

BIOL 458 - Biometry

LAB 6 - SINGLE FACTOR ANOVA and MULTIPLE COMPARISON PROCEDURES

PART 1: INTRODUCTION TO ANOVA

Purpose of ANOVA

Analysis of Variance (**ANOVA**) is an extremely useful statistical method with a misleading name. The purpose of **ANOVA** is not to test for differences in variances; it is to test for differences in means! However, in **ANOVA** one tests for differences in treatment means by comparing the variation among treatment means to the variation in the observations within treatments pooled among treatments. If the treatment means differ, the variation in the treatment means will be large relative to the pooled variation in the observations within treatment. If the means are not different from one another, the variation in treatment means and the variation in the observations within treatments will both be estimates of the variation among observations. In previous lab exercises, we used a t - test to test for difference between means. If just two samples were being compared the t - test would be suitable, but when an experiment involves 3 or more related treatments **ANOVA** is the appropriate test procedure.

Hypothesis testing with ANOVA

In **ANOVA**, our objective is to determine whether significant differences exist among k population means. Consequently, we will test the null hypothesis "all k population means, $\mu_1, \mu_2, \dots, \mu_k$, are equal." Thus, the null hypothesis being examined is $H_0: \mu_1 = \mu_2 = \dots = \mu_k$. Such a test is known as an "omnibus" test since the only conclusion that can be drawn is not specific about which particular treatment means differ. The alternative hypothesis of interest, H_a , is "at least one pair of the population means differ." If the differences among at least one pair of sample means are large enough to indicate differences among the corresponding population means, we will reject the null hypothesis H_0 in favor of the alternative hypothesis H_a ; otherwise, we will not reject H_0 and will conclude that there is insufficient evidence to indicate differences among the population means. Exactly which means differ is not revealed by this test.

One-way or Single-factor ANOVA

Experiments comparing two or more population means can be designed using a variety of sampling schemes. However, in this lab we shall consider only those experiments that use a sampling design in which each treatment is applied to a different or independent group of subjects. (Remember that an independent sampling design is one in which independent random samples are drawn from each of the target populations.) Furthermore, we will focus our attention on experiments that involve a single series of k related experimental treatment. Such designs are also called one-way or single factor **ANOVA** designs.

ANOVA: A Graphical Introduction

Observations from three different groups have been drawn on a number line, each indicated by a number (1-3). The means of each group and the **Grand Mean** (the mean of the ungrouped data) are shown below the line.

1 11 111 1 222 2 2 22 3 33 3 3 3 3
 1 2 GM 3

If we calculate the distance from each observation within a group to its' group mean, square them (since negative and positive distances would cancel out), sum them, and pool this value among the k treatment groups, we have what is called the within group sum of squares (**SS_W**). The **SS_W** is often referred to as the **error sum of squares (SS_e)**. If we follow a similar procedure measuring distances from the grand mean to each observation, regardless of group, we have what is called the total sum of squares (**SS_T**). If we sum the squared distances between the grand mean and each of the group means, we get the between group sum of squares (**SS_B**); sometimes referred to as the among group sum of squares, hypothesis sum of squares, or sum of squares for treatments. The total sum of squared deviations between observations and the grand mean is comprised of both the within groups variation and the between groups variation, and may be partitioned into these components; thus

$$SS_T = SS_B + SS_W,$$

$$\sum_{j=1}^k \sum_{i=1}^n (x_{ij} - \bar{x})^2 = \sum_{j=1}^k \sum_{i=1}^n (\bar{x}_j - \bar{x})^2 + \sum_{j=1}^k \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2,$$

with x_{ij} representing the i th observation in the j th treatment group, \bar{x} the Grand Mean, \bar{x}_j the j th treatment mean, n the number of observations in each treatment group, and k the number of treatments.

Examining the number line above, it is evident that these groups do have different means. It should also be evident that the average distance between an observation and its group mean is much less than the distance from the group means to the grand mean. But what about the number line below?

2 13 131 2 133 2 1 22 3 23 3 1 2 1
 1 2 GM 3

Inspection indicates that these groups are not distinct. Now the **SS_W** is almost as large as **SS_T** and **SS_B** is quite small.

Information about the sums of squares when the groups are distinct versus when they are not provides us a way to test for differences between groups. But we cannot use the sums of squares directly. What we really want to compare is the average squared distance between measures and their respective means. To obtain an average value,

we divide each sum of squares by the number of freely variable observations used in its calculation (its degrees of freedom). For \mathbf{SS}_B this is $k - 1$ where k is the number of groups. For \mathbf{SS}_W this is $(kn - k)$ or $(k(n - 1))$ where n is the total number of observations per treatment, and k , again, is the number of groups. In our example, $df_B = 2$ and $df_W = 18$, since $n = 7$ and $k = 3$.

