Test of Hypothesis and ANOVA using RStudio

Course Taught at SUAD

Tanujit Chakraborty & Madhurima Panja (TA)

MDA Course @ Sorbonne

Code Link: https://github.com/tanujit123/MATH-260

What we quest to achieve through the sessions

- Testing of Hypothesis
- Analysis of Variance (ANOVA)

t-tests for Testing of Hypothesis

In this section, we shall illustrate the usage of performing hypothesis testing using R. This includes the t-tests –

- one sample Student's t-test,
- two-sample t-test,
- pooled t-test
- paired t-test.

Other tests include the large sample tests for proportion(s) and testing of variances.

One sample student's t-test

We start with a random sample $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$. The null hypothesis is given by

$$H_0: \mu \leq \mu_0 \; vs \; H_1: \mu > \mu_0$$

We first generate some simulated data with true value of $\mu = 2$ and $\sigma^2 = 4$, and sample size $n = 25$. We take the null value $\mu_0 = 3$.

```
set.seed(8885)
Xsamp <- rnorm(n = 25, mean = 2, sd = 2)
```

Note: This is a simulated scenario, where we actually are aware of the true value of the parameters. In the real-life scenario, we would have no idea about the true value of the parameters, and our goal would be to test the population mean using the data given. This exercise is to illustrate the t-test procedure, and check whether we are getting desired results or not.
The function that we are going to use is **t.test**.

```
mytest <- t.test(Xsamp, alternative = "greater", mu = 3)
mytest
```

```
##  One Sample t-test
## data:  Xsamp
## t = -1.3467, df = 24, p-value = 0.9047
## alternative hypothesis: true mean is greater than 3
## 95 percent confidence interval:
##  1.954146      Inf
## sample estimates:
## mean of x
##   2.53935
```

One sample student's t-test

To specifically extract the p-value of the test, we write

```
mytest$p.value
```
```
## [1] 0.9046778
```

As we would have expected, the p-value of this test came out to be quite high, since the true value of μ is, in fact, equal to 2. Our data didn't provide evidence that the true mean is greater than 3. Of course, if we had changed our null value to something closer to 2, it would be harder to distinguish. Let's check this.

```
t.test(Xsamp, alternative = "greater", mu = 2.1)
```
```
##  One Sample t-test
## data:  Xsamp
## t = 1.2845, df = 24, p-value = 0.1056
## alternative hypothesis: true mean is greater than 2.1
## 95 percent confidence interval:
##  1.954146      Inf
## sample estimates:
## mean of x
##   2.53935
```

Notice the drop in the p-value of the test.

One sample student's t-test

Can you argue why the p-value has dropped?

Take another case.

```
t.test(Xsamp, alternative = "greater", mu = 1.9)
##  One Sample t-test
## data:  Xsamp
## t = 1.8692, df = 24, p-value = 0.03692
## alternative hypothesis: true mean is greater than 1.9
## 95 percent confidence interval:
##  1.954146      Inf
## sample estimates:
## mean of x
##   2.53935
```

The p-value has dropped below 0.05 now, and if we are using a level $\alpha = 0.05$, we would reject H_0 in favour of H_1 (and we would have taken the correct decision in this case). We would still fail to reject it at $\alpha = 0.01$. Taking the null value away from the actual true value towards the region supported by H_0 would result in lower p-values. Let's check.

One sample student's t-test

```
t.test(Xsamp, alternative = "greater", mu = 1)
##  One Sample t-test
## data:  Xsamp
## t = 4.5004, df = 24, p-value = 7.398e-05
## alternative hypothesis: true mean is greater than 1
## 95 percent confidence interval:
##  1.954146      Inf
## sample estimates:
## mean of x
##   2.53935
```

We can actually find the p-value as a function of μ_0 and the sample size n, keeping the true values of the parameters fixed. In R, we can quickly write a function to do this.

```
func.p.val <- function(n, mu0){
  set.seed(100)
  Xsamp <- rnorm(n, mean = 2, sd = 2)          # generating random sample of size n from N(2,4)
  p.val <- t.test(Xsamp, alternative = "greater", mu = mu0)$p.value          # calculate p-value
  return(p.val)
}
n <- seq(5,25, by = 5)
mu0 <- seq(-5,5, by = 0.1)
pvals <- matrix(0, nrow = length(n), ncol = length(mu0))
```

One sample student's t-test

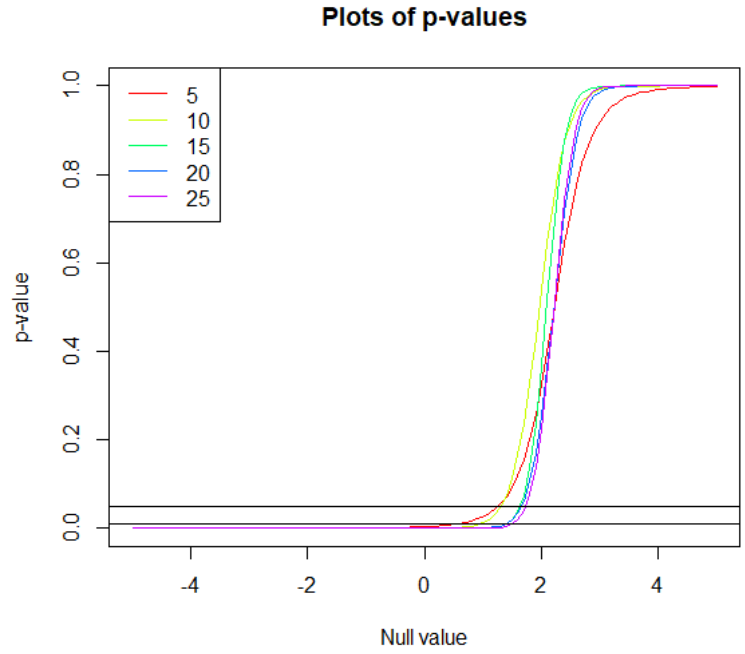
```
for(i in 1:length(n)){
  for(j in 1:length(mu0)){
    pvals[i,j] <- func.p.val(n[i],mu0[j])
  }
}
```

Now let us plot the p-values for different values of the sample size and null values.

```
cl <- rainbow(length(n))

plot(mu0, pvals[1, ], type = "l", main = "Plots of p-values“,
     xlab = "Null value", ylab = "p-value",
     ylim = c(min(pvals),1), col = cl[1])

for(i in 2:length(n)){
  lines(mu0, pvals[i, ], type = "l", col = cl[i])
}
abline(h = 0.05)
abline(h = 0.01)
legend("topleft", legend = n, lty = 1, col = cl)
```



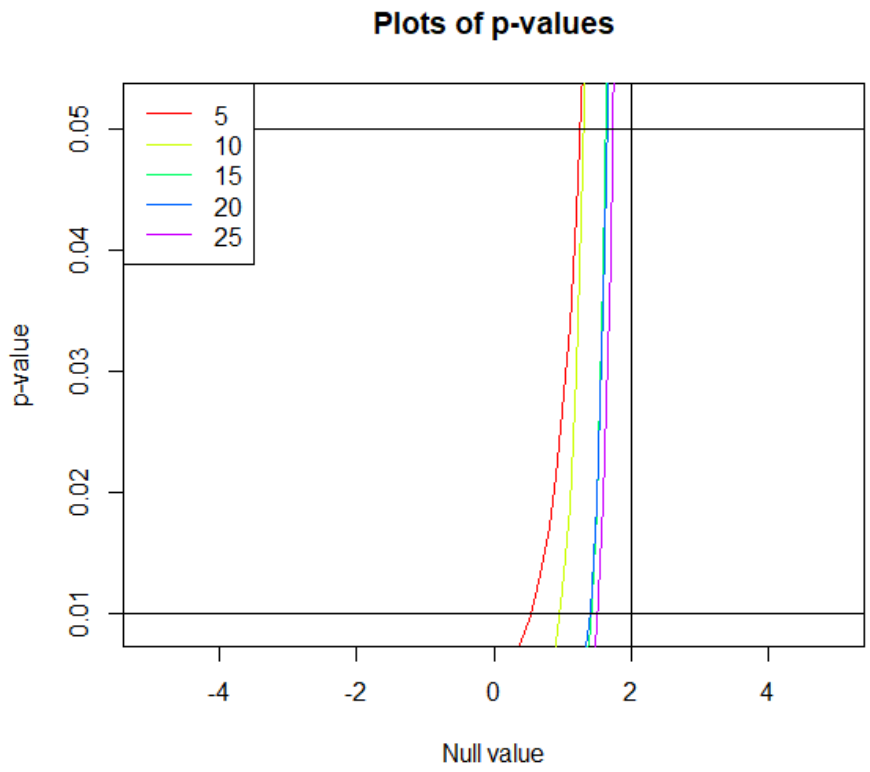
One sample student's t-test

Let us zoom it in near the true value of 2.

```
cl <- rainbow(length(n))

plot(mu0, pvals[1, ], type = "l", main = "Plots of p-values", xlab = "Null value",
     ylab = "p-value", ylim = c(0.009,0.052), col = cl[1])

for(i in 2:length(n)){
  lines(mu0, pvals[i, ], type = "l", col = cl[i])
}
abline(h = 0.05)
abline(h = 0.01)
abline(v = 2)
legend("topleft", legend = n, lty = 1, col = cl)
```



To work with other alternatives, you should change the alternative
to "lesser" for left-tailed tests, and
to "two-sided" for two-tailed tests.

