Mathematics Learning Support Centre

# Statistics: 1.5 Oneway Analysis of Variance

Rosie Cornish. 2006.

## 1 Introduction

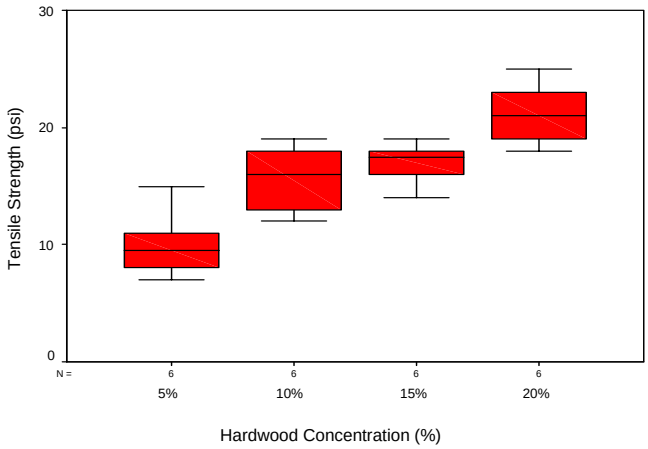
Oneway analysis of variance (ANOVA) is used to compare several means. This method is often used in scientific or medical experiments when treatments, processes, materials or products are being compared.

Example:

A paper manufacturer makes grocery bags. They are interested in increasing the tensile strength of their product. It is thought that strength is a function of the hardwood concentration in the pulp. An investigation is carried out to compare four levels of hardwood concentration: 5%, 10%, 15% and 20%. Six test specimens are made at each level and all 24 specimens are then tested in random order. The results are shown below:

| Hardwood Concentration (%) | Tensile strength (psi) | | | | | | Mean | Standard Deviation |
|---|---|---|---|---|---|---|---|---|
| 5 | 7 | 8 | 15 | 11 | 9 | 10 | 10.00 | 2.83 |
| 10 | 12 | 17 | 13 | 18 | 19 | 15 | 15.67 | 2.81 |
| 15 | 14 | 18 | 19 | 17 | 16 | 18 | 17.00 | 1.79 |
| 20 | 19 | 25 | 22 | 23 | 18 | 20 | 21.17 | 2.64 |
| All | | | | | | | 15.96 | 4.72 |

Source: Applied Statistics and Probability for Engineers - Montgomery and Runger




As stated above, in ANOVA we are asking the question, "Do all our groups come from populations with the same mean?". To answer this we need to compare the sample means. However,

even if all the population means were identical, we would not expect the sample means to be exactly equal — there will be always be some differences due to sampling variation. The question therefore becomes, "Are the observed differences between the sample means simply due to sampling variation or due to real differences in the population means?" This question cannot be answered just from the sample means — we also need to look at the variability of whatever we're measuring. In analysis of variance we compare the variability between the groups (how far apart are the means?) to the variability within the groups (how much natural variation is there in our measurements?). This is why it is called analysis of variance.

ANOVA is based on two assumptions. Therefore, before we carry out ANOVA, we need to check that these are met:

1) The observations are random samples from normal distributions.
2) The populations have the same variance, $\sigma^2$.

Fortunately, ANOVA procedures are not very sensitive to unequal variances — the following rule can be applied:

> If the largest standard deviation (not variance) is less than twice the smallest standard deviation, we can use ANOVA and our results will still be valid.

So, before carrying out any tests we must first look at the data in more detail to determine whether these assumptions are satisfied:

i) Normality: If you have very small samples, it can sometimes be quite difficult to determine whether they come from a normal distribution. However, we can assess whether the distributions are roughly symmetric by (a) comparing the group means to the medians — in a symmetric distribution these will be equal and (b) looking at boxplots or histograms of the data

ii) Equal variances: We can simply compare the group standard deviations.

In our example the medians are very close to the means. Also the standard deviations in the four groups (see table on page 1) are quite similar. These results, together with the boxplots given above, indicate that the distribution of tensile strength at each hardwood concentration is reasonably symmetric and that its variability does not change markedly from one concentration to another. Therefore, we can proceed with the analysis of variance.

# 2 The ANOVA Model

## 2.1 Notation

In general, if we sample n observations from each of $k$ populations (groups), the total number of observations is $N = nk$. The following notation is used:

- $y_{ij}$ represents the $j^{th}$ observation in group $i$ (e.g. $y_{13}$ is the 3rd observation in the first group, $y_{31}$ is the first observation in the third group, and so on).
- $\bar{y}_i$ represents the mean in group $i$
- $\bar{y}$ represents the mean of all the observations

## 2.2 Sums of squares

The total variation of all the observations about the overall mean is measured by what is called the **Total sum of squares**, given by:

$$SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y})^2$$

This variation can be split into two components:

1) the variation of the group means about the overall mean (between-group variation)
2) the variation of the individual observations about their group mean (within-group variation)

It can be shown that:

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y})^2 = n\sum_{i=1}^{k}(\bar{y}_i - \bar{y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2$$

or

$$SS_T = SS_B + SS_W$$

In words this is written as:
Total sum of squares = Between groups sum of squares + Within groups sum of squares

**Degrees of freedom and mean squares**
Each sum of squares has a certain number of degrees of freedom:

$SS_T$ compares $N$ observations to the overall mean, so has $N - 1$ degrees of freedom.
$SS_B$ compares $k$ means to the overall mean, so has $k - 1$ degrees of freedom.
$SS_W$ compares N observations to $k$ sample means, so has $N - k$ degrees of freedom.

