

9. SINGLE FACTOR ANALYSIS OF VARIANCE (ANOVA)

The next step to consider after comparisons of means of two treatments, μ_1 and μ_2 , is comparison of means of multiple treatments: $\mu_1, \mu_2, \dots, \mu_k$. The most powerful method available is the analysis of variance (ANOVA).

The Variance Among Multiple Sample Means

The logic behind the analysis of variance is as follows. If the means of all k populations (e.g., treatment groups) are the same, as the null hypothesis claims ($H_0: \mu_1 = \mu_2 = \dots = \mu_k$ vs. H_a : at least one μ_i is different), then the sample means, $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_k$, based on a single random sample of n_i individuals from each population i , will be similar to one another. When we had only two samples, we used the difference between sample means as our measure of the size of the discrepancy between them. With more than two samples, we measure discrepancy among sample means by the **mean square between groups**:

$$MS_{\text{groups}} = \frac{\sum_{i=1}^k n_i (\bar{X}_i - \bar{\bar{X}})^2}{k - 1}$$

This quantity (also called the **mean square for treatment**) is like a variance, but it is based on squared deviations of sample means \bar{X}_i around the “grand” mean, $\bar{\bar{X}}$ (the mean of all observations from all k samples combined). If the population means are truly the same, then MS_{groups} will not be large. In particular, MS_{groups} will be similar in magnitude to the *pooled sample variance*:

$$MS_{\text{within}} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \dots + (n_k - 1)s_k^2}{N - k}$$

where N is the total sample size, the sum of the n_i (your textbook might write the equation differently, but this format is the one most familiar to you). In the jargon of analysis of variance, the pooled sample variance is called the **mean square within groups** (or **mean squared error**). A test of the null hypothesis is then carried out by comparing the ratio of these two quantities,

$$F = \frac{MS_{\text{groups}}}{MS_{\text{within}}}$$

If the null hypothesis of equal means is correct, then MS_{groups} will not be large, and this ratio will not be too different from 1. More precisely, the F ratio will follow an F -distribution with $k-1$ degrees of freedom for the numerator and $N-k$ degrees of freedom for the denominator. However, if the H_0 is false then we expect MS_{groups} to be *large*, and to exceed MS_{within} . The test involves comparing F with the critical value $F_{0.05(1),k-1,N-k}$. If F is larger than this critical value, then $P < 0.05$ and the null hypothesis is rejected. Note that we use the *one-tailed* critical value for the F -distribution. This is because the alternative hypothesis makes a directional prediction: if H_0 is false, then we expect $MS_{\text{groups}} > MS_{\text{within}}$, and therefore for F to be larger than 1. Otherwise we expect F to be near 1.

In other respects, ANOVA is just like the two-sample t -test. In fact, when $k=2$, either ANOVA or the two-sample t -test may be used (the P -values will be identical). The assumptions are the same: data must be randomly sampled from populations having normal distributions with equal variance. Normality and equality of variance can be assessed for multiple samples in the same way as for two samples (e.g. using the Levene test). ANOVA is a robust test, meaning it can tolerate a reasonable amount of deviation from these assumptions, especially when sample sizes are large and nearly equal in the different groups.

Fixed and Random Effects ANOVA

There are two main types of single factor ANOVA: **fixed effects** (Model I) and **random effects** (Model II). In a fixed effects ANOVA the treatments are specifically chosen (e.g. drug A vs. drug B vs. drug C), treatments are repeatable, and we care about the results for each treatment (e.g. Do drugs A, B and C differ in effectiveness? Which one is best?). In a random effects ANOVA, the treatments are randomly sampled from a distribution of possible treatments. Specific treatments are not repeatable and we won't usually care about the findings for individual treatments. Instead, our goal is to say something general about the population of possible treatments from which the analysed treatments are drawn. For example, to answer the question "does the mean size of offspring differ between females in a population of mice?" we would obtain a random sample of females (=treatments) from the population, breed them, and measure the sizes of each of their offspring. This will tell us about variation among females, but since the females used are simply a random sample of females from a larger population, we will not be interested in the results for individual females.

The calculations for fixed and random effects are the same for single-factor ANOVA. The calculations will differ when more complicated experiments having more than one factor are analysed.

Multiple Comparisons

Rejecting the null hypothesis of equal means only tells us that at least one of the population means is different from the others. In a **fixed effects** experiment (e.g. comparison of the effectiveness of drug A vs. drug B vs. drug C), we usually want to know more: which is the most effective? which is the least effective? are all three drugs different from one another, or does one clearly stand out from the other two? Answering these questions will require a comparison of all pairs of population means (A vs. B; A vs. C; B vs C). Of course, the whole point of using ANOVA is to avoid pairwise comparisons among pairs of means when testing for an overall difference. However, once the ANOVA has rejected H_0 , we need to return to the individual means for additional information.

