

ANalysis Of VAriance (ANOVA)

Analysis of variance (frequently abbreviated to ANOVA) is an extremely powerful statistical technique, that can be used to separate and estimate the different causes of variation when more than two means are to be compared.

More specifically, it can be used to separate any variation which is caused by changing a controlled factor from the variation due to random error. It can thus test whether changing the controlled factor leads to a significant difference between mean values obtained.

Actually, ANOVA can also be used when there is more than one source of random variation.

Consider, for example, the purity testing of a barrelful of sodium chloride. Samples are taken from different parts of the barrel chosen at random and replicated analyses are performed on these samples. In addition to the random error in the measurement of the purity, there may also be variation in the purity of the samples taken from different parts of the barrel. Since the samples are chosen at random, this variation is sometimes known as a random-effect factor. Again, ANOVA can be used to separate and estimate the sources of variation.

Both types of statistical analysis now described, i.e., where there is one factor, either controlled or random, in addition to the random error, are known as one-way ANOVA.

Let us suppose that 13 different groups of students measured the enthalpy variation related to HCl neutralization with NaOH, each of them making 5 replicated measurements. Mean values and variances for each group are easily calculated:

| Group | x_1 | x_2 | x_3 | x_4 | x_5 | \bar{x} | s^2 |
|-------|-------|-------|-------|-------|-------|-----------|-------|
| 1 | 56.9 | 59.2 | 56.3 | 58.0 | 56.9 | 57.46 | 1.32 |
| 2 | 53.8 | 55.4 | 58.0 | 59.6 | 55.5 | 56.46 | 5.34 |
| 3 | 58.4 | 55.0 | 55.7 | 56.6 | 57.2 | 56.58 | 1.74 |
| 4 | 58.0 | 56.4 | 57.6 | 57.5 | 55.0 | 56.90 | 1.48 |
| 5 | 57.7 | 58.5 | 58.9 | 57.8 | 57.4 | 58.06 | 0.38 |
| 6 | 54.8 | 56.4 | 55.2 | 60.3 | 57.1 | 56.76 | 4.76 |
| 7 | 57.1 | 60.4 | 58.9 | 55.5 | 54.7 | 57.32 | 5.55 |
| 8 | 58.6 | 57.8 | 58.0 | 55.5 | 55.6 | 57.10 | 2.09 |
| 9 | 58.9 | 59.8 | 60.0 | 57.1 | 56.4 | 58.44 | 2.61 |
| 10 | 59.5 | 57.7 | 60.0 | 57.6 | 56.8 | 58.32 | 1.86 |
| 11 | 57.2 | 58.2 | 57.4 | 55.7 | 59.1 | 57.52 | 1.60 |
| 12 | 55.4 | 56.1 | 57.7 | 56.9 | 59.2 | 57.06 | 2.17 |
| 13 | 55.1 | 56.8 | 55.7 | 61.6 | 58.3 | 57.50 | 6.74 |

Here the statistical problem is represented by the comparison of the 13 mean values.

In principle, one could make pairwise comparisons between the 13 mean values using a t-test.

However, if a significance level α (Type I error) is considered for each test, it can be demonstrated that the error rate related to the family of data (the ensemble of groups), called Family Wise (FW) error rate, is given by:

$$\alpha_{FW} = 1 - (1 - \alpha)^c$$

where c is the number of comparisons to be made.

As an example, if two independent hypothesis tests are considered, each at a significance level α , the probability that neither is affected by Type I error is $(1-\alpha)^2$

Consequently, the probability that at least one test is affected by Type I error is:

$$\alpha_2 = 1 - (1 - \alpha)^2$$

If every pair of h means had to be tested, a total of $C = h(h-1)/2$ t-tests, each at a significance level α would be required.

The probability of finding at least one erroneous difference would then be:

$$\alpha_C = 1 - (1 - \alpha)^C$$

As an example, for $\alpha = 0.05$ and $h = 3$, which implies $c = 3$, the probability would be:

$$\alpha_{FW} = 1 - (1 - 0.05)^3 = 0.143$$

thus a 14% Type I family wise error would be obtained by making pair-wise comparisons between 3 mean values.

ANOVA is an attempt to keep the family wise error at an acceptable level.

The data table shown before can be generalized, indicating with i the different groups and with j the different replicates in each group:

| | | | | | | | | |
|-----------|-----|-------------------------------|-----------|-------|----------|-------|----------|---------------------------|
| | | $\xrightarrow{\quad j \quad}$ | | | | | | |
| Group 1 | | X_{11} | X_{12} | | X_{1j} | | X_{1n} | $\sim N(\mu_1, \sigma^2)$ |
| Group 2 | | X_{21} | $X_{2,2}$ | | X_{2j} | | X_{2n} | $\sim N(\mu_2, \sigma^2)$ |
| | | | | | | | | |
| Group i | i | X_{i1} | X_{i2} | | X_{ij} | | X_{in} | $\sim N(\mu_i, \sigma^2)$ |
| | | | | | | | | |
| Group h | | X_{h1} | X_{h2} | | X_{hj} | | X_{hn} | $\sim N(\mu_h, \sigma^2)$ |

X_{ij} indicates the j -th replicate of the i -th group

n is the number of replicates for each group (in the specific case n is the same for all groups)

h is the number of groups

$N = n \times h$ is the total number of data

The basic assumption is that data in each group are extracted from a normal population with a specific mean but with the same variance as that of other groups. The latter assumption can be checked preliminarily using one of the tests on multiple variances for normal variables discussed before (specifically, the Hartley's or the Bartlett's tests).

We may describe the observations reported in the table by the **linear statistical model** (known as **effects model**):

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \left\{ \begin{array}{l} i = 1, 2, \dots, h \\ j = 1, 2, \dots, n \end{array} \right. \quad \text{Equation 1}$$

where:

- X_{ij} is a random variable denoting the $(ij)^{th}$ observation
- μ is a parameter common to all groups (treatments), called the overall mean
- α_i is a parameter associated with the i -th group, called the i -th group effect
- ε_{ij} is a random error component.

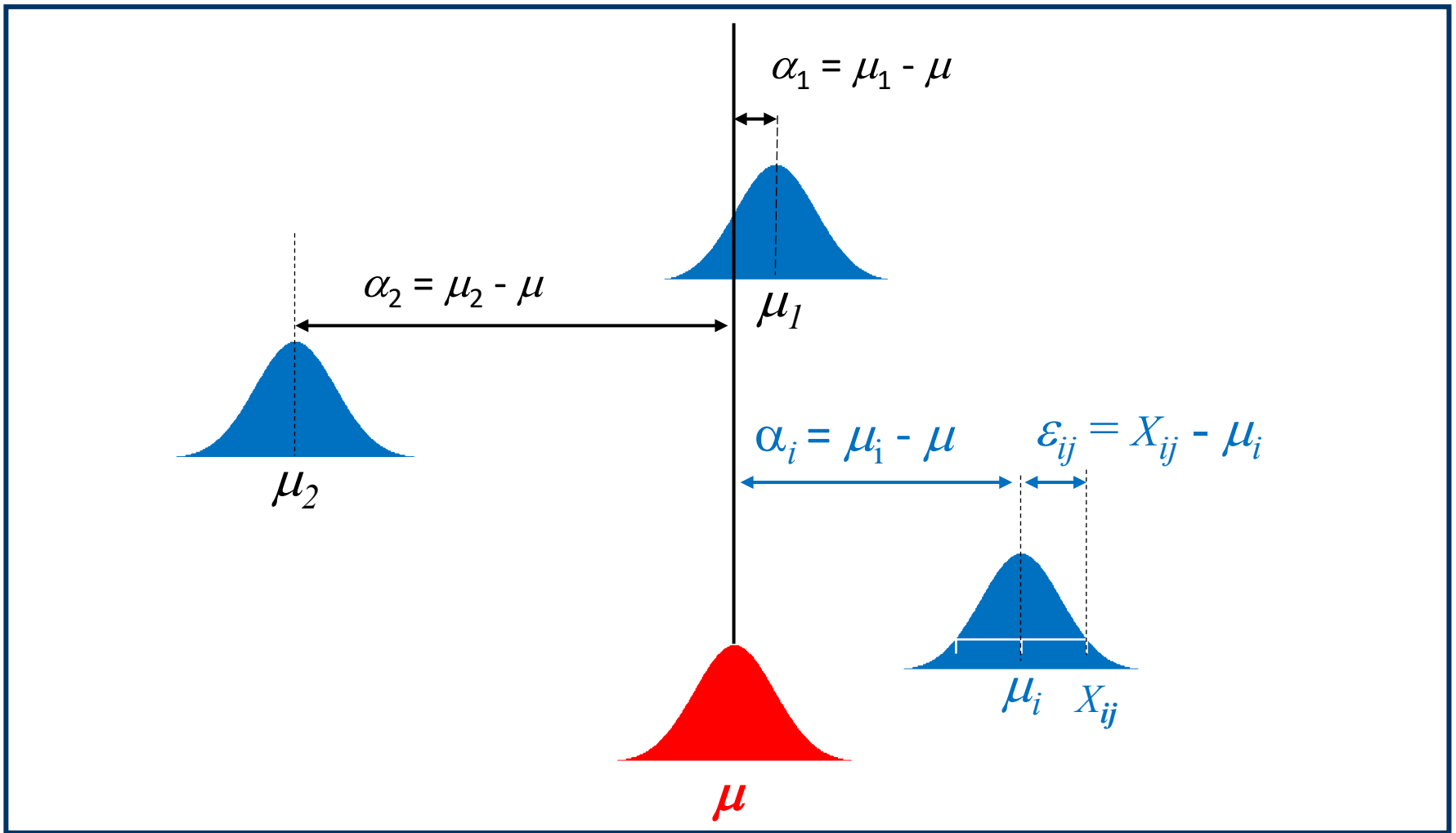
Note that the model can be expressed also as the so-called **means model**:

$$X_{ij} = \mu_i + \varepsilon_{ij} \quad \left\{ \begin{array}{l} i = 1, 2, \dots, h \\ j = 1, 2, \dots, n \end{array} \right. \quad \text{where:}$$

$\mu_i = \mu + \alpha_i$ is the mean of the i -th treatment.

In this form of the model each group defines a population that has mean μ_i , consisting of the overall mean μ plus an effect α_i , that is specific for that particular group.

