**Analysis of Variance**

The term analysis of variance (ANOVA) describes a group of inferential statistical tests that were developed by the British statistician Sir Ronald Fisher. Whereas a t-test is used in statistics to determine if the means of two groups differ significantly, analysis of variance allows for comparison of more than two groups. ANOVA evaluates the null hypothesis that in a set of *k* independent samples (where *k* ≥ 2), all *k* samples are drawn from the same population, with the alternate hypothesis that at least two of the samples are drawn from populations with different mean values. The test statistic computed is based on the *F* distribution. In the case of two samples (i.e., a comparison of 2 means), the t-test and the F-test are equivalent (to be exact, *F* = *t*2); therefore, ANOVA is most frequently used when *k* > 2.

**One-Way Analysis of Variance**

The most commonly used ANOVA is the **one-way analysis of variance,** in which each of *k* sample means is used to estimate the value of the mean of the population from which the sample is drawn. A significant test statistic indicates that a difference exists between at least 2 of the sample means; i.e., that at least two of the samples represent populations with different mean values.

In order to calculate the test statistic for the ANOVA, the total variability in the data is divided into two groups: that which represents the difference among the means of the *k* groups (between group variability) and that which measures the variability that exists among the members in each of the *k* groups (within group variability).

*Notation for a table of data with 3 groups and 4 members in each group.*

|  |  |  |  |
| --- | --- | --- | --- |
| *i* | *Xi1* | *Xi2* | *Xi3* |
| 1 | *X11* | *X12* | *X13* |
| 2 | *X21* | *X22* | *X23* |
| 3 | *X31* | *X32* | *X33* |
| 4 | *X41* | *X42* | *X43* |
| Total | $$\sum\_{i}^{}X\_{i1}$$ | $$\sum\_{i}^{}X\_{i2}$$ | $$\sum\_{i}^{}X\_{i3}$$ |
| Mean | $$\overbar{X}\_{1}$$ | $$\overbar{X}\_{2}$$ | $$\overbar{X}\_{3}$$ |
| Grand Mean | $$̿= \frac{\left(\overbar{X}\_{1}+ \overbar{X}\_{2}+ \overbar{X}\_{3}\right)}{3}$$ |

Between-group variability estimates how much variability there is among the different groups. It is measured by calculating the “grand mean”, i.e., the overall mean of all observations across all groups, and summing the squared differences between the grand mean and the means of each individual group (calculated by averaging across group members for each of the *k*  groups).

Between-group variability = $\sum\_{k}^{} n\left(\overbar{X}\_{.k}- ̿\right)^{2}$

where *n* = the number of observations in each group

 $\overbar{X}\_{.k}=$ the mean of the *k*th group

 $̿$ = the grand mean

 ∑*k* = summation over all groups

Within group variability estimates how much variability is contained within each of the groups. It is measured by summing the squared differences between each individual observation and the mean of the group of which it is a member.

Within group variability = $\sum\_{k}^{} \sum\_{i}^{}\left(X\_{ik}-\overbar{X}\_{.k} \right)^{2}$

where $X\_{ik}$ = the value of the *i*th observation in the *k*th column

 $\overbar{X}\_{.k}$ = the mean of the *k*th group

 $\sum\_{k}^{}\sum\_{i}^{}=$ the squared deviations are first summed over all sample observations within a given group, then summed over all groups

The between-group variability and the within-group variability are the two components of total variability, which is calculated by adding the squared deviations of all the individual observations from the grand mean, $̿$.

Total variability = $\sum\_{k}^{}\sum\_{i}^{}\left(X\_{ik}- ̿ \right)^{2}$

**Degrees of Freedom**

The next step is to determine degrees of freedom (*df*) for the sums of squares calculated for the between-group and within-group variability. Adjusting the sums of square by their *df* will yield the mean squares necessary to calculate a test statistic. The number of *df* (between) is *k* – 1, where *k* is the number of groups being compared (one degree of freedom is lost when the grand mean is calculated). For the within-group variability, there are *n\*k* datapoints: *k* groups with *n* subjects in each. One *df* is lost for each group mean, *k* in all. The *df* is therefore *nk* – *k* (or, *k*(*n –* 1)). Dividing each sum of squares by its *df* yields two mean squares, the comparison of which yields the *F* ratio, which functions as the test statistic:

$$F =\frac{Mean square (between)}{Mean square (within)}$$

where mean square (between) = $\frac{\sum\_{k}^{} n\left(\overbar{X}\_{.k}- ̿\right)^{2}}{k-1}$ and

mean square (within)$ = \frac{\sum\_{k}^{}\sum\_{i}^{}\left(X\_{ik}-\overbar{X}\_{.k} \right)^{2}}{nk-k}$.