Pro tip: Usage of "g","l","t" also works!

Two sample t-test

Behrens-Fisher Problem

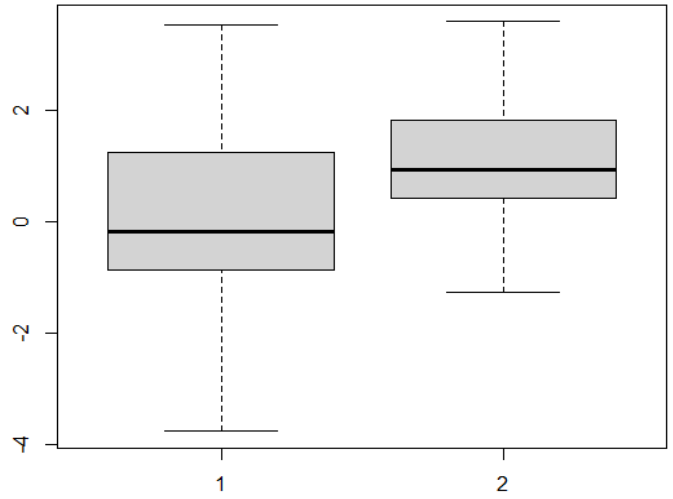
We start off with the Behrens-Fisher problem, which deals with the comparison of means of two Normal populations when the population variances are not equal. We shall be using the Satthertwaite approximation for this, and the corresponding test is called the Welsh t-test.

We shall illustrate this with a simulation.

Consider $X_1, X_2, \dots, X_n \sim N(\mu_1, \sigma_1^2)$, $Y_1, Y_2, \dots, Y_n \sim N(\mu_2, \sigma_2^2)$ and X and Y are independent.
We wish to test the hypothesis $H_0: \mu_1 - \mu_2 \leq \mu_0 \; vs \; H_1: \mu_1 - \mu_2 > \mu_0$.

We first generate some simulated data.

```
Xsamp <- rnorm(25, mean = 0, sd = 2)
Ysamp <- rnorm(20, mean = 1, sd = 1)
boxplot(Xsamp, Ysamp)
```



Two sample t-test

Time for t-testing. Take null value of difference to be zero.

```
mytest2 <- t.test(Xsamp, Ysamp, alternative = "g", mu = 0)
mytest2
##  Welch Two Sample t-test
## data:  Xsamp and Ysamp
## t = -2.0846, df = 41.581, p-value = 0.9784
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -1.747405       Inf
## sample estimates:
## mean of x mean of y
## 0.1096918 1.0767013
```

In case you had to test whether μ_2 is greater than μ1 by 0.5, we can do it in two different ways as follows.

```
t.test(Xsamp, Ysamp, alternative = "l", mu = -0.5)
##  Welch Two Sample t-test
## data:  Xsamp and Ysamp
## t = -1.0068, df = 41.581, p-value = 0.1599
## alternative hypothesis: true difference in means is less than -0.5
## 95 percent confidence interval:
##        -Inf -0.1866143
## sample estimates:
## mean of x mean of y
## 0.1096918 1.0767013
```

```
t.test(Ysamp, Xsamp, alternative = "g", mu = 0.5)
##  Welch Two Sample t-test

## data:  Ysamp and Xsamp
## t = 1.0068, df = 41.581, p-value = 0.1599
## alternative hypothesis: true difference in means is greater than 0.5
## 95 percent confidence interval:
##  0.1866143       Inf
## sample estimates:
## mean of x mean of y
## 1.0767013 0.1096918
```

Pooled t-test

Now we move on to pooled t-test. As you might recall, we should be using pooled t-testing procedure if the two population variances are equal.

We shall illustrate this with a simulation.

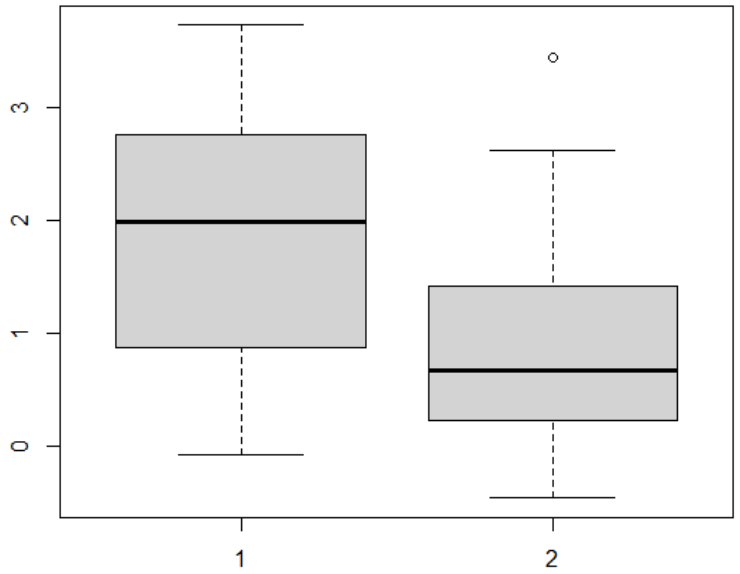
Consider $X_1, X_2, \ldots, X_n \sim N(\mu_1, \sigma^2)$, $Y_1, Y_2, \ldots, Y_n \sim N(\mu_2, \sigma^2)$ and X and Y are independent.

We wish to test the hypothesis $H_0: \mu_1 - \mu_2 \leq \mu_0$ vs $H_1: \mu_1 - \mu_2 > \mu_0$.

We first generate some simulated data.

```
Xsamp <- rnorm(20, mean = 2, sd = 1)
Ysamp <- rnorm(30, mean = 1, sd = 1)
boxplot(Xsamp, Ysamp)
```

You can see that we have used the same variance to generate our samples. Suppose the problem is to check whether the means differ, under the assumption that the variances are unknown but equal.



Pooled t-test

```
mytest.pooled <- t.test(Xsamp, Ysamp, alternative = "t", mu = 0, var.equal = TRUE)
mytest.pooled
```

```
##  Two Sample t-test
## data:  Xsamp and Ysamp
## t = 2.9634, df = 48, p-value = 0.004724
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2928261 1.5287200
## sample estimates:
## mean of x mean of y
## 1.8358720 0.9250989
```

Notice that we have used the additional argument of var.equal = TRUE to incorporate the additional information in our testing procedure.

Paired t-test

Now we move on to paired t-test for samples coming in pairs from a Bivariate Normal Distribution. We shall illustrate this with a simulation.

Consider $(X_1, Y_1), \ldots, (X_n, Y_n) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$.

We wish to test the hypothesis $H_0: \mu_1 - \mu_2 = \mu_0 \; vs \; H_1: \mu_1 - \mu_2 \neq \mu_0$.

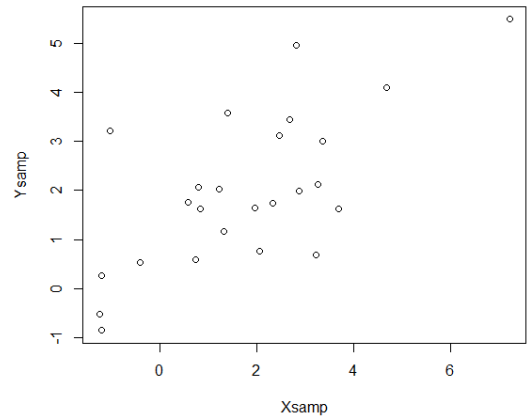
We first generate some simulated data.

Xsamp <- rnorm(25, mean = 2, sd = 2)

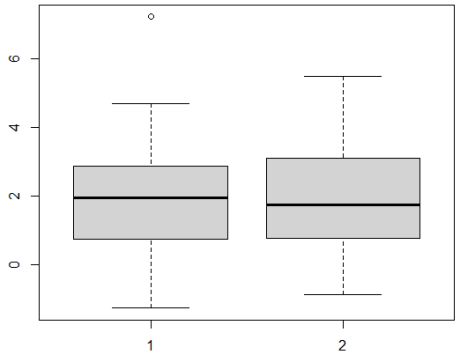
Ysamp <- 0.5*Xsamp + rnorm(25, mean = 1, sd = 1)

Notice that the Y samples are linearly related to X samples, and hence a correlation between them are established. You can plot the data to visualize this.

plot(Xsamp, Ysamp)



boxplot(Xsamp, Ysamp)



Paired t-test

The true means are $\mu_1 = 2$ and $\mu_2 = 0.5 * 2 + 1 = 2$, so that the means are, in fact, equal. We shall test this in light of the data.

```
mytest.paired <- t.test(Xsamp, Ysamp, alternative = "t", mu = 0, paired = TRUE)
mytest.paired
```

```
##  Paired t-test
## data:  Xsamp and Ysamp
## t = -0.89964, df = 24, p-value = 0.3773
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5820331  0.2286589
## sample estimates:
## mean of the differences
##              -0.1766871
```

Note the usage of the argument paired = TRUE in the command above.