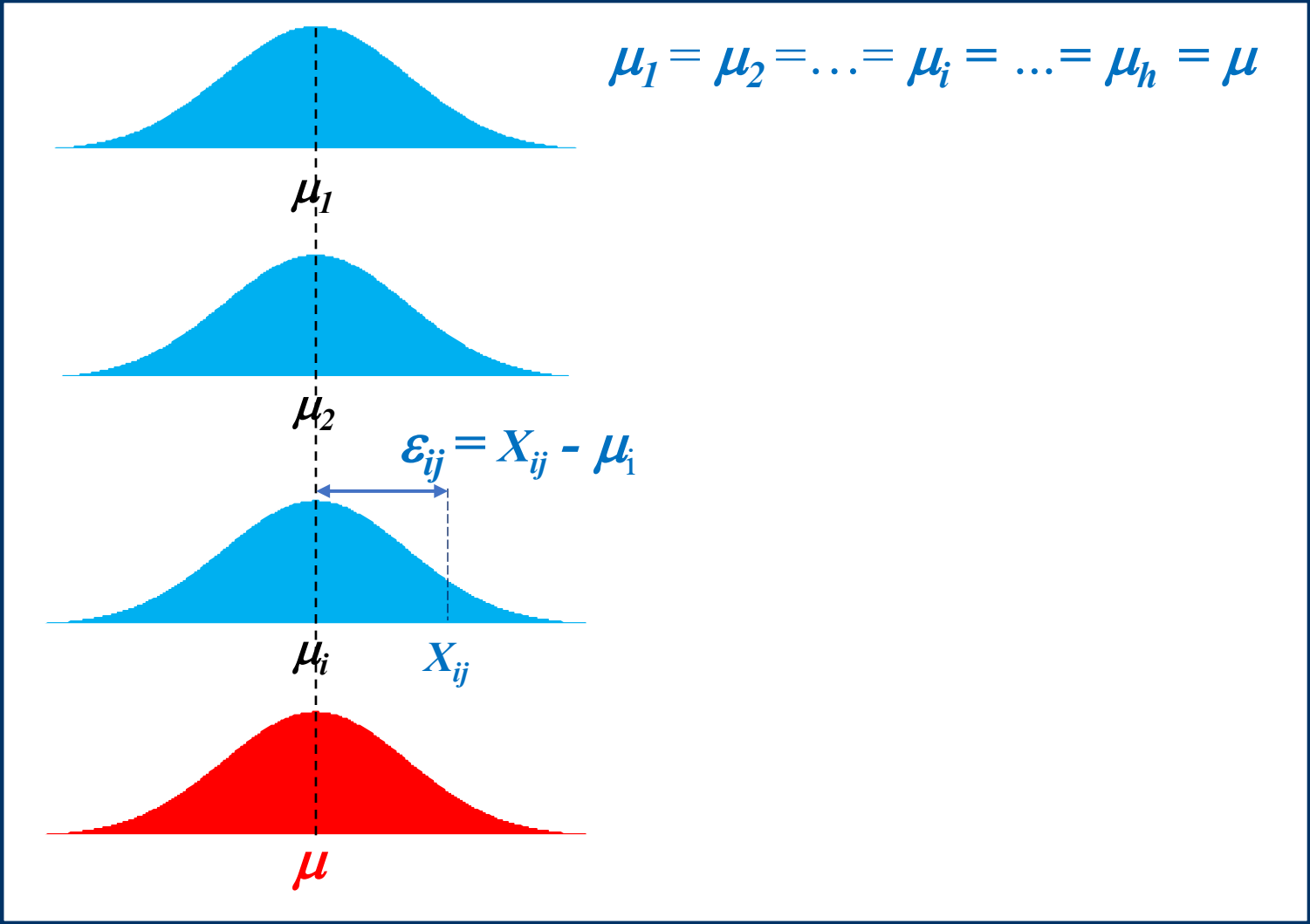
As shown in the following figure, referred to the case in which all α_i are different from 0, the basic assumption is that errors ε_{ij} are normally and independently distributed with mean zero and variance σ^2 .



The model for each observed response is:

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \text{with} \quad \sum_{i=1}^h \alpha_i = 0$$

In the case represented by the following figure, all α_i are equal to 0, which means that all μ_i are statistically equal.



Equation 1:

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \left\{ \begin{array}{l} i = 1, 2, \dots, h \\ j = 1, 2, \dots, n \end{array} \right.$$

is the underlying model for a **single-factor experiment**. Furthermore, since we require that the observations are taken in random order and that the environment in which the treatments are used is as uniform as possible, this design is called a **completely randomized experimental design**.

Since we are interested in testing the equality of the h group/treatment means:

$$\mu_1 = \mu_2 = \dots = \mu_i = \dots = \mu_h$$

the null (H_0) and the alternative (H_1) hypotheses can be formulated as follows:

$$\left\{ \begin{array}{l} H_0 : \mu_1 = \mu_2 = \dots = \mu_i = \dots = \mu_h = \mu \\ H_1 : H_0 \text{ is not true for at least one couple of values} \end{array} \right.$$

Equation 2

Considering the model given by Equation 1 and that equivalently, formulated as follows:

$$\sum_{i=1}^h \alpha_i = 0 \quad H_0 \text{ and } H_1 \text{ can be,}$$

$$\left\{ \begin{array}{l} H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_i = \dots = \alpha_h = 0 \\ H_1 : \alpha_i \neq 0 \quad \text{for at least one } i \end{array} \right.$$

or as:

$$\left\{ \begin{array}{l} H_0 : X_{ij} = \mu + \varepsilon_{ij} \\ H_1 : X_{ij} = \underbrace{\mu + \alpha_i}_{\mu + \alpha_i} + \varepsilon_{ij} \quad \text{for at least one } i \end{array} \right.$$

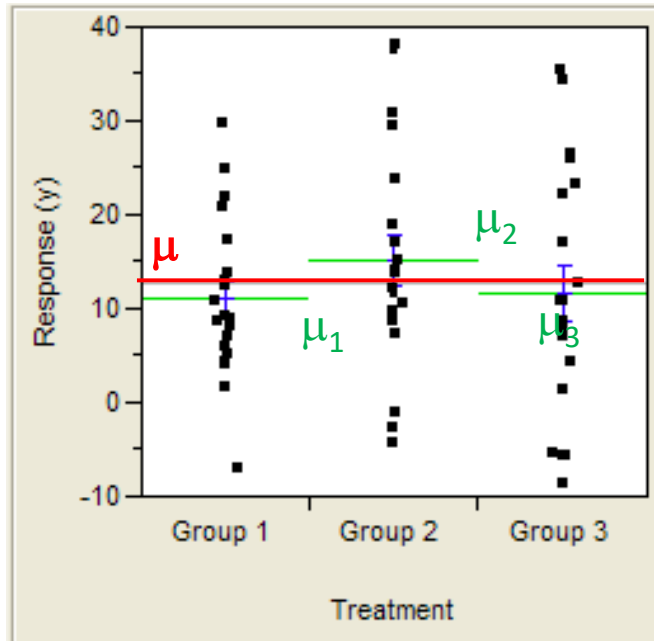
Thus, if the null hypothesis H_0 is true, each observation consists of the overall mean μ plus a realization of the random error component ε_{ij} .

This is equivalent to say that all N observations (or all the h groups) are taken from a normal distribution with mean μ and variance σ^2 .

Therefore, if the null hypothesis is true, changing the levels of the factor (i.e., changing the group) has no effect on the mean response. Note also that:

$$\varepsilon_{ij} \sim N(0, \sigma^2) \quad E(\varepsilon_{ij}) = 0 \quad E(\varepsilon_{ij}^2) = \sigma^2$$

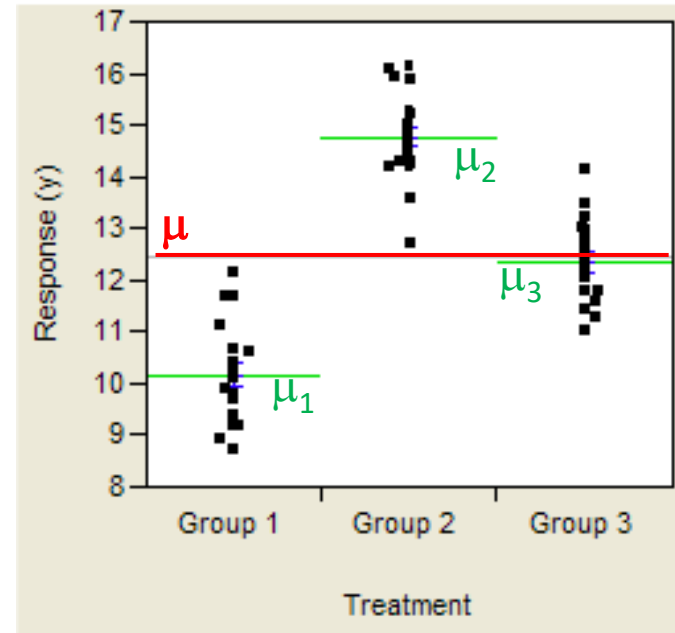
ANOVA partitions the total variability related to sample data into two component parts (*between group and within group variability*).



Between-group variation is small compared to within-group variation



Here we would likely accept the null hypothesis.