Under the null hypothesis that the group means are the same, the between-group variance will be similar to the within-group variance and *F* will approximate 1. Since within-group variability is assumed to be variability that cannot be controlled by the researcher (or, the unexplained variability), if the *k* samples are drawn from the same population, the amount of variability among the sample means will be approximately the same as the amount of variability within any one sample. When the between-group variability is larger than the within-group variability, *F* will be > 1, implying that something in addition to chance factors is contributing to the amount of variability between the sample means.

**EXAMPLE:** We wish to test whether three different anti-hypertension medications differ in effectiveness at controlling blood pressure.

Null hypothesis: µ1 = µ2 = µ3 (all three medications have the same effect)

Alternative hypothesis: The means µ1 , µ2 , and µ3 are not all equal (at least two of the medications differ in their effectiveness).

What we wish to determine is whether the differences among the sample means $\overbar{X}\_{1}, \overbar{X}\_{2}, $ and $\overbar{X}\_{3}$ are too great to be attributed to the chance errors of drawing samples from the same populations having the same means.

Having calculated a test statistic, and using the degrees of freedom determined for the numerator and denominator of the statistic, the *F*-ratio can be compared to a table of the *F* distribution (e.g., see [http://www.socr.ucla.edu/Applets.dir/F\_Table.html](http://www.socr.ucla.edu/Applets.dir/F_Table.html%20) ) or entered into an on-line calculator (one of many available can be found at <http://faculty.vassar.edu/lowry/tabs.html>) to find the level of significance. If the test statistic is significant, the conclusion that the anti-hypertensive medications differ in effectiveness can be drawn.

**Post-hoc Comparisons**

When the null hypothesis is rejected, leading us to conclude that the means of the populations from which the samples were drawn are *not* all equal, we may wish to make additional, more specific comparisons. Such comparisons are called *post hoc* if they are comprised of further (unplanned) investigations of the data after a significant effect has been found. Most of these make two comparisons at a time, while adjusting for multiple comparisons in a slightly less conservative way than the Bonferroni Correction, which divides α by the number of comparisons made. These include (among others) the following:

*Fisher’s LSD (Least Significant Difference) Test :* Multiple *t-*tests are used compare pairs of means but using the variance estimate from the ANOVA and not solely from the two groups being compared. Related to other unplanned comparison procedures, Fisher’s LSD tests is the most powerful for finding differences between pairs of means, because if does not adjust the α level needed to achieve significance in order to account for multiple testing. As a result, it has the greatest chance of resulting in one or more Type I errors.

*Tukey’s HSD (Honestly Significant Difference) Test:* This test is generally recommended when a researcher plans to make all possible pair-wise comparisons, since it controls the Type I error rate so that it will not exceed the α value pre-specified in the analysis. It maintains an acceptable level for α without an excessive loss of power. (to calculate Tukey’s HSD, see <http://web.mst.edu/~psyworld/virtualstat/tukeys/tukeyformulas.html>)

*The Newman-Keuls Test*: This test is similar to, and more powerful than, Tukey’s HSD. However, it does not control for experiment-wise error rate at α. The Newman-Keuls test cannot generate confidence intervals, which are often more informative than significance levels. For further explanation, see <http://en.wikipedia.org/wiki/Post-hoc_analysis>.

*The Scheffe Test*: This test is extremely flexible, allowing for any type of comparison, even between averages; e.g., the average of groups A and B can be compared to the average of groups C, D, and E. This increased versatility results in less power to detect differences between pairs of groups, however. It is the most conservative of the unplanned comparison procedures. The test specifies a fixed value of α which does not depend on the number of comparisons conducted.