Testing of variances

We now illustrate testing of variances of two independent Normal populations. Suppose

$$X_1, X_2, \dots, X_n \sim N(\mu_1, \sigma_1^2), \qquad Y_1, Y_2, \dots, Y_n \sim N(\mu_2, \sigma_2^2)$$

We wish to test the hypothesis concerning their respective variances. Let's consider testing

$$H_0: \sigma_1^2 \leq \sigma_2^2 vs\ H_1: \sigma_1^2 > \sigma_2^2$$

We use the command **var.test** for this.

```
Xsamp <- rnorm(15, mean = 2, sd = 2)
Ysamp <- rnorm(20, mean = 5, sd = 4)

vartest <- var.test(Xsamp, Ysamp, alternative = "g")
vartest
```

```
##  F test to compare two variances
## data:  Xsamp and Ysamp
## F = 0.36659, num df = 14, denom df = 19, p-value = 0.9696
## alternative hypothesis: true ratio of variances is greater than 1
## 95 percent confidence interval:
##  0.1625231       Inf
## sample estimates:
## ratio of variances
##          0.3665894
```

TEST OF HYPOTHESIS using R

TEST OF HYPOTHESIS

Introduction:

In many situations, it is required to accept or reject a statement or claim about some parameter

Example:

1. The average cycle time is less than 24 hours
2. The % rejection is only 1%

The statement is called the hypothesis

The procedure for decision making about the hypothesis is called hypothesis testing

Advantages

1.Handles uncertainty in decision making

2.Minimizes subjectivity in decision making

3.Helps to validate assumptions or verify conclusions

TEST OF HYPOTHESIS

Commonly used hypothesis tests on mean of normal distribution:

- Checking mean equal to a specified value ($mu = mu_0$)
- Two means are equal or not ($mu_1 = mu_2$)

Null Hypothesis:

A statement about the status quo

One of no difference or no effect

Denoted by H0

Alternative Hypothesis:

One in which some difference or effect is expected

Denoted by H1

TEST OF HYPOTHESIS

Methodology demo: To Test Mean = Specified Value ($mu = mu_0$)

Suppose we want to test whether mean of a process characteristic is 5 based on the following sample data from the process

4	4	5	5	6
5	4.5	6.5	6	5.5

Calculate the mean of the sample, xbar = 5.15

Compare xbar with specified value 5

or xbar - specified value = xbar - 5 with 0

If xbar - 5 is close to 0

then conclude mean = 5

else mean ≠ 5

TEST OF HYPOTHESIS

Methodology demo : To Test Mean = Specified Value (mu = mu_0)

Consider another set of sample data. Check whether mean of the process characteristic is 500

400	400	500	500	600
500	450	650	600	550

Mean of the sample, xbar = 515

xbar - 500 = 515 - 500 = 15

Can we conclude mean ≠ 500?

Conclusion:

Difficult to say mean = specified value by looking at xbar - specified value alone

TEST OF HYPOTHESIS

Methodology demo: To Test Mean = Specified Value ($mu = mu_0$)

Test statistic is calculated by dividing (xbar - specified value) by a function of standard deviation

To test Mean = Specified value

Test Statistic t_0 = (xbar - Specified value) / (SD / $\sqrt{n}$)

If test statistic is close to 0, conclude that Mean = Specified value

To check whether test statistic is close to 0, find out p value from the sampling distribution of test statistic

TEST OF HYPOTHESIS

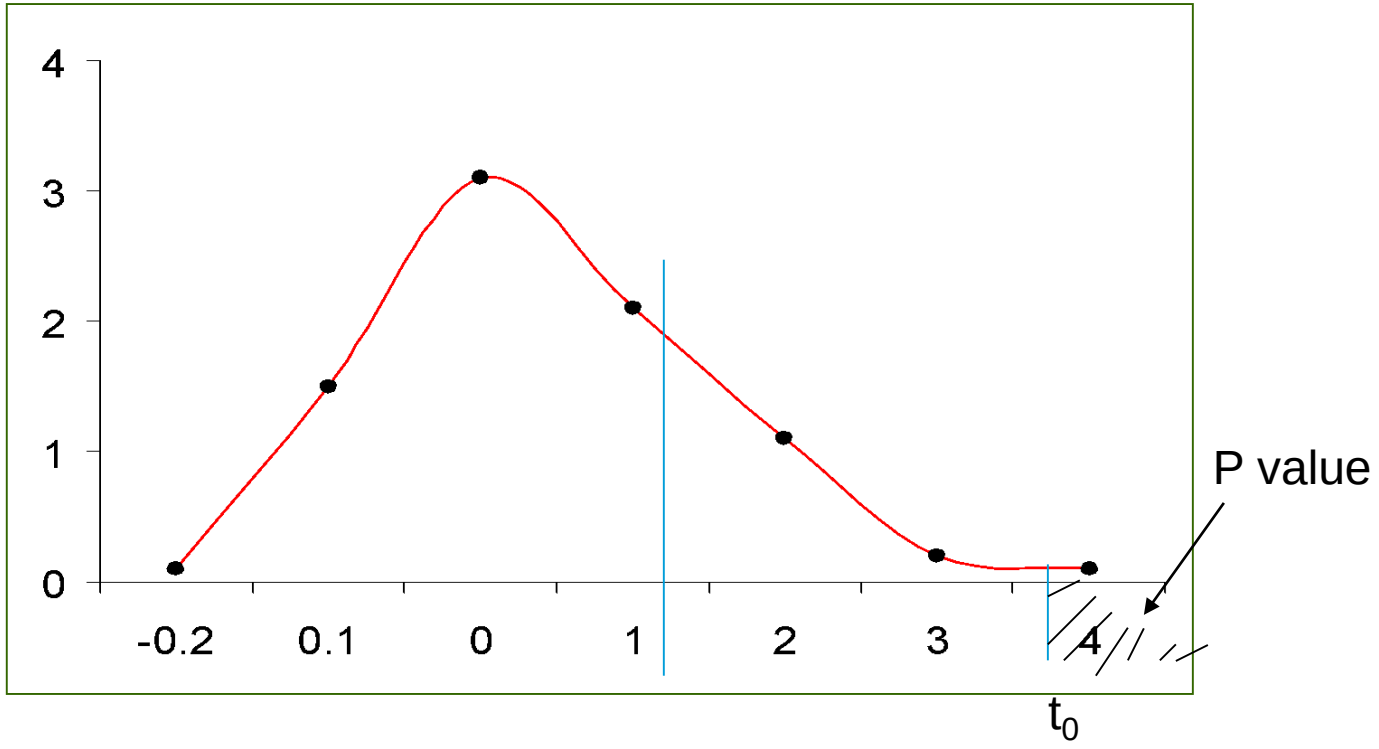
Methodology demo: To Test Mean = Specified Value

P value

The probability that such evidence or result will occur when H0 is true

Based on the reference distribution of test statistic

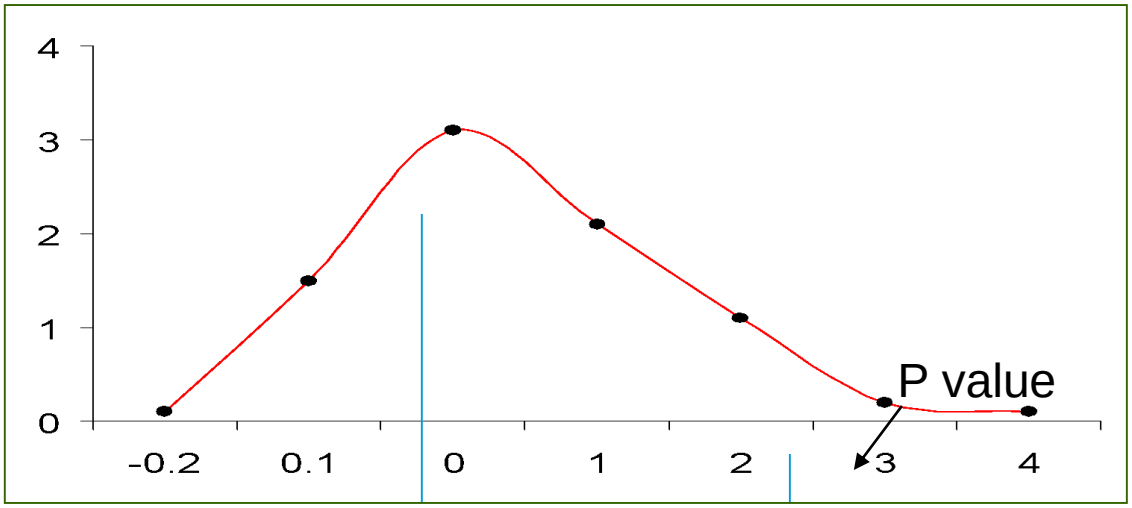
The tail area beyond the value of test statistic in reference distribution



TEST OF HYPOTHESIS

Methodology demo : To Test Mean = Specified Value

P value



If test statistic t_0 is close to 0 then p will be high

If test statistic t_0 is not close to 0 then p will be small

If p is small , $p < 0.05$ (with alpha = 0.05), conclude that $t \neq 0$, then

Mean ≠ Specified Value, H0 rejected

TEST OF HYPOTHESIS

To Test Mean = Specified Value (mu = mu_0)

Example: Suppose we want to test whether mean of the process characteristic is 5 based on the following sample data