Notice that $N - 1 = (N - k) + (k - 1)$ (i.e. the degrees of freedom are related in the same way as the sums of squares: $df_T = df_B + df_W$).

> Degrees of freedom:
> The degrees of freedom basically indicates how many 'values' are free to vary. When we are considering variances or sums of squares, because the sum of the deviations is always zero, the last deviation can be found if we know all the others. So, if we have $n$ deviations, only $n - 1$ are free to vary.

The **mean square** for each source of variation is defined as being the sum of squares divided by its degrees of freedom. Thus:

$$MS_B = SS_B/(k-1) \qquad \text{and} \qquad MS_W = SS_W/(N-k)$$

# 3 The F Test in ANOVA

It can be shown that if the null hypothesis is true and there are no differences between the (unknown) population means, $MS_B$ and $MS_W$ will be very similar. On the other hand, if the (unknown) means are different, $MS_B$ will be greater than $MS_W$ (this makes sense intuitively — if the population means are very different, we would expect the sample means to be quite far apart and therefore the between group variability will be large). Therefore, the ratio

$MS_B/MS_W$ is a statistic that is approximately equal to 1 if the null hypothesis is true but will be larger than 1 if there are differences between the population means.

The ratio $MS_B/MS_W$ is a ratio of variances, and follows what is called an $F$ distribution with $k-1$ and $N-k$ degrees of freedom. In summary:

To test $H_0 : \mu_1 = \mu_2 = ... = \mu_k = \mu$ use the statistic $F = \dfrac{MS_B}{MS_W}$ and compare this to the $F$ distribution with $k-1$ and $N-k$ degrees of freedom.

---

**The F Distribution and F Tests**

A statistical test called the **F test** is used to compare variances from two normal populations. It is tested by the **F-statistic**, the ratio of the two sample variances: $F = \dfrac{s_1^2}{s_2^2}$.

Under the null hypothesis, this statistic follows an $F$ distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom, written $F(n_1 - 1, n_2 - 1)$.

The **F distributions** are a family of distributions which depend on two parameters: the degrees of freedom of the sample variances in the numerator and denominator of the F statistic. The degrees of freedom in the numerator are always given first. The F distributions are not symmetric and, since variances cannot be negative, cannot take on values below zero. The peak of any F-distribution is close to 1; values far from 1 provide evidence against the null hypothesis. Two examples of the $F$ distribution with different degrees of freedom are shown in Figure 1.

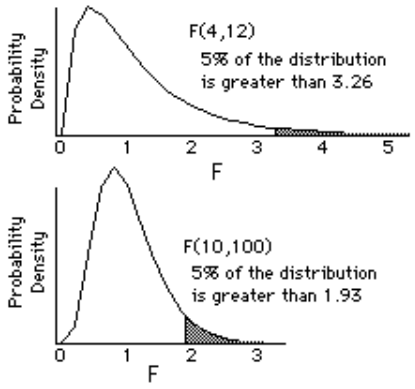



Figure 1: Probability density functions of two $F$ distributions

# 4 The ANOVA Table

When you carry out an analysis of variance on a computer, you will get an **analysis of variance** (or **ANOVA) table**, as shown below.

The ANOVA table: Tensile strength of paper

| Source of variation | Sum of squares | Degrees of freedom | Mean square | F | p |
|---|---|---|---|---|---|
| Between groups | 382.79 | 3 | 127.60 | $\frac{127.60}{6.51} = 19.61$ | $p < 0.001$ |
| Within groups | 130.17 | 20 | 6.51 | | |
| Total | 512.96 | 23 | | | |

From our results we can say that there is strong evidence that the mean tensile strength varies with hardwood concentration. Although the $F$-test does not specify the nature of these differences it is evident from our results that, as the hardwood concentration increases, so does the tensile strength of the paper. It is possible to test more specific hypotheses — for example, that there is an increasing or decreasing trend in the means — but these tests will not be covered in this leaflet.

**What exactly is the p-value?**
If the (true) mean tensile strength of paper made with different concentrations of hardwood were actually constant (i.e. if the hardwood concentration had no effect on tensile strength whatsoever), the probability of getting sample means as far apart as, or further apart than, we did (i.e. means of 10.0, 15.7, 17.0, and 21.1, or values further apart than this) is incredibly small — less than 0.001. The p-value represents this probability.
We turn this around and conclude that the true mean tensile strength is very unlikely to be constant (i.e. we conclude that the hardwood concentration does seem to have an effect on tensile strength).

Note: The ANOVA results given above are based on the assumption that the sample size in each group is equal. Usually this will be the case — most experiments are designed with equal sized samples in each experimental group. However, it is also possible to carry out an analysis of variance when the sample sizes are not equal. In this case, the formulae for the sums of squares etc. are modified to take account of the different sample sizes. If you use a statistical package to carry out your analysis, this is done automatically.

# 5 Carrying out oneway ANOVA in SPSS

– **Analyze**
– **Compare Means**
– **One Way ANOVA**
– Choose your outcome variable (in our case tensile strength) to go in **Dependent List**
– Choose the variable that defines the groups as the **Factor** then click on **OK**.

The output will look something like the ANOVA table given above.