We can't simply carry out a series of two-sample t tests to compare all possible pairs of treatment means. Doing so will badly inflate the probability of making at least one Type I error. The **Tukey test** was invented to circumvent this problem. The Tukey test avoids the inflation of Type I error rates by using a critical value, q , that takes account of the number of pairs of means being compared. Use of this critical value ensures that the probability of making at least one Type I error, when carrying out tests between all pairs of means, is 0.05. This "protection" comes at a price, however: the test is not very powerful. Indeed it is possible for ANOVA to

reject H_0 yet the Tukey test will not find any pairs of means that differ (usually, however, it finds at least one pair of means that differ). The Tukey test is referred to as an *a posteriori* test (one carried out after getting a specific result from another test).

We don't usually carry out Tukey tests with random effects ANOVAs for the simple reason that in this model our treatments are random and unrepeatable and we are uninterested in specific treatment differences.

Transformations

If the assumptions of ANOVA are not met, don't give up yet: consider transforming the data to achieve normality and equal variances. Many transformations are possible, but these three are the standards: log, square root, arcsin. These transformations rescale the measurements but don't otherwise distort them.

Name	Calculation	Uses
Log	$\log_e(X)$ or $\log_e(X+1)$	The most frequently-used transformation. Works for many types of data , especially data that are measured dimensions (size, length, etc.). In general, consider using when group variances are unequal but group coefficients of variation are equal. Use the natural logarithm (base e). Use $X+1$ if the data set includes zeros.
Square root	$\sqrt{X+0.5}$	Often useful for data in the form of counts , when group variances are unequal but group variance:mean ratios are equal. The addition of 0.5 is optional, but might improve the transformation when there are zeros.
Arcsine square root	$\sin^{-1}\sqrt{X}$	Used only when data are proportions (note: first divide by 100 if data are percentages). Arcsine is the inverse sine function.

The types of data for which each transformation is often used provide a guide only. Every data set is different, and a log transformation might work better for your count data set than the square root transformation. Use whatever works, but within reason. If simple transformations fail, then move on to a nonparametric alternative.

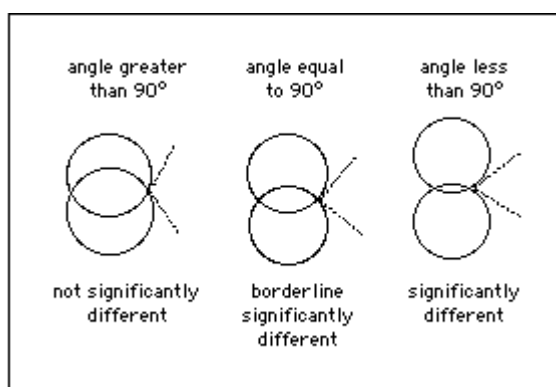
Nonparametric Alternative to ANOVA

The Kruskal-Wallis test is the best nonparametric alternative to ANOVA, and should be used if the assumptions of ANOVA cannot be met (and transformations don't solve the problem). The method is based on ranks, and is the multi-sample equivalent of the Mann-Whitney U -test (Wilcoxon rank sum test) used in the case of two samples. Under the null hypothesis, the test statistic H has an approximate chi-square distribution with $k-1$ degrees of freedom. For small total sample size, however, exact critical values may be obtained from the statistical tables in your textbook.

Using The Program

The required data format is the same as for the two-sample t -test. A category (nominal) variable indicates treatment group (this will be “ X ”). A second column contains the actual measurements (this will be “ Y ”). Use **Fit Y by X** to start the analysis, placing the appropriate variables in the X and Y boxes. This will generate a “oneway” plot in which the observations of Y are displayed for each category of X . Use **Display Options** → **Points Jittered** to spread apart overlapping observations. Click the red “▼” next to the “Oneway” title bar columns to select the following actions:

- **Means/ANOVA/T-test**: Carries out the ANOVA and generates the ANOVA table. Also calculates a 95% confidence interval for each mean (using the individual standard errors rather than the standard error based on the pooled sample variance). Adds the means diamonds to the plot (the vertical span of each diamond represents the 95% confidence interval for the mean of each group).
- **UnEqual Variances**: Carries out tests of the null hypothesis that variances of all populations are equal, using several methods including the Bartlett test (equivalent to the F test in the case of two samples) and the Levene test.
- **Compare Means** → **All Pairs, Tukey HSD**: Carries out the Tukey test between all pairs of means. A table with the results is added to the results window, and a diagram of “comparison circles” is positioned next to the oneway plot. Each circle corresponds to a sample mean. The distance between the upper and lower edges of a circle is wider than the 95% confidence interval of the mean of that group, reflecting the conservative nature of the Tukey test. Clicking one of the circles with your mouse will light up circles corresponding to all other means not significantly different from the chosen mean, based on results from the Tukey test. Overlap of comparison circles is related to statistical differences as follows (from the JMP IN manual; don’t worry if you don’t find the “angles” approach helpful, we didn’t either):



- **Nonparametric** → **Wilcoxon Test**: Carries out the Kruskal-Wallis nonparametric test (the Wilcoxon test in the case of two samples). Note that exact P -values are not provided even when sample sizes are small: JMP IN uses the normal approximation regardless of sample size, and therefore provides only an approximate P -value. This P -values may not be accurate when sample sizes are small.


