling distribution of $\overline{d}$ is approximately normal.

$$\overline{d} \sim N\left(\mu_D, \frac{\sigma_D^2}{n}\right)$$

- σ_D is the standard deviation of $D = X_1 - X_2$.
- If σ_D is unknown, as it would be in most cases, it is estimated by s_d.
- Then the standardised mean difference follows a t-distribution.

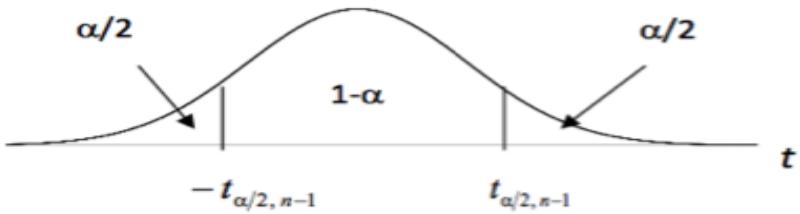
$$t = \frac{\overline{d} - \mu_D}{s_d/\sqrt{n}} \sim t_{n-1}$$

Confidence Interval Estimation - Paired Samples

- Since the standardised mean difference, $\overline{d}$ follows a t-distribution, and

$$P\left(-t_{\alpha/2,n-1} < t < t_{\alpha/2,n-1}\right) = 1-\alpha$$

$$\Rightarrow P\left(-t_{\alpha/2,n-1} < \frac{\overline{d}-\mu_D}{s_d/\sqrt{n}} < t_{\alpha/2,n-1}\right) = 1-\alpha,$$

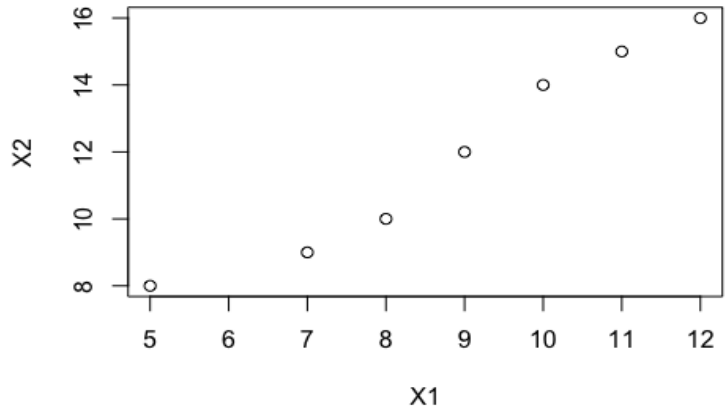


it can be shown that on making μ_D the subject of the formula:

- A $100(1-\alpha)\%$ Confidence Interval estimator of μ_D is

$$\overline{d} \pm t_{\alpha/2,n-1}\frac{s_d}{\sqrt{n}}$$

- In general, the paired-sample procedure is appropriate when
- the samples are naturally paired.
- there is a reasonably large positive correlation between the pairs.



- In this case the paired-sample procedure makes more efficient use of the data and generally narrows the confidence interval.