4	4	5	5	6
5	4.5	6.5	6	5.5

H0: Mean = 5

H1: Mean ≠ 5

Calculate xbar = 5.15

SD = 0.8515

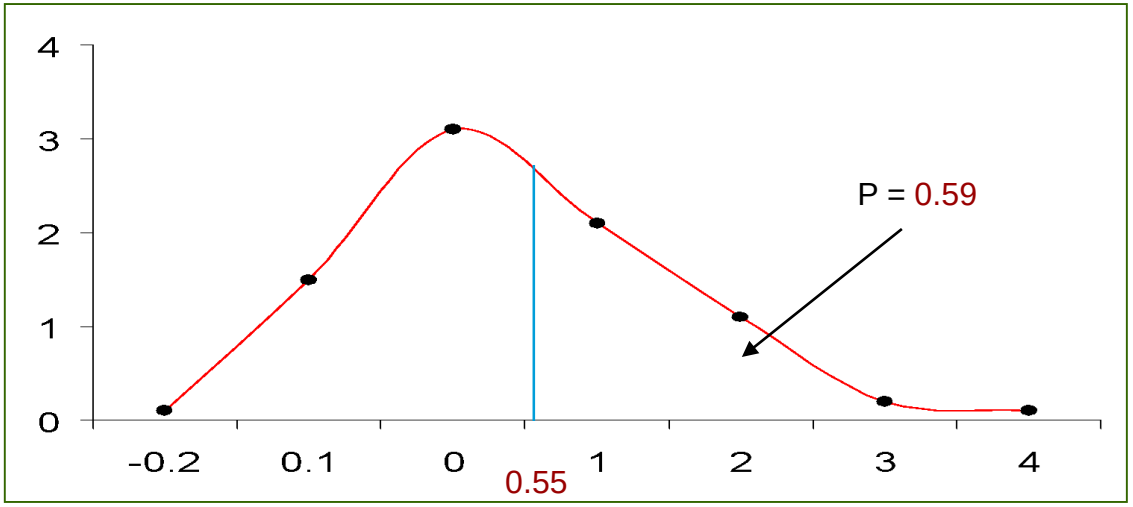
n = 10

Test statistic t_0 = (xbar - 5)/(SD / √n) = (5.15 - 5) / (0.8515 / √10) = 0.5571

TEST OF HYPOTHESIS

Example: To Test Mean = Specified Value (mu = mu_0)

$t_0 = 0.5571$



P ≥ 0.05, hence Mean = Specified value = 5.

H0: Mean = 5 is not rejected

TEST OF HYPOTHESIS

Hypothesis Testing: Steps

1.Formulate the null hypothesis H0 and the alternative hypothesis H1

2.Select an appropriate statistical test and the corresponding test statistic

3.Choose level of significance alpha (generally taken as 0.05)

4.Collect data and calculate the value of test statistic

5.Determine the probability associated with the test statistic under the null hypothesis using sampling distribution of the test statistic

6.Compare the probability associated with the test statistic with level of significance specified

TEST OF HYPOTHESIS

Install the necessary packages

```
> install.packages("car")
> library(car)
> install.packages("gplots")
> library(gplots)
> install.packages("ggplot2")
> library(ggplot2)
> install.packages("qqplotr")
> library(qqplotr)
> install.packages("boot")
> library(boot)
```

TEST OF HYPOTHESIS

One sample t test

Exercise 1 : A company claims that on an average it takes only 40 hours or less to process any purchase order. Based on the data given below, can you validate the claim? The data is given in PO_Processing.csv

Reading data to mydata

```
> mydata = read.csv('PO_Processing.csv',header = T,sep = ",")
> PT = mydata$Processing_Time
```

Performing one sample t test

```
> t.test(PT, alternative = 'greater', mu = 40)
```

Statistics	Value
t	3.7031
df	99
P value	0.0001753

NORMALITY TEST

NORMALITY TEST

Normality test

A methodology to check whether the characteristic under study is normally distributed or not

Two Methods :

Normality test - Quantile – Quantile (Q- Q) plot

Plots the ranked samples from the given distribution against a similar number of ranked quantiles taken from a normal distribution

If the sample is normally distributed then the line will be straight in the plot

Normality test – Shapiro – Wilk test

H0: Deviation from bell shape (normality) = 0

H1 : Deviation from bell shape ≠ 0

If p value ≥ 0.05 (5%), then H0 is not rejected, distribution is normal

NORMALITY TEST

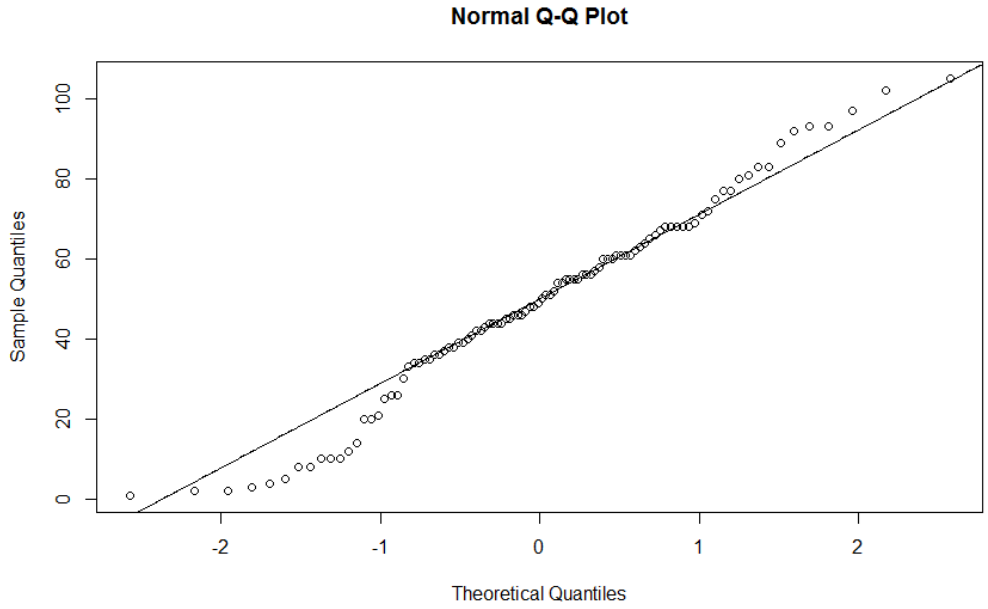
Normality test

Exercise 1 : The processing times of purchase orders is given in PO_Processing.csv. Is the distribution of processing time is normally distributed?

Normality Check using Normal Q – Q plot

> qqnorm(PT)

> qqline(PT)



NORMALITY TEST

Normality test

Exercise 1 : The processing times of purchase orders is given in PO_Processing.csv. Is the distribution of processing time being normally distributed?

Normality Check using Shapiro – Wilk test

```
> shapiro.test(PT)
```

Statistics	Value
W	0.9804
p value	0.1418

Conclusion: The data is Normal if p-value is above 0.05

ANALYSIS OF VARIANCE

One-way Analysis of Variance (One-way ANOVA)

The Analysis of Variance, or ANOVA in short, refers broadly to a collection of experimental situations and statistical procedures for the analysis of quantitative responses from experimental units. The simplest of them is referred to as a single-factor, or one-way ANOVA. It involves the analysis either of data sampled from more than two populations or of data from experiments in which more than two treatments have been used. The characteristic that differentiates the treatments or populations from one another is called the *factor* under study, and the different treatments or populations are referred to as the *levels* of the factor.

Some examples:

- An experiment to study the effect of different fertilizers on the yield of a crop.
- An experiment to study drug effectiveness on a disease.
- An experiment to study effect of different insecticides on pest control.

One-way Analysis of Variance (One-way ANOVA)

Consider the ‘InsectSprays’ dataset in R.

```
data("InsectSprays")
dat <- InsectSprays
head(InsectSprays)
```

```
##   count spray
## 1    10     A
## 2     7     A
## 3    20     A
## 4    14     A
## 5    14     A
## 6    12     A
```

```
str(InsectSprays)
```

```
## 'data.frame':    72 obs. of  2 variables:
##  $ count: num  10 7 20 14 14 12 10 23 17 20 ...
##  $ spray: Factor w/ 6 levels "A","B","C","D",..: 1 1 1 1 1 1 1 1 1 1 ...
```

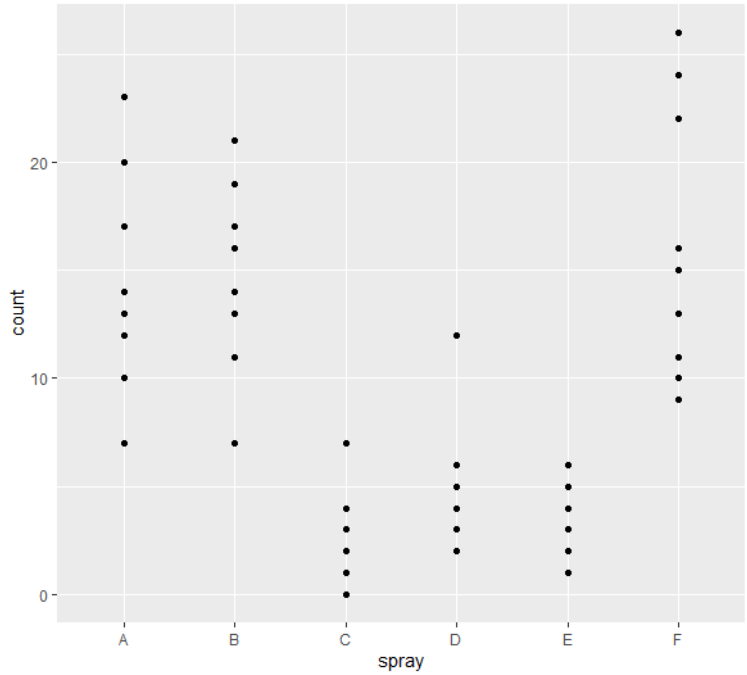
Show the levels of treatment

```
levels(dat$spray)
```

```
## [1] "A" "B" "C" "D" "E" "F"
```

Let us visualize the data.

```
library(ggplot2)
ggplot(dat, aes(x=spray,y=count))+ geom_point()
```



One-way Analysis of Variance (One-way ANOVA)

Single factor ANOVA focuses on a comparison of more than two populations or treatment means. Let

- $I =$ number of treatment levels
- $J =$ number of observations in level $\mathrm{i} = 1,2,\ldots,I$
- $\mu_i =$ mean of treatment level $\mathrm{i} = 1,2,\ldots,I$
- $N = \sum_{i=1}^{I} J_i$ total number of observations

The relevant hypotheses are

$$H_0 : \mu_1 = \ldots = \mu_I \; vs. H_1 : not(H_0)$$

The alternative hypothesis is tantamount to saying that at least one pair of means are different.