Transformations: To transform data, you will need to **create a new column**. Choose **column info** and label this column to identify it as transformed data. Format the column so that it is

based on a **formula**, then in the calculator window set up the equation for the appropriate transform for your data type. The log function is located in the **trancendental** group of functions. The arcsine function is located in the **trigonometric** group of functions.

Problems

1. Open the **fruitflies.jmp** file in the shared directory. These data were seen in an earlier lab. They were collected by Partridge (1981, Nature 294: 580-581) to test whether male *Drosophila melanogaster* suffered a survival cost from mating. The life spans of individual males supplied with 1 or 8 receptive virgin females per day were compared with life spans of two types of control males. The first control consisted of two sets of individual males kept with either 1 or 8 newly inseminated females (newly inseminated females will not re-mate for at least two days, so they controlled for any effect of competition with the male for food or space). The second control was a set of individual males kept with 0 females. The four variables are:

Number of females supplied daily to males (0,1 or 8)
Experimental treatment (0, 1 or 8 virgin females, 1 or 8 newly inseminated females)
Male life span, in days
Male thorax length, in mm


 - a) Examine the histograms of male life span for each group separately (use the **By** box in the **Distribution** window to generate histograms for each treatment group all at once). Is an assumption of normality reasonable?
 - b) Use **Fit Y by X** to view the oneway plot of male lifespans of different experimental treatment groups. Use **Display Options** → **Points Jittered** to spread apart overlapping points. To the eye, do the mean lifespans appear to differ between treatment groups?
 - c) Use ANOVA to test whether experimental treatment influenced the lifespan of males fruit flies.
 - d) Is this a fixed effects ANOVA or a random effects ANOVA? Explain.
 - e) Why is it invalid to test multisample hypotheses by applying two-sample tests to all possible pairs of samples?
 - f) The ANOVA in (b) required an assumption in addition to the assumption of normal populations. Test this assumption using the appropriate method. Comment on the validity of the assumption.
 - g) Assuming that you rejected the null hypothesis in (b), determine which pairs of treatment means were significantly different. Which treatment(s) yielded the lowest mean lifespan? Which treatment(s) yielded the highest mean lifespan?
 - h) What are the assumptions of the Tukey test?
 - i) Examine the histograms of male thorax length for each group separately. On the basis of your results, recommend a strategy for testing whether there are any differences in mean

thorax length of males in different treatment groups. Implement your strategy and report the results.

2. The mimic leatherjacket, *Paraluteres prionurus*, is a small fish of the Great Barrier Reef that resembles the sharp-nosed puffer, *Canthigaster valentini*, a fish with a powerful neurotoxin in its skin. It has been suggested that the leatherjacket has evolved to resemble the toby because of the protection from predators gained. A field study tested this idea by constructing plastic replicas in the body shape of the toby and painting them one of four color patterns: “toby” (the puffer fish pattern), “leatherjacket” (the leatherjacket pattern), “1step”, and “2step” (patterns that were small and medium departures from the leatherjacket pattern, respectively, but using the same colors). Using SCUBA, the researchers tethered the replicas to line and drew them across sections of reef for a two hour period each. The number of times the model was approached by a predatory fish was recorded. This really happened! The data are in the file **mimicry.jmp**. Use these data to test for differences between color patterns in the mean number of approaches by predators.



- What are the assumptions of ANOVA? Examine the data and judge whether or not these assumptions might be met.
 - Try transforming the data to improve the validity of the assumptions. Given the type of data, which transformation would be your first choice? Carry out the transformation and reexamine the data. Are the assumptions of ANOVA met?
 - Think of a second transformation that might also work, and give it a try. Which transformation worked best?
 - After you decide on the best transformation, test whether color pattern influenced the number of approaches by predators.
 - Which color pattern(s) were most attractive to predators, and which pattern(s) were least attractive? Base your answers to these questions on a formal test.
 - Is this a fixed effects ANOVA or a random effects ANOVA? Explain.
3. Open the data file **genotype.jmp** from the shared drive. The data are weight gains of young rats separated from their natural mothers at birth and randomly reassigned to other mothers. The variables in the data set are:



Weight gain of a young rat, in g

Mother ID, the identity of the young rat’s true mother

Foster mother ID, the identity of the young rat’s foster mother

- Test whether mean weight gain of offspring differed between foster mothers. Justify all steps of your analysis.

- b) Test whether mean weight gain of offspring differed between true mothers. Justify all steps of your analysis.
- c) Suggest a biological explanation for the results in (a) and (b)
- d) Is this a fixed effects design or a random effects design?