Hypothesis Testing: Paired Samples

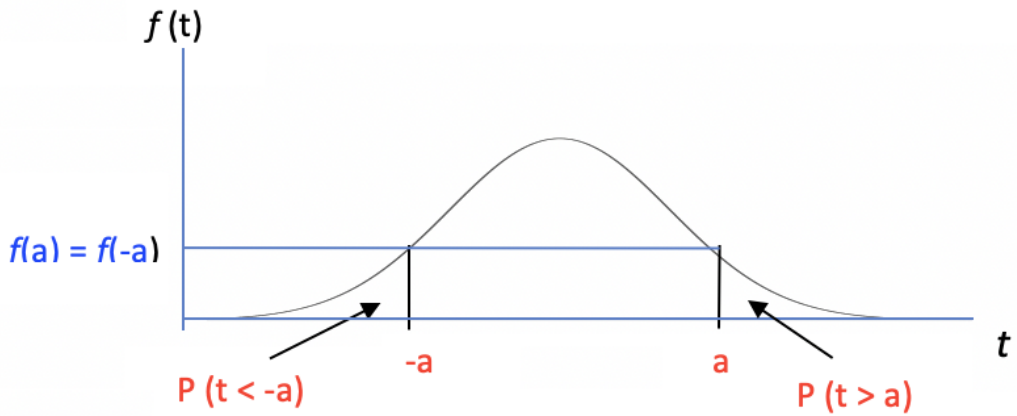
Table: 1

Null Hypothesis	$H_0 : \mu_D = D_0$	$H_0 : \mu_D = D_0$ or $H_0 : \mu_D \leq D_0$	$H_0 : \mu_D = D_0$ or $H_0 : \mu_D \geq D_0$
Alternative Hypothesis	$H_1 : \mu_D \neq D_0$	$H_1 : \mu_D > D_0$	$H_1 : \mu_D < D_0$
Level of Significance	α	α	α
Test Statistic	$t = \frac{\bar{d} - \mu_D}{s_d/\sqrt{n}}$	$t = \frac{\bar{d} - \mu_D}{s_d/\sqrt{n}}$	$t = \frac{\bar{d} - \mu_D}{s_d/\sqrt{n}}$
P-Value Approach Reject H_0 if	$p - value < \alpha$	$p - value < \alpha$	$p - value < \alpha$

P-value approach - Paired samples

$H_0 : \mu_D = D_0$
$H_1 : \mu_D \neq D_0$

$$t = \frac{\overline{d} - \mu_D}{s_d/\sqrt{n}} \sim t_{n-1}$$



a = |numerical value of the test statistic|

$f(t)$ is the density function of the t-distribution distribution and $f(a) = f(-a)$

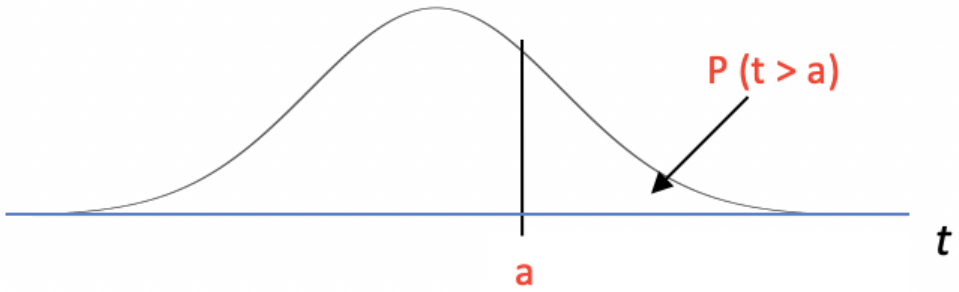
p-value = P(t > a) + P(t < -a)

Reject H_0 if $p - value < \alpha$

P-value approach - Paired samples

$H_0 : \mu_D = D_0$ or $H_0 : \mu_D \leq D_0$
$H_1 : \mu_D > D_0$

$$t = \frac{\overline{d} - \mu_D}{s_d/\sqrt{n}} \sim t_{n-1}$$



a = numerical value of the test statistic

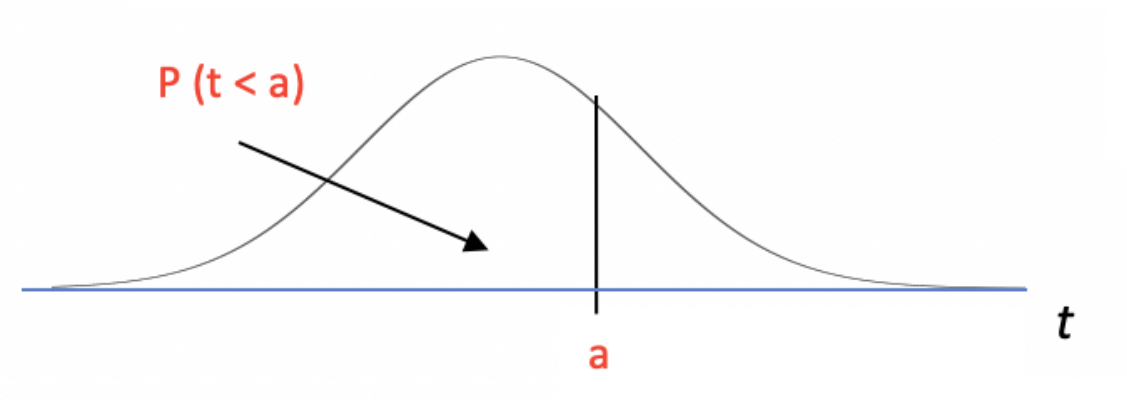
p-value = P(t > a)