The ANOVA Model

Let $X_{i,j}$ denote the random variable representing the j^{th} measurement taken from the i^{th} treatment, and $x_{i,j}$ be the observed value of the same.
The one-way ANOVA model is given by

$$X_{i,j} = \mu_i + \in_{i,j}$$

where $\in_{i,j}$ are the error terms. We assume that $\in_{i,j} \sim N(0, \sigma^2)$.
This gives,

$$E(X_{i,j}) = \mu_i, \qquad Var(X_{i,j}) = \sigma^2$$

An alternative description of the one-way ANOVA is given by

$$X_{i,j} = \mu + \alpha_i + \in_{i,j}$$

where $\mu = \frac{1}{I}\sum_{i=1}^{I} \mu_i$, and $\alpha_i = \mu_i - \mu$; $i = 1,\ldots,I$. Note $\sum_{i=1}^{I} \alpha_i = 0$.

One-way Analysis of Variance (One-way ANOVA)

The null hypothesis above thus becomes

$$H_0 : \alpha_1 = \cdots = \alpha_I = 0 \; vs. H_1 : not(H_0)$$

The individual sample means are denoted as $\bar{X}_{1o}, \ldots, \bar{X}_{Io}$, such that $\bar{X}_{i0} = \frac{1}{J_i}\sum_{j=1}^{J_i} X_{i,j}$ and the grand mean is denoted as $\bar{X}_{00} = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J_i} X_{i,j}$.

Think about the total variation in the data. The observed values of the variable of interest are $X_{i,j}$, and the grand mean is $\bar{X}_{00}$. Thus, the total variation in the data is given by the Total Sum of Squares (SST), defined as

$$SST = \sum_{i=1}^{I}\sum_{j=1}^{J_i}\left(X_{i,j} - \bar{X}_{00}\right)^2$$

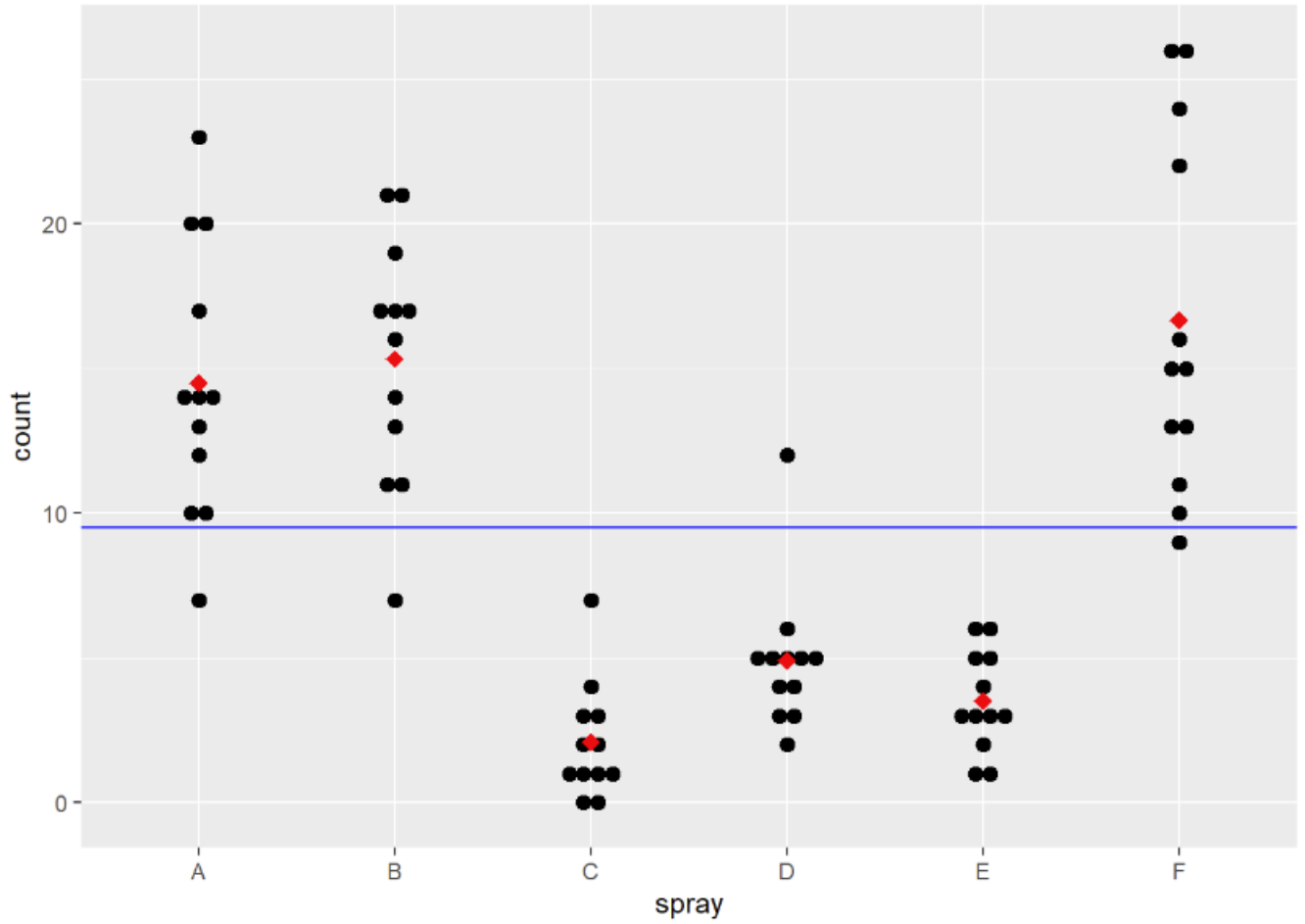
The total SS can be partitioned in to two sums, as $SST = SSTr + SSE$, where

$$SSTr = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(\bar{X}_{i0} - \bar{X}_{00})^2 \; and \; SSE = \sum_{i=1}^{I}\sum_{j=1}^{J_i}\left(X_{i,j} - \bar{X}_{i0}\right)^2$$

The above identity says that the total variation can be partitioned into two parts.
$SSTr$ measures the variation (between levels) that can be explained by possible differences in the μ_i (How would $SSTr$ behave if all the μ_i were identical to each other?)
On the other hand, SSE measures variation (within levels) that would be present irrespective of whether H_0 is true or false.

One-way Analysis of Variance (One-way ANOVA)

Let us look at our 'InsectSprays' data example again.



One-way Analysis of Variance (One-way ANOVA)

The red dots are the within level sample means $\bar{X}_{i0}$ and the blue line corresponds to the overall grand mean $\bar{X}_{00}$. $SSTr$ looks at the variation between levels, taking the squared differences of the red dot with the blue line, whereas SSE looks at the variation within levels, taking squared differences of the black dots with the red dots for each level.

If the null hypothesis is true, then, $SSTr$ would have a significantly smaller contribution to the total variation SST. This intuitive idea forms the basis of ANOVA. You must be wondering why a comparison of means is coined as analysis of variance. To answer this, let us look into the theoretical properties of the quantities $SSTr$ and SSE.
In this context, we define the Mean Squared Treatment ($MSTr$) and Mean Squared Error (MSE) as

$$MSTr = \frac{SSTr}{(I-1)}, \qquad MSE = \frac{SSE}{(N-1)}$$

Note that, if we denote the sample variance for the i^{th} treatment level as S_i^2, then,

$$MSE = \frac{(J_1-1)S_1^2 + (J_2-1)S_2^2 + \cdots + (J_I-1)S_I^2}{(J_1-1) + (J_2-1) + \cdots + + \cdots + (J_I-1)}$$

We can prove that,

$$E(SSTr) = (I-1)\sigma^2 + \sum_{i=1}^{I} J_i\alpha_i^2\,, E(SSE) = (N-1)\sigma^2$$

so that,

$$E(MSTr) = \sigma^2 + \frac{1}{(I-1)}\sum_{i=1}^{I} J_i\alpha_i^2\,, E(MSE) = \sigma^2$$

One-way Analysis of Variance (One-way ANOVA)

If H_0 is true, then, $\alpha_i = 0; i = 1, 2, \ldots, I$ and thus,

$$E(MSTr) = E(MSE) = \sigma^2$$

On the other hand, if H_0 is not true, then, $E(MSTr) > E(MSE) = \sigma^2$.