For a good summary of the mechanics necessary, and examples of, multiple comparison tests including hypotheses, statistical formulae, and decision rules to reject the null hypothesis, see

<http://www.anselm.edu/homepage/jpitocch/anova/multcomp.html>

**Assumptions**

The one-way analysis of variance is used with interval or ratio data and is based on the following assumptions:

1. each sample has been randomly selected from the population from which it is drawn;
2. the variable being studied in the underlying population is normally distributed;
3. the variances of the populations underlying the *k*  sample are equal (**homogeneity of variance assumption)**

Although ANOVA is relatively robust to moderate departures for the assumption of normality, it becomes unreliable if variances are unequal. Before doing an ANOVA, it is important to test whether the variances in the groups are similar. This is generally done using Levene’s test or Bartlett’s test. One option, if variances are not equal among the groups, is to convert the data to ranks and use the non-parametric equivalent test, the Kruskal-Wallis. This test evaluates the hypothesis that in a set of *k* independent samples (where *k* ≥ 2), all of the samples represent populations with the same median values. The alternative hypothesis is that at least two of the samples represent populations with different median values.

**Sample Size and Power**

For fixed effects models, the following elements are necessary to calculate sample size: effect size, number of groups, and desired levels of α and β. In determining effect size, sample size calculations necessarily must take into account multiple means and the possible differences between them, as well as the potential ways in which they may be distributed in relation to each other. Because effect sizes are not as obviously determined as in two sample models, standardized measures of effect are frequently used. A number of effect size measures are available (see <http://en.wikipedia.org/wiki/Analysis_of_variance>) , but one of the most commonly used is **Cohen’s *f* .** To calculate this measure, an effect size is first calculated as follows:

*d* = (δ ÷ *s*)

where δ = the difference between the means of the groups with the highest and lowest means, and

*s* = the standard deviation of the means.

This measure is then transformed into an effect size for ANOVA by adjusting it for the assumed distribution of the means (for example, one possible distribution might assume that in a study with three groups comparing two drugs to a placebo, the means from the two drug groups may be clustered together at one end of the distribution, while the mean of the placebo group lies at the other end of the distribution). Formulas are available for the different possible patterns of means. (Cohen J (1988). *Statistical power analysis for the social sciences* (2nd ed.). New York, Academic Press.) When **Cohen’s *f*** has been calculated, a table, or computer programs, can be used to calculate sample sizes and power.

**Factorial Analysis of Variance**

Additional factors can be introduced into analysis of variance designs, such that their effects can be examined separately (**main effects**) or in combination with one another (**interactions**). The within group variance calculated in ANOVA represents the variability that is unaccounted for. Adding another factor may help explain some of this variability. In the previous example, we may suspect that gender has an effect on blood pressure control, and may interact differently with different medications. If this is the case, adding gender to the analysis may explain some of the previously unexplained variability between members in each group.

The more explanatory variables that are included, the smaller the unexplained variance will be. However, adding variables that do not account for a significant proportion of the variance can result in a less powerful test.

**Factors in Analysis of Variance**

There are two types of factors that can be included in a factorial analysis of variance:

*Fixed Effects:* A fixed factor is one for which all levels of the factor are included in the design, such as gender. If no intention is made of generalizing to other levels of a factor (for example, with race), then the inclusion of a subset of all levels of the factor may also be considered fixed.

*Random Effects:*  A random factor contains a sample of all the possible levels of a factor. The 3 anti-hypertensive medications in the previous example would be random, since they are only a sample of the many anti-hypertensive medications available. When using random factors, the presumption is made that if no difference exists between the levels sampled, then generalization to other levels of the factor may be possible. If differences are found, then generalization outside of the levels included in the analysis may not be valid.

**Assumptions**

Factorial ANOVA rests on the same assumptions as one-way ANOVA. In addition, the issue of balanced designs (equal sample sizes) must be taken into account. A design is balanced if (a) every cell has an equal number of observations or (b) if the same proportion of individuals appears at each level of a factor (for example, all levels of a factor might have the same ratio of male to females).

**Sample Size and Power**

Calculating sample size for a factorial ANOVA is generally done by using the same strategy for sample size calculations as are used for main effects as in the one-way ANOVA, with the effect or effects of most importance treated as a difference among means. This is then used to calculate a sample size.

See also:

 <http://en.wikipedia.org/wiki/Kruskal-Wallis_one-way_analysis_of_variance>).

<http://faculty.vassar.edu/lowry/ch14a.html>

<http://homepages.ucalgary.ca/~jefox/Kruskal%20and%20Wallis%201952.pdf>

References:

1. Kruskal and Wallis. [Use of ranks in one-criterion variance analysis](http://homepages.ucalgary.ca/~jefox/Kruskal%20and%20Wallis%201952.pdf). *Journal of the American Statistical Association* **47** (260): 583–621, December 1952)