A sum of squares divided by its degrees of freedom is called a mean square (\mathbf{MS}). If the null hypothesis of no differences between treatment means is true, then both the \mathbf{MS}_B and the \mathbf{MS}_W should be estimates of the common observational variability, and therefore should be equal. The ratio of these two mean squares, $\mathbf{MS}_B/\mathbf{MS}_W$, follows an F -distribution with numerator and denominator degrees of freedom equal to the degrees of freedom of the \mathbf{MS}_B and \mathbf{MS}_W , respectively. If the null hypothesis is true this ratio should equal one. However, if the null hypothesis is false this ratio will be large. To determine the probability of obtaining a particular value of the F -statistic, or one more extreme, one enters tables of the F -distribution for a specific α , and numerator and denominator degrees of freedom. Unlike the t -distribution that depends on a single parameter (its degrees of freedom), the shape of the F -distribution depends on two parameters - the degrees of freedom of the numerator of the ratio and the degrees of freedom of the denominator of the ratio.

Understanding ANOVA: A Numerical Example

To see how **ANOVA** works, consider the following example. The means (μ_1 and μ_2) of two populations are to be compared using independent random samples of size $n_1 = n_2 = 5$ from each population.

POP1	POP2
6	5
-1	1
0	3
3	2
2	4
$\bar{x}_1 = 2.0$	$\bar{x}_2 = 3.0$

The question at hand is “do you think these data provide sufficient evidence to indicate a difference between the population means?” One approach to test for such a difference is to examine the spread (or variation) between the sample means \bar{x}_1 and \bar{x}_2 , and to compare it to a measure of variability within the samples. The extent to which the variation between group means is greater than the variation of observations within groups, the greater will be the evidence to indicate a difference between μ_1 and μ_2 . For the data above, you can see that the difference between the sample means is small relative to the variability within the sample observations. Thus, I think you will agree that the difference between \bar{x}_1 and \bar{x}_2 is not large enough to indicate a difference between μ_1 and μ_2 .

Now consider:

POP1	POP2
2	3
2	3
2	3
2	3
2	3
$\bar{x}_1 = 2.0$	$\bar{x}_2 = 3.0$

Notice that the difference between the sample means for the data is identical to the previous example, however, since there is now no variability within the sample observations, the difference between the sample means is large compared to the variability within the sample observations. Thus, the data appear to give clear evidence of a difference between μ_1 and μ_2 .

We can apply this principle to the general problem of comparing k population means: if the variability among the k sample means is large relative to the variability within the k samples, then there is evidence to indicate that a difference exists among the k population means.

Computing the Test Statistic

As stated before, the criterion for testing the equality of means involves a comparison of two measures of variability: (1) the variation between the k sample means, and (2) the variation within the k samples.

To determine how large the value of MS_B must be in order to reject the null hypothesis, we must compare its value to the variability within the samples themselves, the MS_W . This is the test statistic for comparing k means (independent random samples); called the F - statistic. This is given by: MS_B/MS_W .

The F distribution

The ratio of two variances has a sampling distribution known as the F - distribution. The shape of the F - distribution will depend upon two quantities: numerator degrees of freedom and denominator degrees of freedom. In a one-way **ANOVA** procedure, the F - distribution has $(k - 1)$ numerator degrees of freedom and $(kn - k)$ denominator degrees of freedom.

A portion of an F - table is shown below, giving $\alpha = 0.05$ upper tail areas for different pairs of degrees of freedom. In order to establish the rejection region for our hypothesis test, we need to be able to find F values corresponding to the tail areas of this distribution. We need to find only upper - tail F - values, however, because we will only

reject H_0 if the value of the computed F - statistic is too large. Consider an F - distribution with 7 and 9 df (numerator and denominator, respectively).

$F_{(\alpha = 0.05)}$, Numerator Degrees of Freedom

Denominator DF	1	2	3	4	5	6	7	8	9
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.25	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65

We see that the F - value is given by $F_{7,9}(\alpha = 0.05) = 3.29$. Thus, the probability that the F - statistic will exceed 3.29 is $\alpha = 0.05$. Here, $F_{critical}$ is the F - value such that $P(F > F_{\alpha}) = \alpha$.