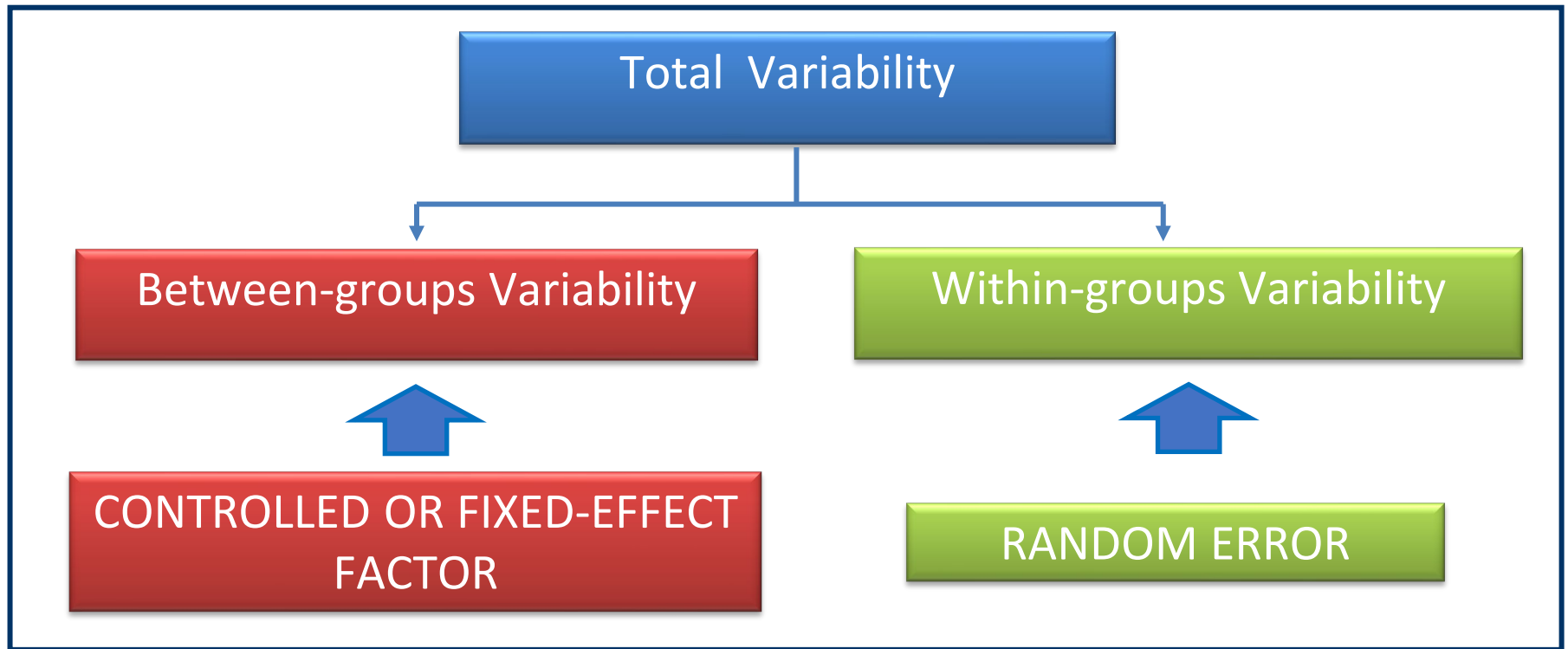
Between-group variation is large compared to within-group variation



Here we would almost certainly reject the null hypothesis.

Then, the test of the hypothesis expressed by *Equation 2* is based on **a comparison of two independent estimates of the population variance**.

The distinction between Total Variability components is expressed graphically in the following figure:



If the null hypothesis is true, a not significant contribution of between-groups variability should be expected, since the observed variability would be due only to random error.

If the null hypothesis is false, both variabilities would be expected to contribute to the total variability, being the between-groups variability too high to be explained only by random error.

Partitioning of the total variability

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \text{Equation 1 - model for the observed response}$$



$$X_{ij} - \mu = \alpha_i + \varepsilon_{ij} = (\mu_i - \mu) + (X_{ij} - \mu_i)$$

Considering the estimators of μ and μ_i , i.e., $\bar{\bar{X}}$ and \bar{X}_i , the following equation is obtained:

$$(X_{i,j} - \bar{\bar{X}}) = (\bar{X}_i - \bar{\bar{X}}) + (X_{i,j} - \bar{X}_i)$$

The deviation of each observation from the **grand mean**, i.e., the mean of all the values grouped together, can thus be partitioned into the deviation of the corresponding group's mean from the grand mean and the deviation of that observation from its group's mean.

If both members of the equation are squared and the sums over indexes i and j are calculated, the following equation can be obtained:

$$\sum_{i=1}^h \sum_{j=1}^n (X_{ij} - \bar{\bar{X}})^2 = n \sum_{i=1}^h (\bar{X}_i - \bar{\bar{X}})^2 + \sum_{i=1}^h \sum_{j=1}^n (X_{ij} - \bar{X}_i)^2 \quad \text{Equation 3}$$

as demonstrated in the following slide.

$$(X_{i,j} - \bar{\bar{X}}) = (\bar{X}_i - \bar{\bar{X}}) + (X_{i,j} - \bar{X}_i) \quad \rightarrow \quad (X_{i,j} - \bar{\bar{X}})^2 = [(\bar{X}_i - \bar{\bar{X}}) + (X_{i,j} - \bar{X}_i)]^2$$

$$\sum_{i=1}^h \sum_{j=1}^n (X_{i,j} - \bar{\bar{X}})^2 = \sum_{i=1}^h \sum_{j=1}^n [(\bar{X}_i - \bar{\bar{X}}) + (X_{i,j} - \bar{X}_i)]^2$$

$$\sum_{i=1}^h \sum_{j=1}^n (X_{i,j} - \bar{\bar{X}})^2 = n \sum_{i=1}^h (\bar{X}_i - \bar{\bar{X}})^2 + \sum_{i=1}^h \sum_{j=1}^n (X_{i,j} - \bar{X}_i)^2 + 2 \sum_{i=1}^h \sum_{j=1}^n (\bar{X}_i - \bar{\bar{X}})(X_{i,j} - \bar{X}_i)$$

Since:

$$\begin{aligned} 2 \sum_{i=1}^h \sum_{j=1}^n (\bar{X}_i - \bar{\bar{X}})(X_{i,j} - \bar{X}_i) &= 2 \sum_{i=1}^h (\bar{X}_i - \bar{\bar{X}}) \sum_{j=1}^n (X_{i,j} - \bar{X}_i) = \\ &= 2 \sum_{i=1}^h (\bar{X}_i - \bar{\bar{X}}) \left(\sum_{j=1}^n X_{i,j} - n\bar{X}_i \right) = 2 \sum_{i=1}^h (\bar{X}_i - \bar{\bar{X}}) \left(\sum_{j=1}^n X_{i,j} - n \frac{\sum_{j=1}^n X_{i,j}}{n} \right) = 0 \end{aligned}$$

Equation 3 is obtained.

$$\sum_{i=1}^h \sum_{j=1}^n (X_{ij} - \bar{X})^2 = n \sum_{i=1}^h (\bar{X}_i - \bar{X})^2 + \sum_{i=1}^h \sum_{j=1}^n (X_{ij} - \bar{X}_i)^2$$



**Total sum of
Squares, SS_{tot}**



**Between-group sum
of squares, $SS_{between}$**



**Within-group sum
of squares, SS_{within}**

If the **abbreviated notation for sum of squares** is used, **Equation 3** can be written as:


$$SS_{tot} = SS_{between} + SS_{within}$$


$$SS_{tot} = SS_{treatment} + SS_{error}$$

To convert *Sums of Squares (SS)* into comparable measures of variance we need to divide each of them by the respective *degrees of freedom (df)*.

| <i>Sum of Squares (SS)</i> | <i>Degrees of freedom (df)</i> |
|----------------------------|--------------------------------|
| <i>Between group</i> | $h-1$ |
| <i>Within group</i> | $h(n-1) = N-h$ |
| Total | $N-1$ |

Note that the degrees of freedom for the within-group sum of squares can be calculated by considering that the degrees of freedom are additive, thus it results:

$$N-1 = (h-1) + (N-h)$$

A **Sum of Squares** divided by **df** provides the respective **Mean Square (MS)**:

| <i>Source of variation</i> | <i>Sums of Squares (SS)</i> | <i>Degrees of freedom (df)</i> | <i>Mean Squares (MS)</i> |
|----------------------------|--|--------------------------------|---------------------------|
| <i>Between-group</i> | $n \sum_{i=1}^h (\bar{X}_i - \bar{\bar{X}})^2$ | <i>h-1</i> | $MS_B = \frac{SS_B}{h-1}$ |
| <i>Within-group</i> | $\sum_{i=1}^h \sum_{j=1}^n (X_{ij} - \bar{X}_i)^2$ | <i>N-h</i> | $MS_W = \frac{SS_W}{N-h}$ |
| Total | $\sum_{i=1}^h \sum_{j=1}^n (X_{ij} - \bar{\bar{X}})^2$ | N-1 | |

Interestingly, **Sum of Squares** can be calculated also using the following values:

1) Row (group) totals $T_i = \sum_{j=1}^n X_{ij}$

2) Grand total $T = \sum_{i=1}^h T_i$

| | | | | | | | |
|----------|----------|-----|----------|-----|----------|-------|---------|
| X_{11} | X_{12} | ... | X_{1j} | ... | X_{1n} | T_1 | T_1^2 |
| X_{21} | X_{22} | ... | X_{2j} | ... | X_{2n} | T_2 | T_2^2 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| X_{i1} | X_{i2} | ... | X_{ij} | ... | X_{in} | T_i | T_i^2 |
| ... | ... | ... | | ... | ... | ... | ... |
| X_{h1} | X_{h2} | ... | X_{hj} | ... | X_{hn} | T_h | T_h^2 |
| | | | | | | T | T^2 |

This can be demonstrated as follows:

Between-group *Sum of Squares*

$$\begin{aligned} n \sum_{i=1}^h (\bar{X}_i - \bar{\bar{X}})^2 &= n \sum_{i=1}^h (\bar{X}_i^2 + \bar{\bar{X}}^2 - 2\bar{X}_i\bar{\bar{X}}) = \\ &= n \left(\sum_{i=1}^h \bar{X}_i^2 + h\bar{\bar{X}}^2 - 2\bar{\bar{X}} \sum_{i=1}^h \bar{X}_i \right) = \\ &= n \sum_{i=1}^h \frac{\left(\sum_{j=1}^n X_{ij} \right)^2}{n^2} + nh\bar{\bar{X}}^2 - 2n\bar{\bar{X}} \sum_{i=1}^h \frac{\sum_{j=1}^n X_{ij}}{n} = \\ &= \sum_{i=1}^h \frac{\left(\sum_{j=1}^n X_{ij} \right)^2}{n} + N \frac{\left(\sum_{i=1}^h \sum_{j=1}^n X_{ij} \right)^2}{N^2} - 2n \frac{\sum_{i=1}^h \sum_{j=1}^n X_{ij}}{N} \frac{\sum_{i=1}^h \sum_{j=1}^n X_{ij}}{n} = \\ &= \sum_{i=1}^h \frac{\left(\sum_{j=1}^n X_{ij} \right)^2}{n} + \frac{\left(\sum_{i=1}^h \sum_{j=1}^n X_{ij} \right)^2}{N} - 2 \frac{\left(\sum_{i=1}^h \sum_{j=1}^n X_{ij} \right)^2}{N} = \\ &= \sum_{i=1}^h \frac{\left(\sum_{j=1}^n X_{ij} \right)^2}{n} - \frac{\left(\sum_{i=1}^h \sum_{j=1}^n X_{ij} \right)^2}{N} = \\ &= \sum_{i=1}^h \frac{T_i^2}{n} - \frac{T^2}{N} \end{aligned}$$