Denoting $E(MSTr)$ as σ_*^2, we can reformulate the testing of hypothesis problem as

$$H_0 : \sigma_*^2 = \sigma^2 \; vs. \; H_1 : \sigma_*^2 > \sigma^2$$

Thus, the hypothesis of testing of means has boiled down to testing of variances. Also, it is a right-tailed test.

The test statistic in this case is given by $F = \frac{MSTr}{MSE}$. We reject H_0 at level of significance α if observed

$$F > F_{\alpha, I-1, N-1}$$

We illustrate the ANOVA testing procedure in **R**. We shall use the *aov* function in **R**, and the *summary*() command to get the **ANOVA Table**.

One-way Analysis of Variance (One-way ANOVA)

```
anova.fit <- aov(count ~ spray, data = dat)
summary(anova.fit)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## spray        5   2669   533.8    34.7 <2e-16 ***
## Residuals   66   1015    15.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple comparisons in ANOVA

Our analysis is terminated if we fail to reject the null hypothesis, owing to the fact that there are no differences in the treatment means across levels. However, if the null hypothesis is rejected, the next step would be to determine which pairs of treatment means differ. We can either perform pairwise testing one at a time, or a multiple comparisons procedure. In the former case, our hypotheses look like

$$H_0 : \mu_r - \mu_s = 0 \; vs. \; H_1 : \mu_r - \mu_s \neq 0$$

We can adopt a two-sample t-test procedure for carrying out the above test. It will be a pooled t-test (why?), with the estimator of σ^2 given by MSE. The $100(1-\alpha)\%$ confidence interval for $\mu_r - \mu_s$ is given by

$$\bar{X}_{r0} - \bar{X}_{s0} \pm t_{\alpha,N-1} \sqrt{MSE(1/J_r + 1/J_s)}$$

The *summary.lm* command gives us the level-specific estimates and significance results.

Multiple comparisons in ANOVA

summary.lm(anova.fit)

```
## Call:
## aov(formula = count ~ spray, data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.333 -1.958 -0.500  1.667  9.333
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  14.5000     1.1322  12.807  < 2e-16 ***
## sprayB        0.8333     1.6011   0.520    0.604
## sprayC      -12.4167     1.6011  -7.755 7.27e-11 ***
## sprayD       -9.5833     1.6011  -5.985 9.82e-08 ***
## sprayE      -11.0000     1.6011  -6.870 2.75e-09 ***
## sprayF        2.1667     1.6011   1.353    0.181
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.922 on 66 degrees of freedom
## Multiple R-squared:  0.7244, Adjusted R-squared:  0.7036
## F-statistic:  34.7 on 5 and 66 DF,  p-value: < 2.2e-16
```

Interpretation of the R Output

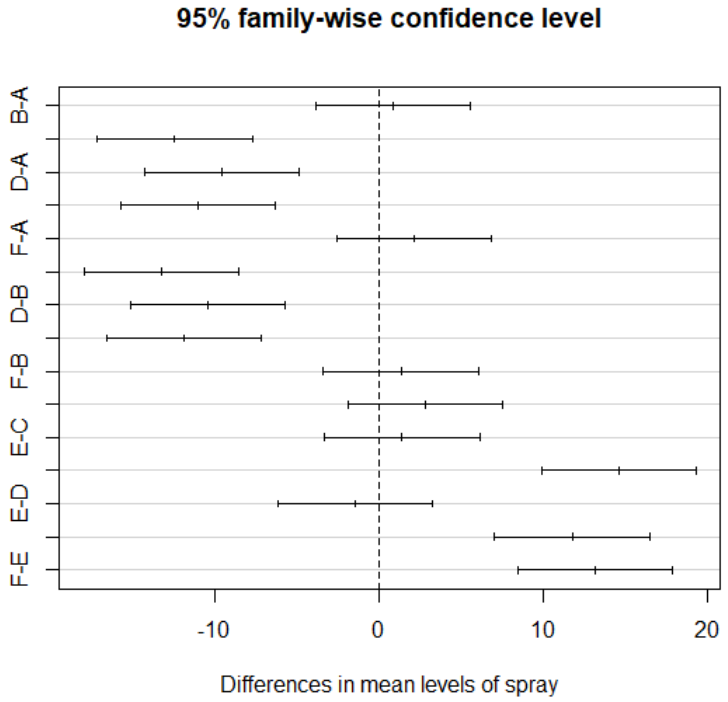
- The ANOVA Table (obtained by the *summary()* command) is the standard ANOVA table.
- Interesting thing is the one obtained using the *summary.lm* command.
- Look at the 'Coefficients' table. The first row (Intercept) corresponds to the baseline control group (Spray A in this case). The value of 14.5 is the mean of the baseline control group and the corresponding t-test tests whether the effect of that spray is zero or not. In this example, the null hypothesis of the mean effect for the control group = 0 is rejected owing to the extremely low p-value. The next rows correspond to the difference of the other levels of treatment with the control group (like Spray B - Spray A, Spray C - Spray A, etc.). The estimates reported are that of the respective differences of the other treatment levels with the control group.
- Notice that the standard errors are the same across all the differences, and this is due to the fact that the expression for the standard error involves the MSE with equal number of observations per group.
- The corresponding t-tests test for the difference of the treatment levels with the control group.
- Now look at the Residual standard error in the bottom of the table. It can be obtained from the ANOVA table above by taking the square root of the MSE.
- Also, the Multiple-R-squared R^2 value is obtained as $SSTr/(SSTr + SSE)$.

Tukey's HSD Method

However, the above method provides Confidence intervals at the desired level for individual pairs of means, but not a simultaneous one which controls the overall confidence level for all the pairs. To answer this, we resort to the multiple comparisons procedure. The intervals are based on **Studentized range statistic**, Tukey's Honest Significant Difference method.

TukeyHSD(anova.fit)

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
## Fit: aov(formula = count ~ spray, data = dat)
## $spray
##           diff       lwr       upr     p adj
## B-A   0.8333333  -3.866075  5.532742 0.9951810
## C-A -12.4166667 -17.116075 -7.717258 0.0000000
## D-A  -9.5833333 -14.282742 -4.883925 0.0000014
## E-A -11.0000000 -15.699409 -6.300591 0.0000000
## F-A   2.1666667  -2.532742  6.866075 0.7542147
## C-B -13.2500000 -17.949409 -8.550591 0.0000000
## D-B -10.4166667 -15.116075 -5.717258 0.0000002
## E-B -11.8333333 -16.532742 -7.133925 0.0000000
## F-B   1.3333333  -3.366075  6.032742 0.9603075
## D-C   2.8333333  -1.866075  7.532742 0.4920707
## E-C   1.4166667  -3.282742  6.116075 0.9488669
## F-C  14.5833333   9.883925 19.282742 0.0000000
## E-D  -1.4166667  -6.116075  3.282742 0.9488669
## F-D  11.7500000   7.050591 16.449409 0.0000000
## F-E  13.1666667   8.467258 17.866075 0.0000000
```



plot(TukeyHSD(anova.fit))

ANALYSIS OF VARIANCE

ANOVA

Analysis of Variance is a test of means for two or more populations

Partitions the total variability in the variable under study to different components

H0 = $Mean_1 = Mean_2 = - - - = Mean_k$

Reject H0 if p – value < 0.05

Example:

To study location of shelf on sales revenue

ANALYSIS OF VARIANCE

One Way ANOVA : Example

An electronics and home appliance chain suspect the location of shelves where television sets are kept will influence the sales revenue. The data on sales revenue in lakhs from the television sets when they are kept at different locations inside the store are given in sales revenue data file. The location is denoted as 1:front, 2: middle & 3: rear. Verify the doubt? The data is given in Sales_Revenue_Anova.csv.

Factor: Location(A)

Levels : front, middle, rear

Response: Sales revenue

ANALYSIS OF VARIANCE

One Way Anova : Example

Step 1: Calculate the sum, average and number of response values for each level of the factor (location).

Level 1 Sum(A_1):

Sum of all response values when location is at level 1 (front)

= 1.55 + 2.36 + 1.84 + 1.72

= 7.47

nA_1: Number of response values with location is at level 1 (front)

= 4

ANALYSIS OF VARIANCE

One Way Anova : Example

Step 1: Calculate the sum, average and number of response values for each level of the factor (location).

Level 1 Average:

Sum of all response values when location is at level 1 / number of response values with location is at level 1

= A_1 / nA_1 = 7.47 / 4 = 1.87

ANALYSIS OF VARIANCE

One Way Anova : Example

Step 1: Calculate the sum, average and number of response values for each level of the factor (location).