Now we are able to find the rejection region for the **ANOVA F - test**. This is best described by an example. Suppose we are comparing samples from 3 populations, thus $k = 3$; and that there are 8 measurements in each sample, thus $n = 24$. The numerator df is given by $(k - 1) = 2$, and the denominator df is given by $(kn - k) = 21$. Using $\alpha = 0.05$, we will reject the null hypothesis that the k (in this case 3) means are equal if:

$$F_{computed} > F_{critical}; \text{ in this case if } F_{computed} > F_{2,21}(\alpha = 0.05).$$

There are several assumptions that are made when performing an F - test. They should look familiar and are:

1) Independent observations

2) Homogeneous variances (the same population variance for each group)

3) Normal distribution

ANOVA is robust with respect to violation of the assumption of normality of the underlying populations, and to some extent to heterogeneity of variances. However, **ANOVA** is not robust with respect to violation of the assumption of independence of observations, or to heterogeneity of variances when the sample sizes within treatment groups are unequal.

PART 2: *a priori* AND *a posteriori* TESTS

When we conduct a one-way **ANOVA** with more than 2 groups we are testing for differences between each possible pair of groups simultaneously. However, there are times when the general question "Are at least one pair of means different?" is not really the question of interest. Similarly, sometimes after we have determined that some differences do exist between treatment groups, we want to determine which particular groups differ. These situations are referred to respectively as *a priori* testing and post hoc or *a posteriori* testing.

A. *a priori* tests

A priori tests are appropriate only when the specific hypotheses being tested can be specified prior to looking at the data. For instance, sometimes only one of many possible comparisons is of interest. In comparing insect damage at different heights in trees, we might only want to make a comparison between the highest and lowest heights. Often in experimental situations, we may want to compare the average of two experimental groups to a control. Let's look at such an example:

An experiment has been conducted where 3 different chemicals were applied to block pituitary function in rodents. The sample size was 100 individuals, 25 animals were randomly assigned to each of the three treatment groups and the control group. The experimental design is represented in the following table:

	Chemical			
	Control	Chem 1	Chem 2	Chem 3
pituitary function	μ_1	μ_2	μ_3	μ_4

The hypothesis tested by a normal one-way **ANOVA** would be that the treatment groups all have the same population mean (*i.e.*, treatment has no effect):

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4.$$

But, a more interesting question might be:

$$H_0: \mu_1 = (\mu_2 + \mu_3 + \mu_4)/3$$

or, is the control group different from the average of the 3 experimental groups. This hypothesis could also be stated as:

$$H_0: 3 \times \mu_1 + (-1 \times \mu_2) + (-1 \times \mu_3) + (-1 \times \mu_4) = 0$$

or, in shorthand: 3 -1 -1 -1.

The **SPSS** one-way program allows us to specify such special tests for a *priori* **ANOVA** testing through the use of user specified contrasts. Basically, a contrast weights the appropriate sample statistic (in this case, the sample mean) such that a specific null hypothesis is tested. In order to perform the test, we would enter the contrast coefficients (3 -1 -1 -1) using the Contrasts option in the one-way **ANOVA** procedure in **SPSS**. This procedure is found in the **Analyze Menu** under **Compare Means**.

Running this procedure would produce the output for both the "regular" **ANOVA** and special output for the contrast(s) (more than one contrast may be specified in a single one - way command). Since we specified the *a priori* contrast we would IGNORE the output for the regular **ANOVA**. This is the price we pay for asking a more specific question. What we gain is that the contrast has more power to detect the difference in which we are interested than does the omnibus **ANOVA F** - test followed by the use of a posteriori contrasts.

Since the contrasts have more power to distinguish differences, why don't we just use them instead of the **ANOVA**? The reason is that every time we test a contrast we have a 5% chance of making a mistake (assuming that we have set $\alpha = 0.05$). If we test many different contrasts, the probability of getting at least one that by chance alone is significant, is very high.

Contrasts may be divided into several classes. The user specified contrasts outlined above are linear contrasts that can test for differences between different weighted combinations of groups of means. Polynomial contrasts test for trends (but we will not deal with them here). Both of these are special cases of the general term "contrast."

There are $(k(k - 1))/2$ possible pair-wise comparisons (where k is the number of groups) and an infinite number of possible alternative contrasts for any analysis. However, since there is not an infinite amount of information content in the data, most contrasts are redundant to other contrasts. Often we want to avoid making redundant contrasts. Sets of contrasts that are not redundant are said to be orthogonal, and a set of orthogonal contrasts cannot have more than $k - 1$ tests.