Total Sum of Squares

$$\begin{aligned} \sum_{i=1}^h \sum_{j=1}^n (X_{ij} - \bar{\bar{X}})^2 &= \sum_{i=1}^h \sum_{j=1}^n (X_{ij}^2 + \bar{\bar{X}}^2 - 2X_{ij}\bar{\bar{X}}) = \\ &= \sum_{i=1}^h \sum_{j=1}^n X_{ij}^2 + \sum_{i=1}^h \sum_{j=1}^n \bar{\bar{X}}^2 - 2\bar{\bar{X}} \sum_{i=1}^h \sum_{j=1}^n X_{ij} = \\ &= \sum_{i=1}^h \sum_{j=1}^n X_{ij}^2 + N \frac{\left(\sum_{i=1}^h \sum_{j=1}^n X_{ij}\right)^2}{N^2} - 2 \frac{\sum_{i=1}^h \sum_{j=1}^n X_{ij}}{N} \sum_{i=1}^h \sum_{j=1}^n X_{ij} = \\ &= \sum_{i=1}^h \sum_{j=1}^n X_{ij}^2 - \frac{\left(\sum_{i=1}^h \sum_{j=1}^n X_{ij}\right)^2}{N} = \\ &= \sum_{i=1}^h \sum_{j=1}^n X_{ij}^2 - \frac{T^2}{N} \end{aligned}$$

The **within-group *Sum of Squares*** is obtained by subtraction:

$$\underbrace{\sum_{i=1}^h \sum_{j=1}^n X_{ij}^2 - \frac{T^2}{N}}_{\text{Total}} - \underbrace{\left(-\frac{\sum_{i=1}^h T_i^2}{n} + \frac{T^2}{N} \right)}_{\text{Between-group}} = \underbrace{\sum_{i=1}^h \sum_{j=1}^n X_{ij}^2 - \frac{\sum_{i=1}^h T_i^2}{n}}_{\text{Within-group}}$$

Expectation of Mean Squares in one-way (single factor) ANOVA

Expectation of Mean Squares can be calculated using the properties of expectation. Starting from the **Between-group mean square, MS_B** , the following equations can be written:

$$E(MS_B) = E\left(\frac{SS_B}{h-1}\right) \quad \text{where:} \quad SS_B = \frac{\sum_{i=1}^n T_i^2}{n} - \frac{T^2}{N}$$

Thus:

$$E(SS_B) = E\left(\frac{\sum_{i=1}^h T_i^2}{n}\right) - E\left(\frac{T^2}{N}\right)$$

According to the one-way ANOVA model the following equations can be written:

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \left\{ \begin{array}{l} T_i = \sum_{j=1}^n X_{ij} = \sum_{j=1}^n (\mu + \alpha_i + \varepsilon_{ij}) \\ T_i^2 = \left(\sum_{j=1}^n (\mu + \alpha_i + \varepsilon_{ij}) \right)^2 \end{array} \right.$$

Thus:

$$E\left(\frac{\sum_{i=1}^h T_i^2}{n}\right) = \frac{1}{n} \sum_{i=1}^h E(T_i^2) = \frac{1}{n} \sum_{i=1}^h E\left(\sum_{j=1}^n (\mu + \alpha_i + \varepsilon_{ij})\right)^2 = \frac{1}{n} \sum_{i=1}^h E\left(n\mu + n\alpha_i + \sum_{j=1}^n \varepsilon_{ij}\right)^2$$

Note that: $E(\varepsilon_{ij}) = 0$

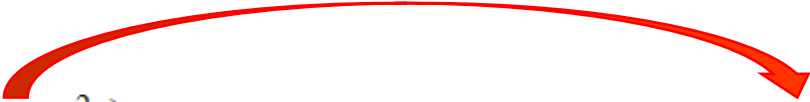
Consequently, when the square of the trinomial reported in the last member is calculated, **all cross-product terms involving ε_{ij} can be canceled.**

As for **the further cross-product term resulting from the square of trinomial, i.e., $2 n^2 \mu \alpha_i$,** the following equation can be easily obtained:

$$\sum_{i=1}^h E(2 n^2 \mu \alpha_i) = 2 n^2 \mu E\left(\sum_{i=1}^h \alpha_i\right) = 0$$

Thus:

Expectations of cross-terms products resulting from the square of ε_{ij} sum are equal to 0

$$E\left(\frac{\sum_{i=1}^h T_i^2}{n}\right) = \frac{1}{n} \sum_{i=1}^h \left(n^2 \mu^2 + n^2 \alpha_i^2 + E\left(\sum_{j=1}^n \varepsilon_{ij}\right)^2 \right) = \frac{1}{n} \sum_{i=1}^h \left(n^2 \mu^2 + n^2 \alpha_i^2 + \sum_{j=1}^n E(\varepsilon_{ij}^2) \right)$$


$$\text{Since: } \sum_{j=1}^n E(\varepsilon_{ij}^2) = n\sigma^2$$

the previous equation can be written as follows:

$$E\left(\frac{\sum_{i=1}^h T_i^2}{n}\right) = \frac{1}{n} \sum_{i=1}^h (n^2 \mu^2 + n^2 \alpha_i^2 + n\sigma^2) = \frac{1}{n} \left(hn^2 \mu^2 + n^2 \sum_{i=1}^h \alpha_i^2 + hn\sigma^2 \right)$$

Finally:

$$E\left(\frac{\sum_{i=1}^h T_i^2}{n}\right) = hn\mu^2 + n \sum_{i=1}^h \alpha_i^2 + h\sigma^2$$

Let us now consider the **second term in the expression of $E(SS_B)$** :

$$\begin{aligned} E\left(\frac{T^2}{N}\right) &= \frac{1}{nh} E(T^2) = \frac{1}{nh} E\left(\sum_{i=1}^h \left(\sum_{j=1}^n (\mu + \alpha_i + \varepsilon_{ij})\right)\right)^2 \\ &= \frac{1}{nh} E\left(hn\mu + n\sum_{i=1}^h \alpha_i + \sum_{i=1}^h \sum_{j=1}^n \varepsilon_{ij}\right)^2 = \frac{1}{nh} E\left(hn\mu + \sum_{i=1}^h \sum_{j=1}^n \varepsilon_{ij}\right)^2 \end{aligned}$$

As before, the expected value of the cross-terms resulting from the square of the binomial in the last expression is equal to zero, therefore:

$$E\left(\frac{T^2}{N}\right) = \frac{1}{nh} \left(h^2 n^2 \mu^2 + E\left(\sum_{i=1}^h \sum_{j=1}^n \varepsilon_{ij}\right)^2 \right) = \frac{1}{nh} (h^2 n^2 \mu^2 + hn\sigma^2) = hn\mu^2 + \sigma^2$$

Combining the expressions evidenced by red rectangles in the last two slides, the following expression for $E(SS_B)$ is obtained:

$$E(SS_B) = E\left(\frac{\sum_{i=1}^h T_i^2}{n}\right) - E\left(\frac{T^2}{N}\right) = hn\mu^2 + n\sum_{i=1}^h \alpha_i^2 + h\sigma^2 - (hn\mu^2 + \sigma^2) = \sigma^2(h-1) + n\sum_{i=1}^h \alpha_i^2$$

Therefore:

$$E(MS_B) = E\left(\frac{SS_B}{h-1}\right) = \frac{\sigma^2(h-1) + n\sum_{i=1}^h \alpha_i^2}{h-1} = \sigma^2 + \frac{n\sum_{i=1}^h \alpha_i^2}{h-1}$$

If σ_F^2 is defined as:
$$\sigma_F^2 = \frac{\sum_{i=1}^h \alpha_i^2}{h-1}$$

The following expression is finally obtained:

$$E(MS_B) = \sigma^2 + n\sigma_F^2$$

If the **Within-group mean square, MS_W** , is considered, the following equations can be written:

$$E(MS_W) = E\left(\frac{SS_w}{N-h}\right) = \frac{1}{N-h} E\left[\sum_{i=1}^h \sum_{j=1}^n X_{ij}^2 - \frac{1}{n} \sum_{i=1}^h T_i^2\right]$$

Thus:

$$E(MS_W) = \frac{1}{N-h} E\left[\sum_{i=1}^h \sum_{j=1}^n (\mu + \alpha_i + \varepsilon_{ij})^2 - \frac{1}{n} \sum_{i=1}^h \left(\sum_{j=1}^n (\mu + \alpha_i + \varepsilon_{ij})\right)^2\right]$$

The following equations can be written:

$$E\left[\frac{1}{n} \sum_{i=1}^h \left(\sum_{j=1}^n (\mu + \alpha_i + \varepsilon_{ij})\right)^2\right] = E\left[\frac{1}{n} \sum_{i=1}^h \left(n\mu + n\alpha_i + \sum_{j=1}^n \varepsilon_{ij}\right)^2\right] = \frac{1}{n} \left(hn^2\mu^2 + n^2 \sum_{i=1}^h \alpha_i^2 + hn\sigma^2\right)^2$$

Hence:

$$E(MS_W) = \frac{1}{N-h} \left(N\mu^2 + n \sum_{i=1}^h \alpha_i^2 + N\sigma^2 - N\mu^2 - n \sum_{i=1}^h \alpha_i^2 - h\sigma^2\right) = \sigma^2$$

The results obtained so far can be summarized using the following table:

| <i>Source of variation</i> | <i>SS</i> | <i>df</i> | <i>MS</i> | <i>E(MS)</i> |
|----------------------------|---|------------|---------------------------|--------------------------|
| <i>Between group</i> | $\frac{\sum_{i=1}^h T_i^2}{n} - \frac{T^2}{N}$ | <i>h-1</i> | $MS_B = \frac{SS_B}{h-1}$ | $\sigma^2 + n\sigma_F^2$ |
| <i>Within group</i> | $\sum_{i=1}^h \sum_{j=1}^n X_{ij}^2 - \frac{\sum_{i=1}^h T_i^2}{n}$ | <i>N-h</i> | $MS_W = \frac{SS_W}{N-h}$ | σ^2 |
| Total | $\sum_{i=1}^h \sum_{j=1}^n X_{ij}^2 - \frac{T^2}{N}$ | N-1 | | |

If H_0 is true (i.e., all $\alpha_i = 0$) $\sigma_F^2 = 0$ ($\sigma_F^2 = \frac{\sum_{i=1}^h \alpha_i^2}{h-1}$) thus the between-group mean square MS_B is an unbiased estimator of σ^2 , i.e., of the random error.

If H_1 is true (i.e., some, or all, $\alpha_i \neq 0$) $\sigma_F^2 > 0$ and MS_B estimates σ^2 plus a positive term that incorporates the variation due to the systematic difference in group means.