	Level 1 (front)	Level 2 (middle)	Level 3 (rear)
Sum	A_1: 7.47	A_2: 30.31	A_3: 15.55
Number	nA_1: 4	nA_2: 8	nA_3: 6
Average	1.87	3.79	2.59

ANALYSIS OF VARIANCE

One Way Anova : Example

Step 2: Calculate the grand total (T)

T = Sum of all the response values

= 1.55 + 2.36 + - - - + 2.72 + 2.07 = 53.33

Step 3: Calculate the total number of response values (N)

N = 18

Step 4: Calculate the Correction Factor (CF)

CF = (Grand Total)^2 / Number of Response values

= T^2 / N = $(537.33)^2$ / 18 = 158.0049

ANALYSIS OF VARIANCE

One Way Anova : Example

Step 5: Calculate the Total Sum of Squares (TSS)

TSS = Sum of square of all the response values - CF

$= 1.55^2 + 2.36^2 + - - - + 2.72^2 + 2.07^2 – 158.0049$

$= 15.2182$

Step 6: Calculate the between (factor) sum of square

$SS_A = A_1^2 / nA_1 + A_2^2 / nA_2 + A_3^2 / nA_3 - CF$

$= 7.47^2 / 4 + 30.31^2 / 8 + 15.55^2 / 4 – 158.0049$

$= 11.0827$

Step 7: Calculate the within (error) sum of square

SS_e = Total sum of square – between sum of square

$= TSS - SS_A = 15.2182 – 11.0827 = 4.1354$

ANALYSIS OF VARIANCE

One Way Anova : Example

Step 8: Calculate degrees of freedom (df)

Total df = Total Number of response values - 1

= 18 - 1 = 17

Between df

= Number of levels of the factor - 1

= 3 - 1 = 2

Within df = Total df – Between df

= 17 - 2 = 15

ANALYSIS OF VARIANCE

One Way Anova : R Code

Reading data and variables to R

```
> mydata = read.csv('Sales_Revenue_Anova.csv',header = T,sep = ",")
> location = mydata$Location
> revenue = mydata$Sales.Revenue
```

Converting location to factor

```
> location = factor(location)
```

Computing ANOVA table

```
> fit = aov(revenue  ~  location)
> summary(fit)
```

ANALYSIS OF VARIANCE

One Way Anova : Example

Anova Table:

Source	df	SS	MS	F	F Crit	P value
location	2	11.08272	5.541358	20.09949	3.68	0.0000
Residuals	15	4.135446	0.275696			
Total	17	15.21816				

MS = SS / df

$F = MS_{Between} / MS_{Within}$

F Crit =finv (probability, between df, within df) , probability = 0.05

P value = fdist (F, between df, within df)

ANALYSIS OF VARIANCE

One Way Anova : Decision Rule

If p value < 0.05, then

The factor has significant effect on the process output or response.

Meaning:

When the factor is changed from 1 level to another level,

there will be significant change in the response.

ANALYSIS OF VARIANCE

One Way Anova : Example Result

For factor Location, $p = 0.000 < 0.05$

Conclusion:

Location has significant effect on sales revenue

Meaning:

The sales revenue is not same for different locations like front, middle & rear

ANALYSIS OF VARIANCE

One Way Anova : Example Result

The expected sales revenue for different location under study is equal to level averages.

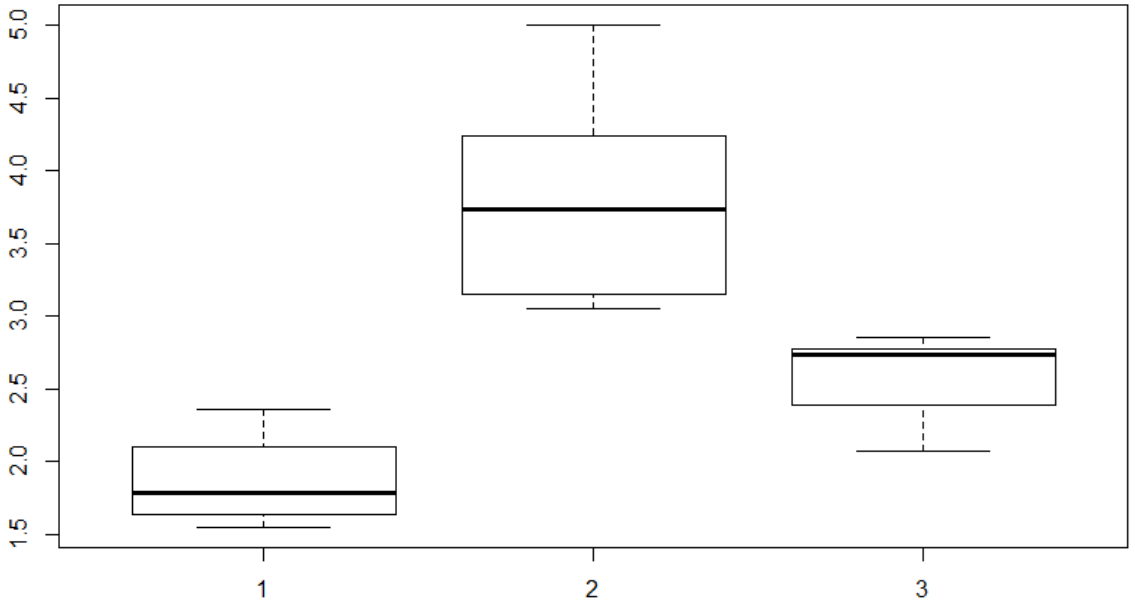
Location	Expected Sales Revenue
Front	1.8675
Middle	3.78875
rear	2.591667

```
> aggregate(revenue ~ location, FUN = mean)
```

ANALYSIS OF VARIANCE

One Way Anova : Example Result

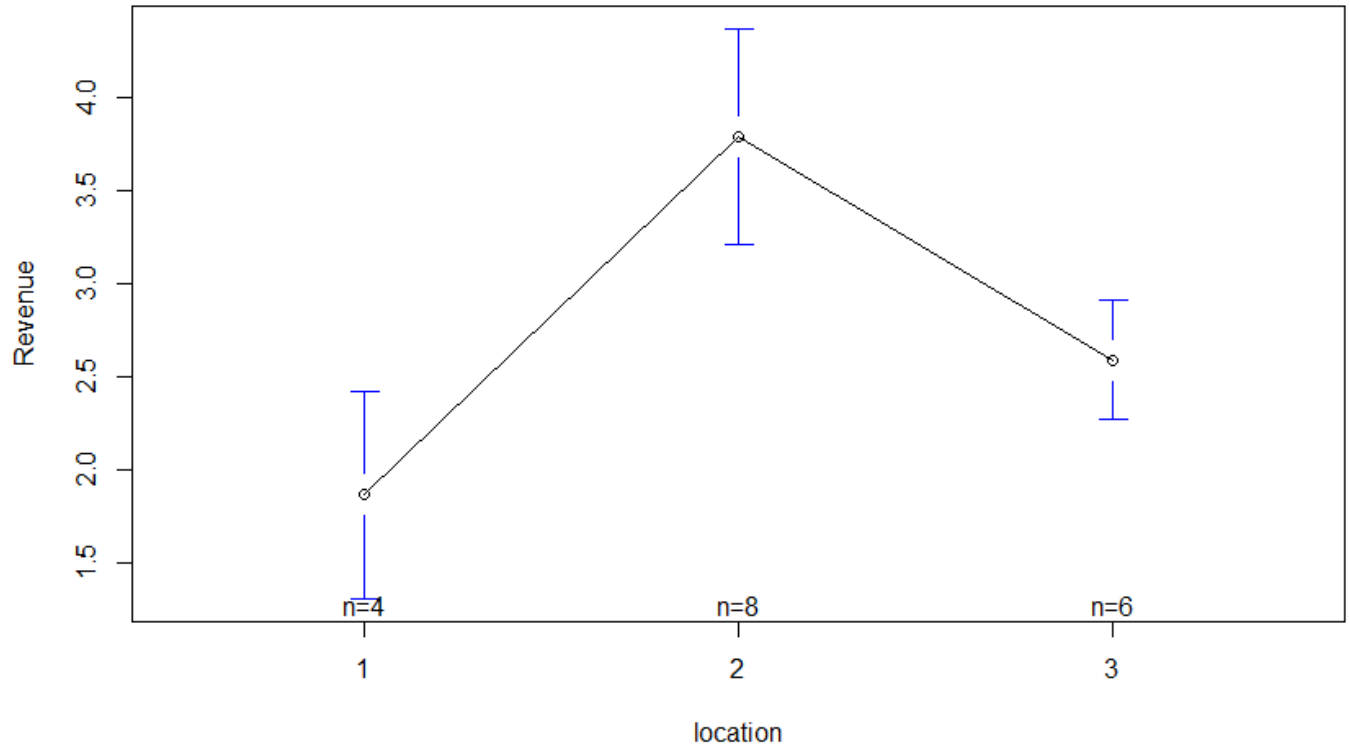
```
> boxplot(revenue ~ location)
```



ANALYSIS OF VARIANCE

One Way Anova : Example Result

```
> library(gplots)
> plotmeans(revenue ~ location)
```



ANALYSIS OF VARIANCE

One Way Anova : Tukey's Honestly Significant Difference (HSD) Test

Used to do pair wise comparison between the levels of factors

R code

>TukeyHSD(fit)

Comparison	Mean difference	Lower	Upper	p value
2 - 1	1.92125	1.086067	2.756433	0.0000
3 - 1	0.724167	-0.15619	1.604527	0.1158
3 - 2	-1.19708	-1.93365	-0.46052	0.0020

ANALYSIS OF VARIANCE

Anova logic:

Two Types of Variations:

1. Variation within the level of a factor
2. Variation between the levels of factor

ANALYSIS OF VARIANCE

Anova logic :

Variation between the level of a factor:

The effect of Factor.