An easy way to find out if a pair of contrasts is orthogonal is to calculate:

$$\sum_{i=1}^k (C_{ai} \times C_{bi}) / n_i ,$$

C_{ai} is the contrast coefficient of group i in contrast a , C_{bi} is the contrast coefficient of group i in contrast b and n_i is the sample size of group i and see if it sums to zero. If the sum is zero the contrasts are orthogonal. In the case where the sample sizes of each treatment group are identical, this simplifies to calculating:

$$\sum_{i=1}^k (C_{ai} \times C_{bi}) .$$

For example, is the following pair of contrasts orthogonal if group sizes are the same?

$$C_a \quad 1 \quad -1 \quad 0 \quad 0$$

$$C_b \quad 1 \quad 1 \quad -1 \quad -1$$

Since, $(1) \times (1) + (-1) \times (1) + (0) \times (-1) + (0) \times (-1) = 1 - 1 + 0 + 0 = 0$, these contrasts are orthogonal.

But these contrasts are not orthogonal:

$$C_a \quad 1 \quad -1 \quad 0 \quad 0$$

$$C_b \quad 1 \quad 0 \quad 0 \quad -1$$

There is no hard and fast rule that non-orthogonal contrasts may not be used, if they ask pertinent questions. However, in general, sets of orthogonal contrasts are more commonly used since they are not redundant.

B. *a posteriori* tests

Sometimes we are not able to come up with good questions without first examining the data. Such situations are common when we have performed an **ANOVA** on several groups. If we obtain a significant result, we know that the means of one or more of the groups are different from one another, but we want to know which ones. One possibility is to perform all possible comparisons of group means using a t - test. The problem with such a procedure is that the number of such comparisons can be very large and the probability of erroneously rejecting the null hypothesis at least once will be quite high (to be precise, the probability of one or more Type I errors = $1 - (1 - \alpha)^j$, where α is the per contrast significance level and j is the number of contrasts).

What we want is to be able to perform these comparisons and still be able to control the experiment - wise error rate. The experiment - wise error rate is defined as:

$$E_w = \text{probability of at least one Type I error in the set of tests performed}$$

When considering contrasts *a posteriori*, a greater penalty is assessed with respect to the comparison and experimental wise α levels. This penalty is assessed because each possible form of each contrast must be accounted for in order to compensate for the bias in choosing contrasts. The Dunn-Sidak correction for the *a posteriori* comparison - wise α is:

$$\alpha_c = 1 - (1 - \alpha_e)^{1/N},$$

where:

α_c = Dunn-Sidak comparison - wise α , the adjusted confidence level

α_e = Experiment-wise error rate

N = The number of possible forms of the contrast

In order to maintain experiment - wise error at an acceptable level a variety of *post hoc* or *a posteriori* tests have been developed. The basic strategy followed by all the procedures is to change the per comparison α_c - value such that the experiment - wise error, α_e , remains at a specified level. The **SPSS one-way** procedure provides a variety of *a posteriori* tests. Only three, however, actually control the experiment - wise error, keeping it at the specified level.

The LSDMOD procedure is the simplest example. It is an approach known as the Bonferroni procedure, and is exact for unequal group sizes. Here, you set the per comparison α_c - level at α_e / j where α_e is the desired experiment - wise error rate and j is the number of individual tests performed. For example, if we wanted to test for all possible differences between 5 groups, we would set our per comparison significance level at $\alpha_c = 0.05/10 = 0.005$. This approach is useful primarily when the number of comparisons is small. As the number of comparisons becomes large, however, the per comparison significance level becomes very small.

The TUKEY procedure (sometimes called the "Honestly significant difference" procedure) controls experiment - wise error for all contrasts performed, but not all possible contrasts. It is only approximate when group sizes are unequal.

The SCHEFFÉ procedure controls experiment - wise error for all possible contrasts. It is exact even when group sizes are unequal, but generally thought to be a conservative test.

The use of these linear contrasts or Multiple Comparisons Procedures is quite controversial and the behavior of many of the available procedures is poorly known. For further reading on which contrasts are well behaved and other issues in the use and interpretation of these procedures see the papers by Jones, Chew, and Day and Quinn in the supplemental readings. **Also, read the notes in this lab manual on "Multiple Comparisons."**

Further Instructions on Lab 6

Parts 1 and 2a of the exercise can be accomplished without SPSS. However, students invariably get the Mean Square within wrong for question 1.3 if they do not use SPSS.