Interestingly, the within-group mean square MS_W is an unbiased estimator of σ^2 regardless of whether or not H_0 is true, which is reasonable.

Under the assumption that each of the h groups can be modeled as a normal distribution, it can be shown that:

$$F_0 = \frac{MS_B}{MS_W} \text{ is distributed as } F_{h-1, N-h}$$

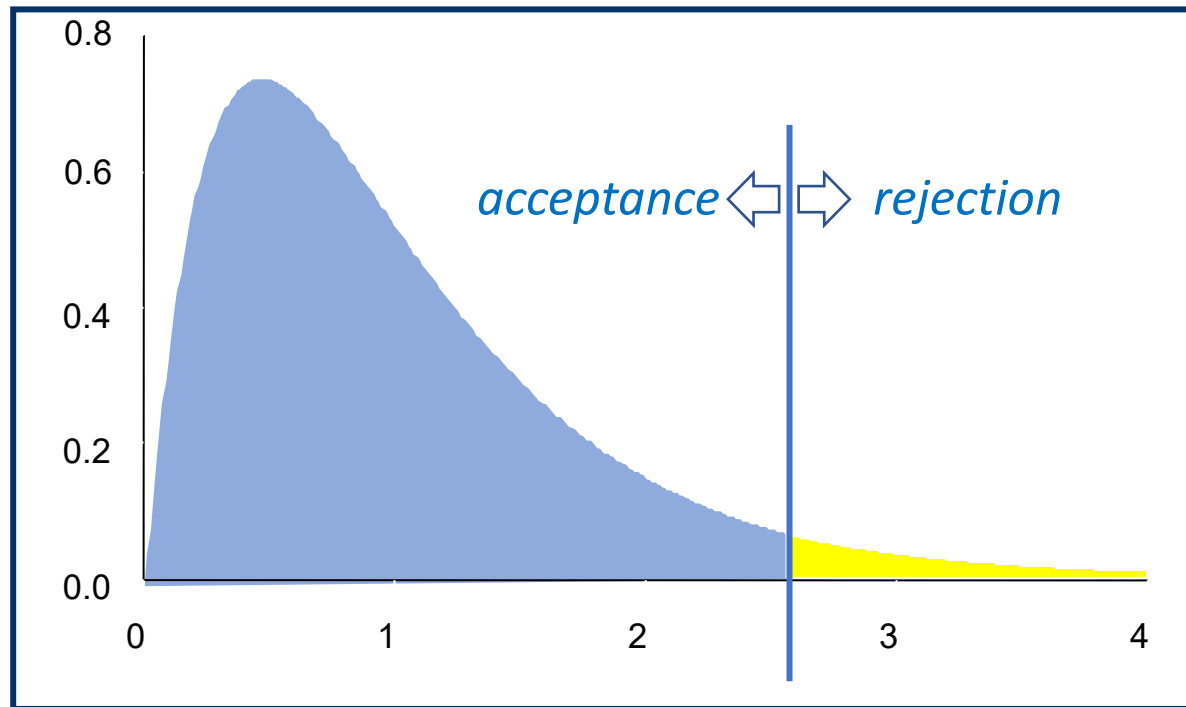
If the alternative hypothesis, H_1 , is true, the expected value of F_0 numerator is higher than that of the denominator.

Consequently, H_0 should be rejected if the realization of F_0 is larger than the critical value. This implies the consideration of an upper-tail (one tail) critical region, i.e., the consideration of a $F_{h-1, N-h, (1-\alpha)}$ as critical value.

As an example, let us consider the $F_{4,30}$ distribution:

If $\alpha = 0.05$, the critical value corresponds to 2.69.

Consequently, if $F_0 < 2.69$, H_0 is accepted at a 5% significance level.



Note that the same test can be made also using the **P-Value**, which corresponds to the **area** underlying the F curve, as calculated from the value assumed by F_0 to infinity.

If $P\text{-value} > \alpha$ H_0 is accepted, if $P\text{-value} < \alpha$ H_0 is rejected.

The **typical format of an ANOVA table** provided by a statistical software is:

| <i>Source of variation</i> | <i>SS</i> | <i>df</i> | <i>MS</i> | <i>F ratio</i> | <i>P value</i> | <i>F_{critical}</i> |
|----------------------------|-----------|-----------|-----------|---------------------------|----------------------------------|-----------------------------|
| <i>Between groups</i> | SS_B | $h-1$ | MS_B | $F_0 = \frac{MS_B}{MS_W}$ | Tail area from F_0 to infinity | $F_{h-1, N-h, (1-\alpha)}$ |
| <i>Within groups</i> | SS_W | $N-h$ | MS_W | | | |
| Total | SS_T | $N-1$ | | | | |



| <i>Source of variation</i> | <i>SS</i> | <i>df</i> | <i>MS</i> | <i>F</i> | <i>P-value</i> | <i>F crit</i> |
|----------------------------|-----------|-----------|-----------|----------|----------------|---------------|
| Between-sample | 7.84 | 4 | 1.96 | 30 | 5.34E-07 | 3.056 |
| Within-sample | 0.98 | 15 | 0.0653 | | | |
| Total | 8.82 | 19 | | | | |

Fixed versus Random Factors in the Analysis of Variance

In the preceding sections the standard analysis of variance (ANOVA) for a single-factor experiment was discussed by assuming that the factor was a **fixed** factor. The term “fixed factor” means that the levels of the factor of interest could be set appropriately during the experiment.

Sometimes the levels of a factor are selected at random from a large (theoretically infinite) population of values. This leads to the **random-effects ANOVA model**.

The model can be formally expressed as before:

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

However, the treatment effects, α_i , are random variables, assumed to be normally and independently distributed, with mean zero and variance σ_α^2 .

The **variance of an observation** can be then expressed as:

$$V(X_{ij}) = V(\mu + \alpha_i + \varepsilon_{ij}) = V(\alpha_i) + V(\varepsilon_{ij}) = \sigma_\alpha^2 + \sigma^2$$

All of the computations in the random effect model are the same as in the fixed effect model, but since an entire population of treatments is studied, **it does not make much sense to formulate hypotheses about the individual factor levels selected in the experiment.**

Instead, the following hypotheses about the variance of the treatment effects are tested:

$$H_0: \sigma_{\alpha}^2 = 0$$

$$H_1: \sigma_{\alpha}^2 > 0$$

The test statistic for these hypotheses is the usual F -ratio: $MS_{Treatments}/MS_{Error}$.

If the null hypothesis is accepted there is no significant variability in the population of treatments (i.e., related to the random factor), while **if the null hypothesis is rejected there is a significant variability among the treatments in the population that was sampled.**

Notice that **the conclusions of ANOVA extend to the entire population of treatments.**

The expected mean squares in the random model are different from their counterparts in the fixed effects model. It can be shown that:

$$E(MS_{Treatments}) = \sigma^2 + n \sigma_{\alpha}^2$$

$$E(MS_{Error}) = \sigma^2$$

Frequently, the goal of an experiment involving random factors is to estimate the variance components.

A logical way to do this is to equate the expected values of the mean squares to their observed values and solve the resulting equations. This leads to:

$$\hat{\sigma}^2 = E(MS_{Error})$$

$$\hat{\sigma}_{\alpha}^2 = \frac{E(MS_{Treatments}) - E(MS_{Error})}{n}$$

This is an example of a **completely randomized single-factor experiment with four levels of the factor.**

The role of randomization in this experiment is extremely important. **By randomizing the order of the 24 runs, the effect of any nuisance variable that may influence the observed tensile strength is approximately balanced out.**

An example of nuisance variable could be **a warm-up effect on the tensile strength testing machine, leading to increasing values with operating time.**

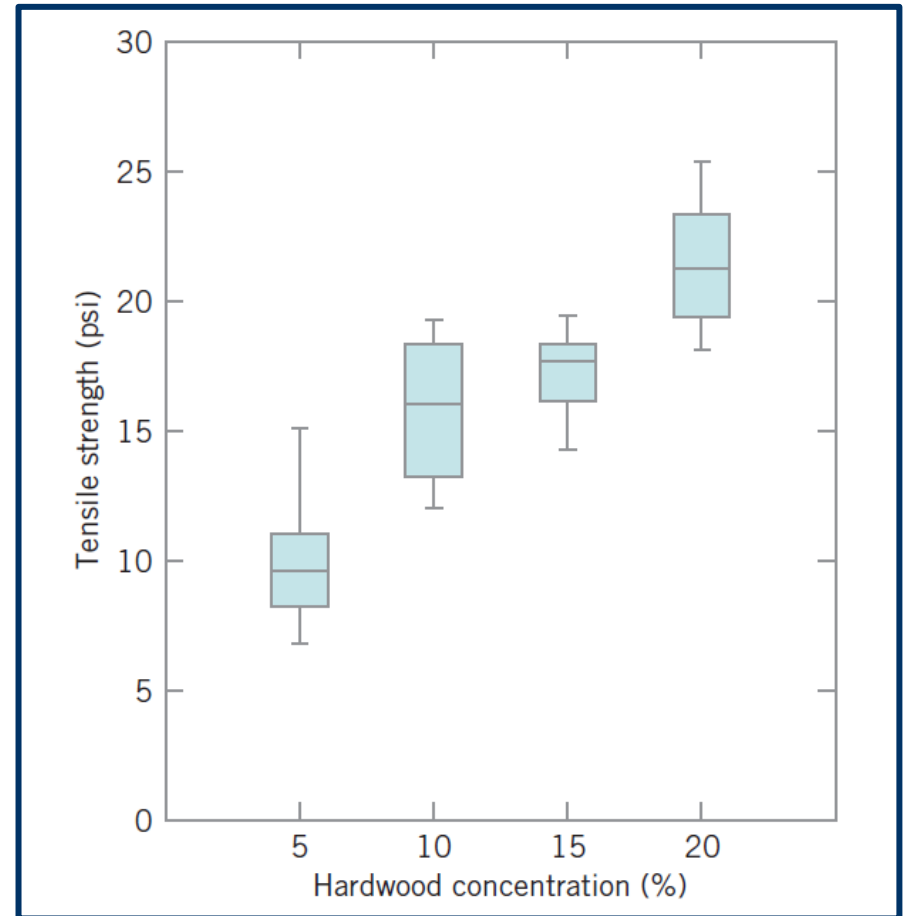
If all 24 runs were made in order of increasing hardwood concentration (that is, all six 5% concentration specimens were tested first, followed by all six 10% concentration specimens, etc.), then **the observed differences in tensile strength might also be due to the warm-up effect.**

A **Box-and-Whisker plot** can be adopted to represent data in a useful way.