Reject H_0 if $p - value < \alpha$

P-value approach - Paired samples

$H_0 : \mu_D = D_0$ or $H_0 : \mu_D \geq D_0$
$H_1 : \mu_D < D_0$

$$t = \frac{\overline{d} - \mu_D}{s_d/\sqrt{n}} \sim t_{n-1}$$



a = numerical value of the test statistic

p-value = P(t < a)

Reject H_0 if $p - value < \alpha$

Example 2

The managing partner of a major consulting firm is trying to assess the effectiveness of extensive computer skills training given to all new entry-level professionals. She administers a computer skills test to each of 40 randomly chosen employees, immediately before and after the training program. The pre-training and post-training scores of these 40 individuals are recorded and are given in the file *skills.csv*. Assume that the difference in the post-training and pre-training scores is normally distributed.

1. Using a 5% level of significance, test if the given sample data supports the claim that the firm's training program is increasing the new employees' computing skills.
2. Obtain the 95% confidence interval estimate of the mean difference between the post-training and pre-training skills for all employees.
3. Perform diagnostic checks for outliers, and to assess the validity of the assumption of normality of the difference between the post-training and pre-training scores.

Example 2 - Part 1

1. Hypotheses:
   $H_0 : \mu_D = 0$
   $H_1 : \mu_D > 0$
2. Level of significance: $\alpha = 0.05$
3. Test statistic: t $= 7.495$
4. p-value ≈ 0.000
5. Decision Rule: Reject H_0 if p-value < 0.05
6. Conclusion: Since p-value < 0.05, reject H_0 at the 5% level or indeed any reasonable level of significance. There is sufficient evidence to conclude that the firm's training is increasing the new employees' computing skills.

D $= X_1 - X_2$
$X_1 =$ Post-training test scores
$X_2 =$ Pre-training test scores

```
> #t-test
> t.test(x,y, alternative="greater", paired = TRUE)

        Paired t-test

data:  x and y
t = 7.4946, df = 39, p-value = 2.265e-09
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 9.922401      Inf
sample estimates:
mean of the differences
                  12.8
```

Example 2 - Part 2

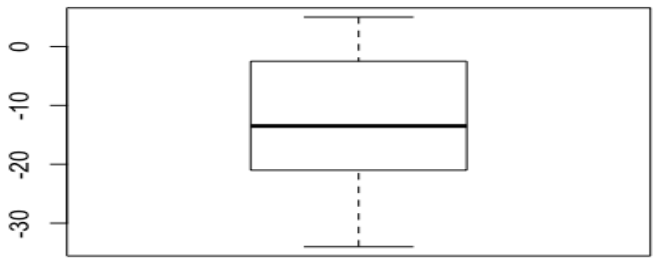
Note that the 95% confidence interval derives from the fact that approximately 95% of intervals, computed in this manner for repeated samples of sizes n will cover the true mean difference μ_D.

- The 95% confidence interval estimate of the difference in population means, μ_D is $[9.345, 16.255]$
- We conclude with 95% confidence, that the mean difference between the post-training and pre-training test scores is at least 9.35 units and can be as much as 16.26 units.
- Because the confidence interval does not contain zero, the null hypothesis $H_0 : \mu_D = 0$ is rejected in favour of the alternative $H_1 : \mu_D \neq 0$.

```
> # this will enable us to get a 95% two-sided confidence iterval
> t.test(x,y, alternative="two.sided", paired = TRUE)

        Paired t-test

data:  x and y
t = 7.4946, df = 39, p-value = 4.531e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  9.345446 16.254554
sample estimates:
mean of the differences
                  12.8
```

Example 2 - Part 3

```
> # difference between post-training and pre-training scores
> d <- y-x
> #check for outliers
> boxplot(d)
> #assess normality assumptions of x and y
> qqnorm(d, ylab = "d Sample Quantiles")
> shapiro.test(x)

        Shapiro-Wilk normality test

data:  x
W = 0.97257, p-value = 0.4324
```



- From the boxplot, we observe that the difference between the post-training and pre-training test scores do not display any outliers.
- From the Normal QQ-plot as well as from the results of the Shapiro-Wilk normality test, we conclude that the assumption of normality of the difference between the post-training and pre-training test scores is valid.