To use SPSS to do a single factor ANOVA (between subjects), you first need to enter your data the same way you would for an independent groups t -test, but with the integer code column with k different integers (one for each treatment group). Go to the **ANALYZE** Menu is **SPSS**. Choose **Compare Means**, and within that submenu choose **One-Way ANOVA**. The sub-window that appears will display your variables. One column is your dependent or response variable the other is your “factor” or column of integers codes that indicate treatment group. Click your dependent variable over to the box labeled “Dependent list.” Click your variable name associated with your column of integer codes over to the “Factor” box. If you want to get descriptive statistics for your treatments (means and sd’s) then choose the “Options” button and click on the box for “descriptive statistics.”

SPSS will produce an ANOVA Table which will contain the relevant sums of squares, mean squares, degrees of freedom, F -ratios and significance values. These values are all you need for your ANOVA and the MS within is needed when performing linear contrasts (multiple comparisons).

To perform a contrast in SPSS, click on the “Contrast” button in the One-way ANOVA sub-window. Click on “Polynomial.” Set the degree of the polynomial to be “linear.” Then in the Coefficients box enter each coefficient for the contrast you are interested in. For example to compare the one treatment to the average of 3 other treatments $[(\mu_1) - (\mu_2 + \mu_3 + \mu_4)/3]$ enter the coefficients +3 -1 -1 -1 since they sum to 0 and all the coefficient in each contrast group are equal (could have used: 300 -100 -100 -100). After entering each coefficient click the “Add button.” The SPSS results will include the value of the contrasts (the actual weighted difference between the treatment means, its standard error, a t -statistic, its degrees of freedom, and its significance level. Any experimentwise error rate correction you must then apply by hand calculation of the per contrast α , and comparison of the computed α to the adjusted critical α .

Lab 6 - ASSIGNMENT

PART 1- Introduction to ANOVA

1.1) Find $F_{(\alpha=0.05)}$ for an F ratio with:

- a) Numerator df = 7, denominator df = 25
- b) Numerator df = 10, denominator df = 8
- c) Numerator df = 30, denominator df = 60

1.2) Find $F_{(\alpha)}$ for an F ratio with 15 numerator and 12 denominator df for the following values of α :

- a) $\alpha = 0.025$
- b) $\alpha = 0.050$
- c) $\alpha = 0.10$

1.3) Independent random samples were selected from three populations, shown in the table below:

<u>Sample 1</u>	<u>Sample 2</u>	<u>Sample 3</u>
2.1	4.4	1.1
3.3	2.6	0.2
0.2	3.0	2.0
	1.9	

- a) Calculate MS_B for the data. What type of variability is measured by this quantity? How many degrees of freedom are associated with this quantity?
- b) Calculate MS_w for the data. What type of variability is measured by this quantity? How many degrees of freedom are associated with this quantity?

ASSIGNMENT PART 2a; *a priori* testing

2a.1) Given contrasts *a* through *d* below and equal group sample sizes, which pairs of contrasts are orthogonal?

$$C_a \quad 1 \quad -1 \quad 0 \quad 0 \quad 0$$

$$C_b \quad 1 \quad 1 \quad 1 \quad 1 \quad -4$$

$$C_c \quad 1 \quad -1 \quad 0 \quad 1 \quad -1$$

$$C_d \quad 0 \quad 1 \quad -2 \quad 1 \quad 0$$

2a.2) Data for the pituitary function experiment can be found in the data file [pit.dat](#) (an ASCII file). Use the **SPSS** one-way program to test the following sets of orthogonal contrasts. Knowing that these contrasts were developed *a priori*, interpret the results. Include all relevant output. Describe the null hypotheses to be tested by these contrasts. Did you reject or accept these hypotheses, and why? What do these results indicate with respect to the context of the problem at hand (e.g. relate this to pituitary function, chemicals, and control).

	Grp 1	Grp 2	Grp 3	Grp 4
Contrast 1	3.0	-1.0	-1.0	-1.0
Contrast 2	1.0	0.0	0.0	-1.0
Contrast 3	-1.0	-1.0	-1.0	3.0

ASSIGNMENT PART 2b; *a posteriori* testing

2b.1) Examine the contrasts from problem 2a.2, under the assumption that these were conducted *a posteriori*. Keep the experiment - wise error rate at 5%. Compute the new comparison - wise α 's using the Dunn-Sidak correction and interpret the results in relation to these. (Hint: Set $\alpha_e = 5\%$, and solve for α_c .) Again, do you accept or reject the null hypotheses at hand? Interpret these results. Compare these results to the simple **ANOVA** (*since the tests are a posteriori* it is assumed that you examined the **ANOVA** first).

2b.2) If you were responsible for really analyzing the data from the pituitary function experiment, what procedure would you use; *a priori* contrasts, or one of the *a posteriori* procedures? Briefly justify your choice.