The figure suggests that changing the hardwood concentration has an effect on tensile strength; specifically, higher hardwood concentrations produce higher observed tensile strengths.

Furthermore, the distribution of tensile strength at a particular hardwood level is generally not very asymmetric, and the variability in tensile strength does not seem to change dramatically as the hardwood concentration changes.

A comparison between variances can be made preliminarily to confirm this hypothesis.

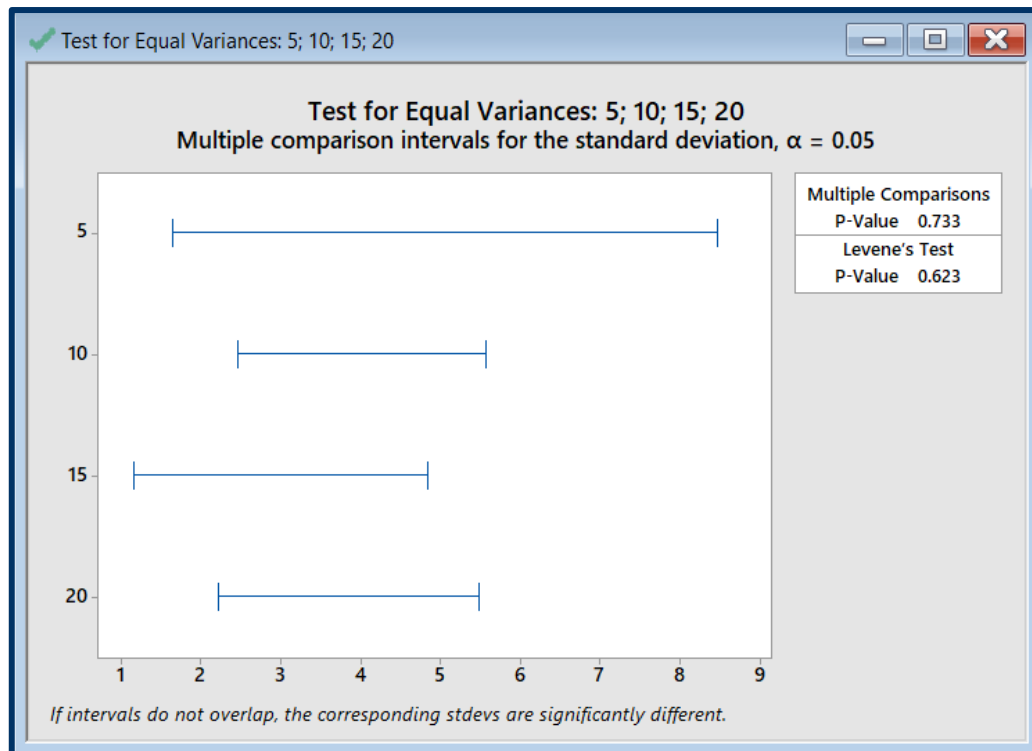


As shown before, the Minitab 18 software can be exploited to compare variances. They can be selected by accessing the “Test for equal variances” option in the Stat > ANOVA> menu.

As an example, the **Bartlett test**, assuming a normal distribution for data in each group, indicates a **p-value 0.769**, thus suggesting that no significant difference is present between the variances.

The **Levene’s test** is consistent, giving a **p-value 0.623**.

The Minitab 18’s plot reporting **confidence intervals for standard deviations of different groups**, emphasizes the fact that **an overlap occurs between them**:



Manual ANOVA calculations

$$SS_{Total} = \sum_{i=1}^h \sum_{j=1}^n X_{ij}^2 - \frac{T^2}{N} = (7)^2 + (8)^2 + \dots + (20)^2 - \frac{(383)^2}{24} = 512.96$$

$$SS_{Treatments} = \frac{\sum_{i=1}^h T_i^2}{n} - \frac{T^2}{N} = \frac{(60)^2 + (94)^2 + (102)^2 + (127)^2}{6} - \frac{(383)^2}{24} = 382.79$$

$$SS_{Error} = \sum_{i=1}^h \sum_{j=1}^n X_{ij}^2 - \frac{\sum_{i=1}^h T_i^2}{n} = 512.96 - 382.79 = 130.17$$

$$MS_{Treatments} = 382.79/3 = 127.597$$

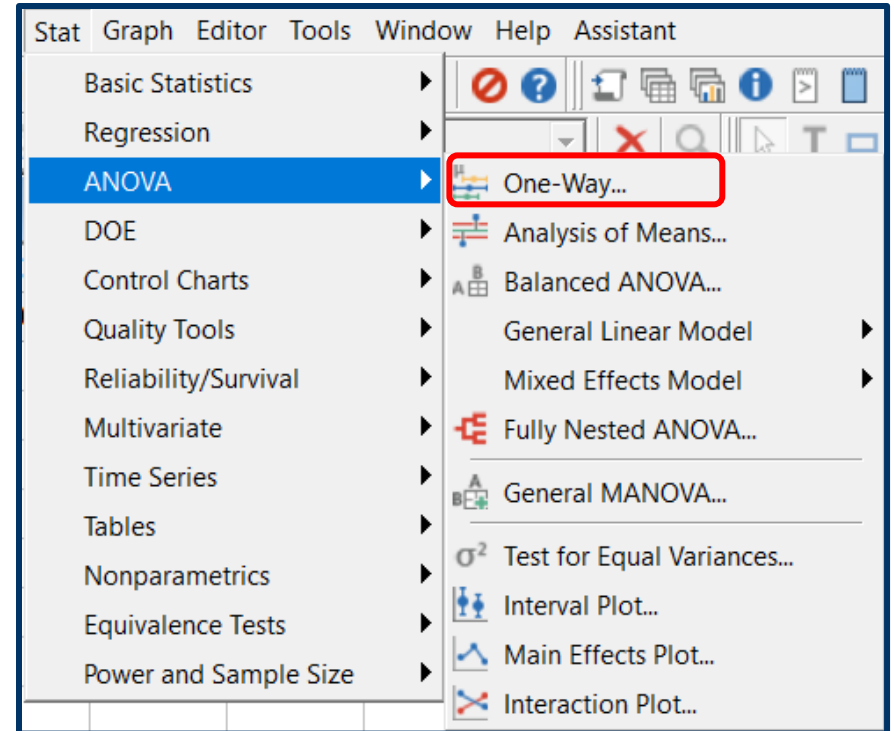
$$MS_{Error} = 130.17/20 = 6.508$$

$$F_0 = 127.597/6.508 = 19.61$$

Since the critical value for $\alpha = 0.01$ is $F_{3,20(0.99)} = 4.94$, H_0 should be rejected, thus the sample means differ significantly. This can be inferred also from the P-value, which is 3.59×10^{-6} .

ANOVA calculations with the Minitab 18 software

| Worksheet 1 *** | | | | | |
|-----------------|----|----|----|----|----|
| ↓ | C1 | C2 | C3 | C4 | C5 |
| | 5 | 10 | 15 | 20 | |
| 1 | 7 | 12 | 14 | 19 | |
| 2 | 8 | 17 | 18 | 25 | |
| 3 | 15 | 13 | 19 | 22 | |
| 4 | 11 | 18 | 17 | 23 | |
| 5 | 9 | 19 | 16 | 18 | |
| 6 | 10 | 15 | 18 | 20 | |
| 7 | | | | | |
| 8 | | | | | |



In the worksheet different factors (treatments) are represented by different columns, whereas levels (replicates) are represented by different rows.

One-Way Analysis of Variance

| | |
|----|----|
| C1 | 5 |
| C2 | 10 |
| C3 | 15 |
| C4 | 20 |

Response data are in a separate column for each factor level

Responses:
'5'-'20'

Options... Comparisons... Graphs... Results... Storage... Select Help OK Cancel

One-Way Analysis of Variance: Graphs

Data plots

- Interval plot
- Individual value plot
- Boxplot of data

Residual plots

- Individual plots
 - Histogram of residuals
 - Normal probability plot of residuals
 - Residuals versus fit
- Three in one

Help OK Cancel

One-Way Analysis of Variance: Options

Assume equal variances

Confidence level: 95 (for table of means and interval plot)

Type of confidence interval: Two-sided

Help OK Cancel

Different options (in the present case variances are assumed to be equal, based on the comparison of variances) and graphical representations of results can be selected.

One-way ANOVA: 5; 10; 15; 20

Method

Null hypothesis All means are equal
 Alternative hypothesis Not all means are equal
 Significance level $\alpha = 0.05$

Equal variances were assumed for the analysis.

Factor Information

| Factor | Levels | Values |
|--------|--------|---------------|
| Factor | 4 | 5; 10; 15; 20 |

Analysis of Variance

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|--------|----|--------|---------|---------|---------|
| Factor | 3 | 382.8 | 127.597 | 19.61 | 0.000 |
| Error | 20 | 130.2 | 6.508 | | |
| Total | 23 | 513.0 | | | |