Variation within the levels of a factor:

The inherent variation in the process or Process Error.

	Location		
	Front	Middle	rear
Sales Revenue	1.34 1.89 1.35 2.07 2.41 3.06	3.20 2.81 4.52 4.40 4.75 5.19 3.42 9.80	2.30 1.91 1.40 1.48

ANALYSIS OF VARIANCE

Anova logic :

If the variation between the levels of a factor is significantly higher than the inherent variation

then the factor has significant effect on response

To check whether a factor is significant:

Compare variation between levels with variation within levels

ANALYSIS OF VARIANCE

Anova logic :

Measure of variation between levels: MS of the factor ($MS_{between}$)

Measure of variation within levels: MS Error (MS_{within})

To check whether a factor is significant:

Compare MS of between with MS within

i.e. Calculate F = $MS_{between}$ / MS_{within}

If F is very high, then the factor is significant.

ANALYSIS OF VARIANCE

Variation Within levels:

Ideally variation within all the levels should be same

To check whether variation within the levels are same or not

Do Bartlett's test

If p value ≥ 0.05, then variation within the levels are equal, otherwise not

R Code for Bartlett's test

```
> bartlett.test(revenue, location, data = mydata)
```

Bartlett's Test result for sales revenue (location of TV sets) example

Bartlett's K^2 Statistic	df	p value
3.8325	2	0.1472

Since p value = 0.1472 > 0.05, the variance within the levels are equal

Appendix

Proof of SST = SSTr + SSE

We have,

$$SST = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(X_{ij} - \bar{X}_{00})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(X_{ij} - \bar{X}_{i0} + \bar{X}_{i0} - \bar{X}_{00})^2$$

$$= \sum_{i=1}^{I}\sum_{j=1}^{J_i}(\bar{X}_{i0} - \bar{X}_{00})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J_i}(X_{ij} - \bar{X}_{i0})^2 = SSTr + SSE.$$

Expression for E(SSTr)

Note that

$$\bar{X}_{io} = \frac{1}{J_i}\sum_{j=1}^{J_i} X_{ij} = \frac{1}{J_i}\sum_{j=1}^{J_i}(\mu + \alpha_i + \epsilon_{ij}) = \mu + \alpha_i + \bar{\epsilon}_{i0},$$

and,

$$\bar{X}_{00} = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J_i} X_{ij} = \frac{1}{J_i}\sum_{j=1}^{J_i}(\mu + \alpha_i + \epsilon_{ij}) = \mu + \bar{\epsilon}_{00}$$

Thus,

$$SSTr = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(\bar{X}_{i0} - \bar{X}_{00})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(\alpha_i + \bar{\epsilon}_{i0} - \bar{\epsilon}_{00})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J_i}\alpha_i^2 + \sum_{i=1}^{I}\sum_{j=1}^{J_i}(\bar{\epsilon}_{i0} - \bar{\epsilon}_{00})^2.$$

$$= \sum_{i=1}^{I} J_i\alpha_i^2 + \sum_{i=1}^{I}\sum_{j=1}^{J_i}(\bar{\epsilon}_{io}^2 + \bar{\epsilon}_{00}^2 - 2\bar{\epsilon}_{i0}\,\bar{\epsilon}_{00}) = \sum_{i=1}^{I} J_i\alpha_i^2 + \sum_{i=1}^{I} J_i\,\bar{\epsilon}_{i0}^2 + N\bar{\epsilon}_{00}^2 - 2\bar{\epsilon}_{00}\sum_{i=1}^{I} J_i\,\bar{\epsilon}_{i0}$$

$$= \sum_{i=1}^{I} J_i\alpha_i^2 + \sum_{i=1}^{I}\bar{\epsilon}_{io}^2 + N\bar{\epsilon}_{00}^2 - 2\bar{\epsilon}_{00}\sum_{i=1}^{I}\sum_{j=1}^{J_i}\epsilon_{ij} = \sum_{i=1}^{I} J_i\alpha_i^2 + \sum_{i=1}^{I} J_i\,\bar{\epsilon}_{i0}^2 + N\bar{\epsilon}_{00}^2 - 2N\bar{\epsilon}_{00}^2 = \sum_{i=1}^{I} J_i\alpha_i^2 + \sum_{i=1}^{I} J_i\,\bar{\epsilon}_{i0}^2 - N\bar{\epsilon}_{00}^2$$

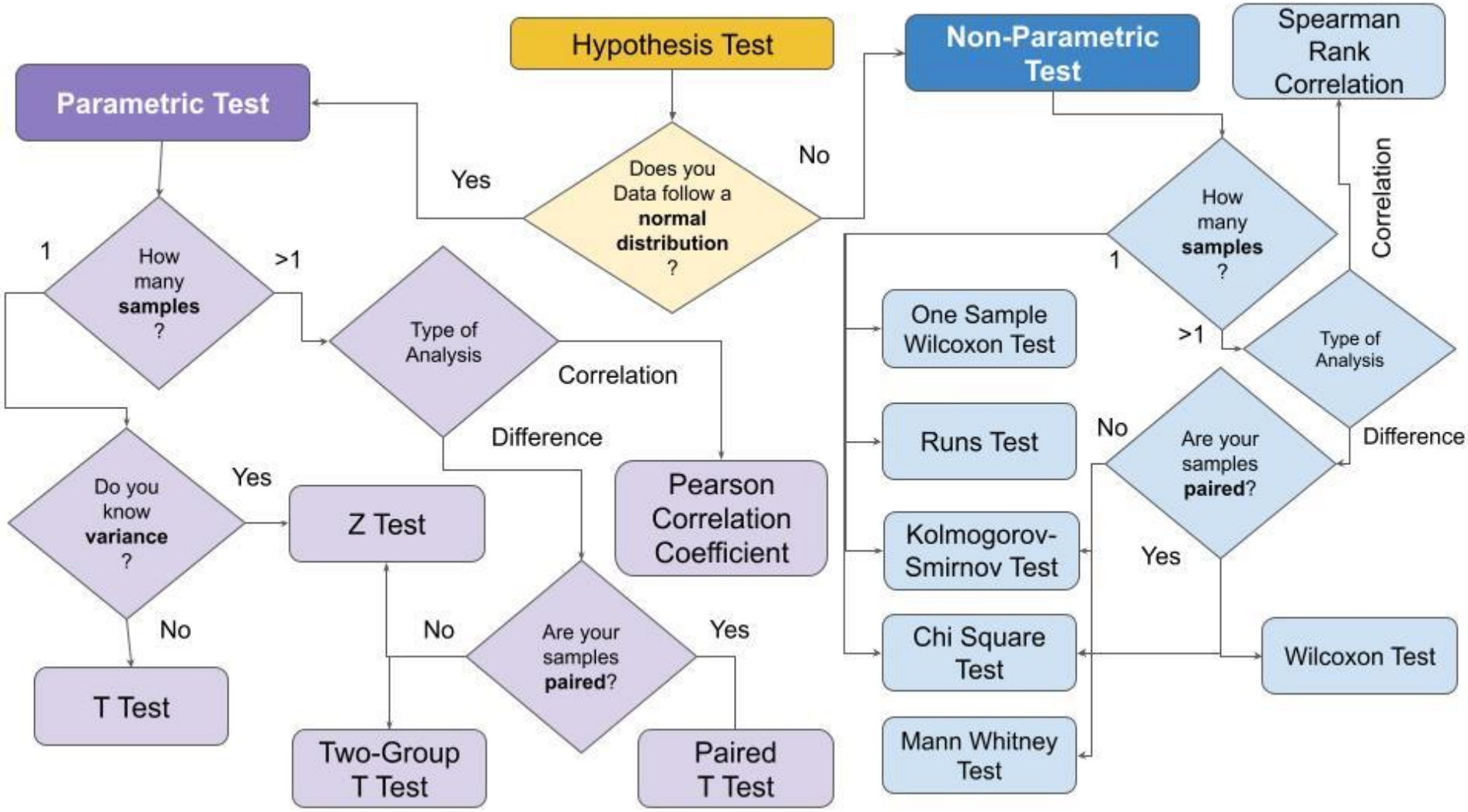
Note that,

$$\bar{\epsilon}_{i0} \overset{indep}{\sim} N(0, \frac{\sigma^2}{J_i}),\ \bar{\epsilon}_{00} \sim N(0, \frac{\sigma^2}{N}).$$

Thus,

$$E(SSTr) = \sum_{i=1}^{I} J_i\alpha_i^2 + \sum_{i=1}^{I} J_i\frac{\sigma^2}{J_i} - N\frac{\sigma^2}{N} = \sum_{i=1}^{I} J_i\alpha_i^2 + (I-1)\sigma^2.$$

CHEATSHEET


Hypothesis Test
Parametric Test
Yes
Does you Data follow a normal distribution ?
No
Non-Parametric Test
Spearman Rank Correlation
1
How many samples ?
>1
Type of Analysis
Correlation
Difference
Do you know variance ?
Yes
Z Test
No
T Test
No
Are your samples paired?
Yes
Two-Group T Test
Paired T Test
Pearson Correlation Coefficient
1
How many samples ?
>1
Type of Analysis
Correlation
Difference
No
Are your samples paired?
Yes
One Sample Wilcoxon Test
Runs Test
Kolmogorov-Smirnov Test
Chi Square Test
Mann Whitney Test
Wilcoxon Test