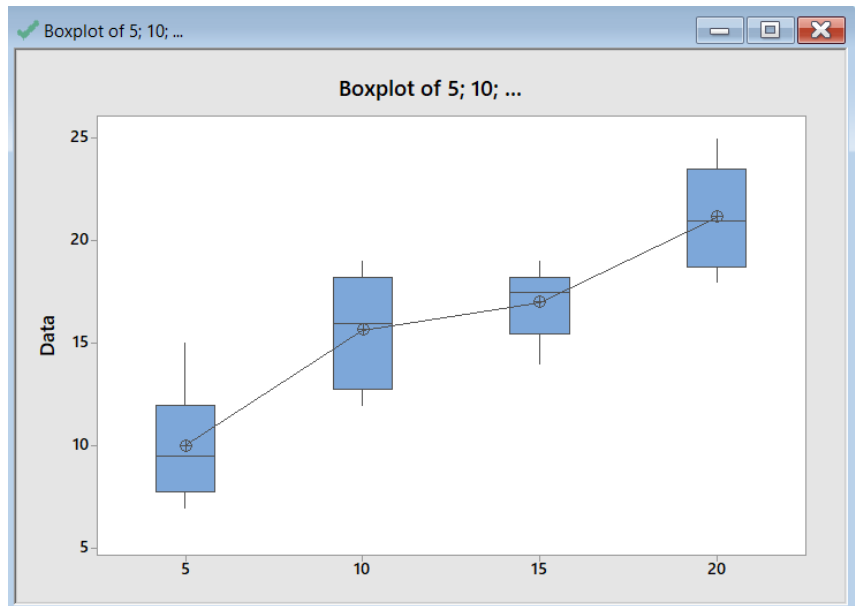
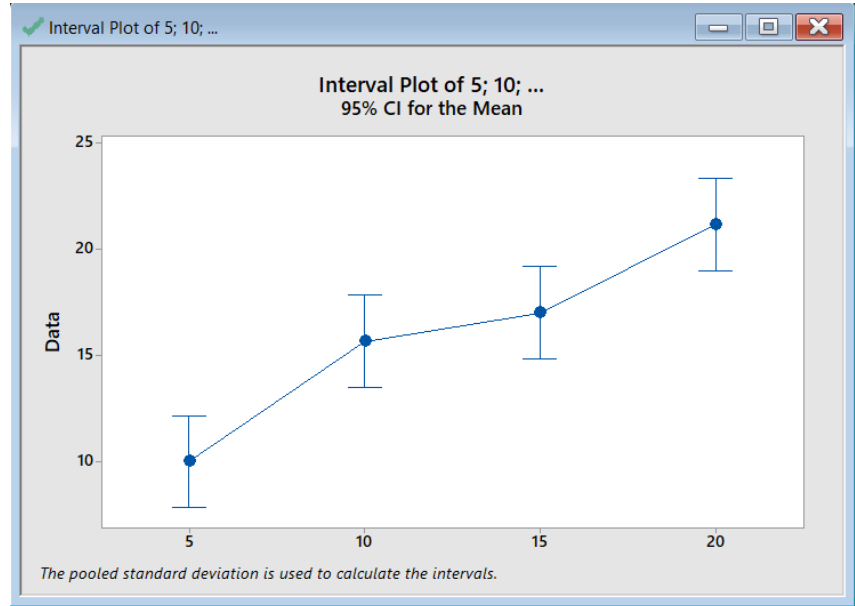
Model Summary

| S | R-sq | R-sq(adj) | R-sq(pred) |
|---------|--------|-----------|------------|
| 2.55114 | 74.62% | 70.82% | 63.46% |

Means

| Factor | N | Mean | StDev | 95% CI |
|--------|---|--------|-------|------------------|
| 5 | 6 | 10.00 | 2.83 | (7.83; 12.17) |
| 10 | 6 | 15.67 | 2.80 | (13.49; 17.84) |
| 15 | 6 | 17.000 | 1.789 | (14.827; 19.173) |
| 20 | 6 | 21.17 | 2.64 | (18.99; 23.34) |

Pooled StDev = 2.55114



2) Stability of a fluorescent reagent stored under different conditions

Table of data:

| Conditions | Replicate measurements | Mean |
|--------------------------------------|------------------------|------|
| A Freshly prepared | 102, 100, 101 | 101 |
| B Stored for 1 hour in the dark | 101, 101, 104 | 102 |
| C Stored for 1 hour in subdued light | 97, 95, 99 | 97 |
| D Stored for 1 hour in bright light | 90, 92, 94 | 92 |
| Overall mean | | 98 |

Mean squares can be calculated according to the first definition:

| Source of variation | Sum of squares | Degrees of freedom |
|---------------------|---|--------------------|
| Between-sample | $n \sum_i (\bar{x}_i - \bar{x})^2 = 186$ | $h - 1 = 3$ |
| Within-sample | $\sum_i \sum_j (x_{ij} - \bar{x}_i)^2 = 24$ | $h(n - 1) = 8$ |
| Total | $\sum_i \sum_j (x_{ij} - \bar{x})^2 = 210$ | $hn - 1 = 11$ |



$$MS_B = 62$$

$$MS_W = 3$$

Since $F_0 = 62/3 = 20.7$ and the critical value is 4.066 ($\alpha = 0.05$), H_0 should be rejected, thus the sample means differ significantly.

Minitab 18 output

One-way ANOVA: Freshly prepared; 1 hour in the dark; 1 ... bright light

Method

Null hypothesis All means are equal
Alternative hypothesis Not all means are equal
Significance level $\alpha = 0.05$

Equal variances were assumed for the analysis.

Factor Information

| Factor | Levels | Values |
|--------|--------|---|
| Factor | 4 | Freshly prepared; 1 hour in the dark; 1 hour in subdued light; 1 hour in bright light |

Analysis of Variance

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|--------|----|--------|--------|---------|---------|
| Factor | 3 | 186.00 | 62.000 | 20.67 | 0.000 |
| Error | 8 | 24.00 | 3.000 | | |
| Total | 11 | 210.00 | | | |

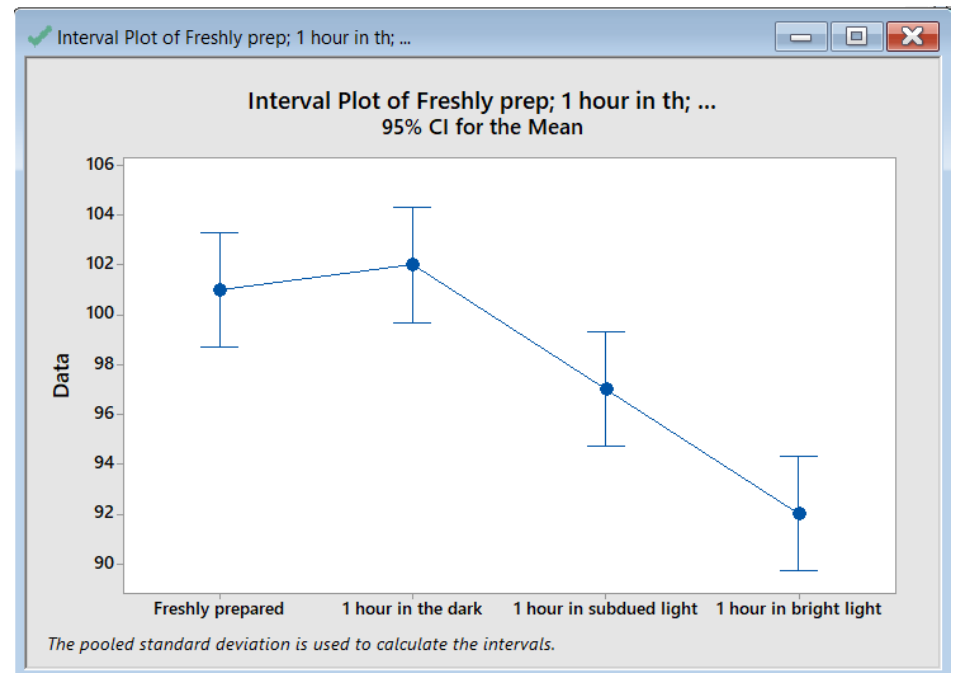
Model Summary

| S | R-sq | R-sq(adj) | R-sq(pred) |
|---------|--------|-----------|------------|
| 1.73205 | 88.57% | 84.29% | 74.29% |

Means

| Factor | N | Mean | StDev | 95% CI |
|-------------------------|---|---------|-------|-------------------|
| Freshly prepared | 3 | 101.000 | 1.000 | (98.694; 103.306) |
| 1 hour in the dark | 3 | 102.00 | 1.73 | (99.69; 104.31) |
| 1 hour in subdued light | 3 | 97.00 | 2.00 | (94.69; 99.31) |
| 1 hour in bright light | 3 | 92.00 | 2.00 | (89.69; 94.31) |

Pooled StDev = 1.73205



Note that the P value in the ANOVA table, rounded off to the third decimal place, is 0.

3) Example of ANOVA with a random-effect factor: purity testing of a barrel of sodium chloride.

The following values were obtained after replicating four times the purity testing on five samples of sodium chloride taken from different parts of a barrel, chosen at random:

| Sample | Purity (%) | Mean |
|--------|------------------------|------|
| A | 98.8, 98.7, 98.9, 98.8 | 98.8 |
| B | 99.3, 98.7, 98.8, 99.2 | 99.0 |
| C | 98.3, 98.5, 98.8, 98.8 | 98.6 |
| D | 98.0, 97.7, 97.4, 97.3 | 97.6 |
| E | 99.3, 99.4, 99.9, 99.4 | 99.5 |

In this case two possible sources of variation can be hypothesized for the observed purity:

- 1) random error in the measurement of purity, given by the measurement variance σ^2
- 2) variations in the sodium chloride purity at different points in the barrel, accounted for by sampling variance, corresponding to σ_α^2 defined before.

These are the results obtained from ANOVA calculations:

| <i>Source of variation</i> | <i>SS</i> | <i>df</i> | <i>MS</i> | <i>F</i> | <i>P-value</i> | <i>F crit</i> |
|----------------------------|-----------|-----------|-----------|----------|----------------|---------------|
| Between-sample | 7.84 | 4 | 1.96 | 30 | 5.34E-07 | 3.056 |
| Within-sample | 0.98 | 15 | 0.0653 | | | |
| Total | 8.82 | 19 | | | | |

As apparent, since the realization of the F statistic, 30, is much higher than the critical value (3.056), a significant difference exists between the purity of the different samples.

It is worth recalling that in this case the expected value for the between-sample-mean square is given by:

$$E(MS_B) = \sigma^2 + n\sigma_a^2$$

Consequently, considering that σ^2 corresponds to MS_w , the following calculation can be made to obtain an estimate of the sampling variance:

$$\sigma_a^2 = (E(MS_B) - \sigma^2)/n = (1.96 - 0.0653)/4 = 0.47$$

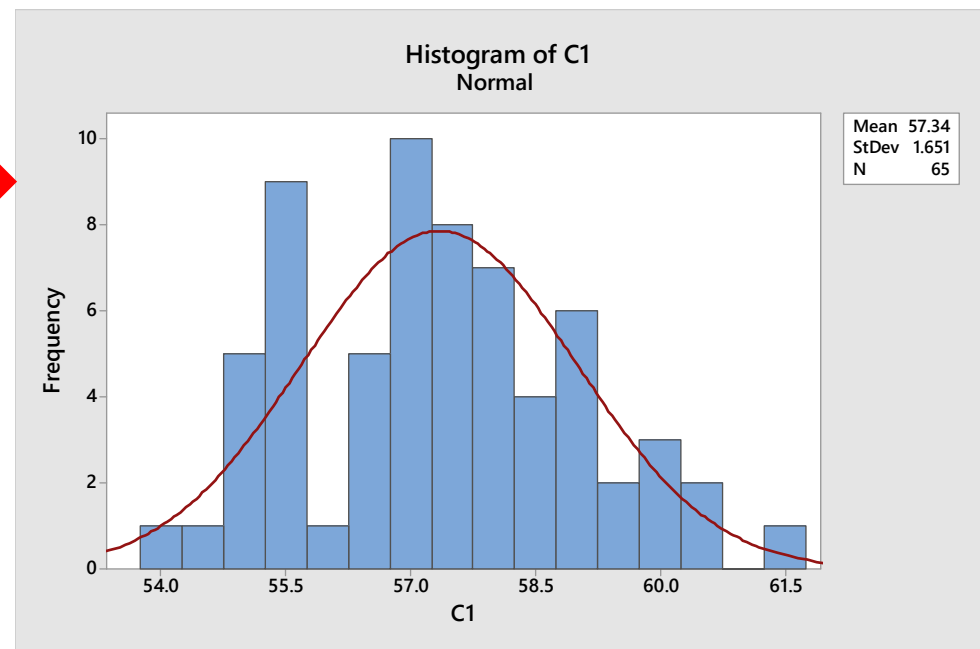
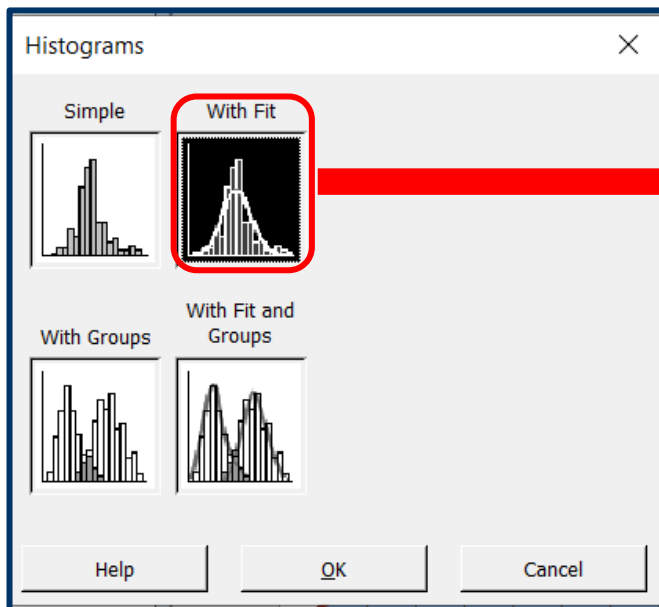
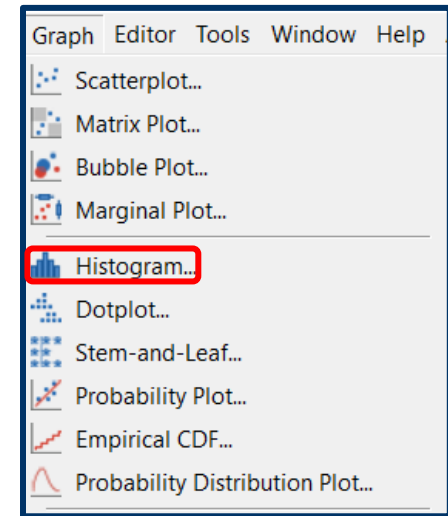
4) Example of ANOVA with a complete evaluation of assumptions

Let us reconsider the set of 65 (5 replicates obtained by each of 13 groups) enthalpy variations (kJ/mol) values measured for the neutralization of NaOH with HCl:

| Group | x_1 | x_2 | x_3 | x_4 | x_5 | \bar{x} | s^2 |
|-------|-------|-------|-------|-------|-------|-----------|-------|
| 1 | 56.9 | 59.2 | 56.3 | 58.0 | 56.9 | 57.46 | 1.32 |
| 2 | 53.8 | 55.4 | 58.0 | 59.6 | 55.5 | 56.46 | 5.34 |
| 3 | 58.4 | 55.0 | 55.7 | 56.6 | 57.2 | 56.58 | 1.74 |
| 4 | 58.0 | 56.4 | 57.6 | 57.5 | 55.0 | 56.90 | 1.48 |
| 5 | 57.7 | 58.5 | 58.9 | 57.8 | 57.4 | 58.06 | 0.38 |
| 6 | 54.8 | 56.4 | 55.2 | 60.3 | 57.1 | 56.76 | 4.76 |
| 7 | 57.1 | 60.4 | 58.9 | 55.5 | 54.7 | 57.32 | 5.55 |
| 8 | 58.6 | 57.8 | 58.0 | 55.5 | 55.6 | 57.10 | 2.09 |
| 9 | 58.9 | 59.8 | 60.0 | 57.1 | 56.4 | 58.44 | 2.61 |
| 10 | 59.5 | 57.7 | 60.0 | 57.6 | 56.8 | 58.32 | 1.86 |
| 11 | 57.2 | 58.2 | 57.4 | 55.7 | 59.1 | 57.52 | 1.60 |
| 12 | 55.4 | 56.1 | 57.7 | 56.9 | 59.2 | 57.06 | 2.17 |
| 13 | 55.1 | 56.8 | 55.7 | 61.6 | 58.3 | 57.50 | 6.74 |

As a first step, **the overall normality of data can be assessed**. A histogram of data, accompanied by a gaussian fit, can be generated for a preliminary evaluation.

In the **Minitab 18 software**, the graph can be drawn by accessing the **Graph > Histogram...** path and then selecting the **With fit** option:



In this case a gaussian probability density function with **mean = 57.34** and **standard deviation = 1.651** seems to fit experimental data appropriately.

The test for normality can be also performed using **Minitab 18** , by accessing the **Normality Test...** option in the **Stat > Basic Statistics** menu. The **Kolmogorov-Smirnov** test is available, among others.

The screenshot shows the Minitab 18 interface. The 'Stat' menu is open, and the 'Basic Statistics' submenu is displayed. The 'Normality Test...' option is highlighted with a red box. The 'Normality Test' dialog box is open, showing the following settings:

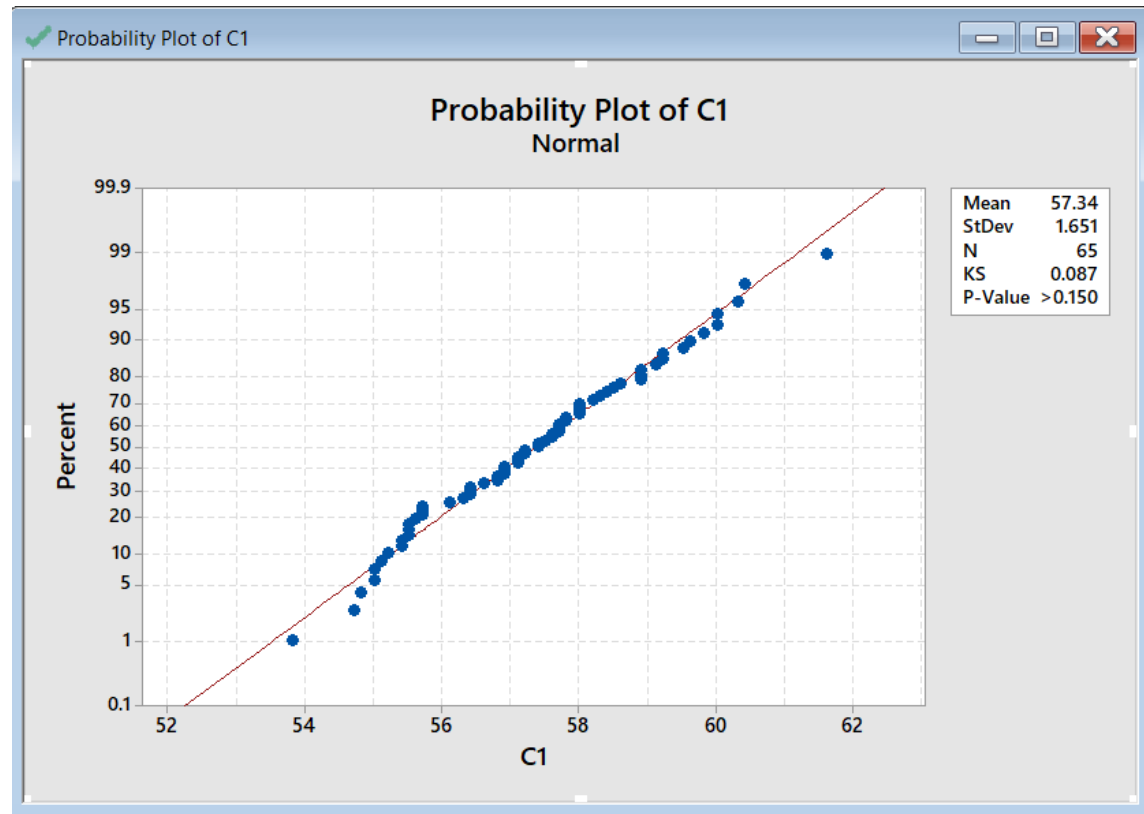
- Variable: C1
- Percentile Lines: None
- Tests for Normality: Kolmogorov-Smirnov
- Title: (empty field)

The background shows a worksheet with the following data:

| | C1 | C2 | C3 | C4 |
|----|------|----|----|----|
| 1 | 56.9 | | | |
| 2 | 53.8 | | | |
| 3 | 58.4 | | | |
| 4 | 58.0 | | | |
| 5 | 57.7 | | | |
| 6 | 54.8 | | | |
| 7 | 57.1 | | | |
| 8 | 58.6 | | | |
| 9 | 58.9 | | | |
| 10 | 59.5 | | | |

In the case of Minitab 18 the Kolmogorov-Smirnov (KS) plot is linearized by using an appropriate vertical scale.

Dots correspond to steps in the typical KS plots, whereas the red line correspond to the sigmoidal curve for the theoretical normal cumulative distribution function.



If dots are closed to the red line, as in this case, data are likely to be distributed according to a Gaussian function at a 5% significance. This outcome is confirmed, in mathematical terms, by the fact that the P-value in this case is greater than 0.150.

As for variances, the Bartlett's test provides a p-value of 0.485, whereas the Levene's test provides a p-value 0.770, thus indicating that variances related to the 13 groups of data are not significant different.

The ANOVA table generated by the Minitab 18 software shows a relatively low value for the F-Value (the name used for F_0 in the program), leading to a p-value of 0.755:

| Analysis of Variance | | | | | |
|----------------------|----|--------|--------|---------|---------|
| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
| Factor | 12 | 23.90 | 1.991 | 0.69 | 0.755 |
| Error | 52 | 150.58 | 2.896 | | |
| Total | 64 | 174.48 | | | |

As a consequence, no significant difference can be inferred between the 13 groups of data in terms of mean values.