

March 5, 2009

Chapter 9: Analysis of Variance (ANOVA)

Recall that the two-sample t -test is used to compare the means of two populations or two experimental groups. In many experiments there are more than two conditions or treatments to compare in which case the two sample t -test will not suffice. This chapter introduces *analysis of variance* (ANOVA) which provides a classical test of equality of more than two means.

1 Terminology

We begin by introducing the terminology used in ANOVA models. An example will be used to illustrate.

Example. An experiment is carried out to see how the amount of phosphorus and genotype affect plant growth. The biomass of the plant is the measured response. Three genotypes and two phosphorus levels (high and low) were used in the experiment. There were 10 plants grown at each experimental combination.

This example is known as a 3×2 factorial experiment.

- **Response Variable:** The response or dependent variable is the biomass of the plant.
- **Factors:** Factors are the independent variables. Factors influence the response variable. In this experiment there are two factors: genotype and phosphorus amount.
- **Levels:** The levels are the different values of the individual factors. The genotype factor has three levels and the phosphorus factor has two levels (high and low). One can also consider factors with a continuum of levels. For instance, suppose we ran the experiment over a continuum of phosphorus values (instead of just high and low). If phosphorus were the only independent variable, then we would just do a regression analysis: $\text{biomass} = \beta_0 + \beta_1(\text{phos}) + \epsilon$. However, incorporating genotype (with three levels) along with a continuous variable requires an *Analysis of Covariance* (ANCOVA) which is beyond the scope of these notes.
- **Treatment:** A combination of the levels of the factors. In this experiment there are 6 possible treatments:
 - genotype 1 & low phosphorus,
 - genotype 2 & low phosphorus,
 - genotype 3 & low phosphorus,
 - genotype 1 & high phosphorus,
 - genotype 2 & high phosphorus,
 - genotype 3 & high phosphorus.
- **Replicates:** Number of experimental units (i.e. plants in this example) per treatment. In this example there are 10 replicates because there are 10 plants grown at each treatment combination.
- **Balanced Design:** Occurs when there are an equal number of replicates per treatment.

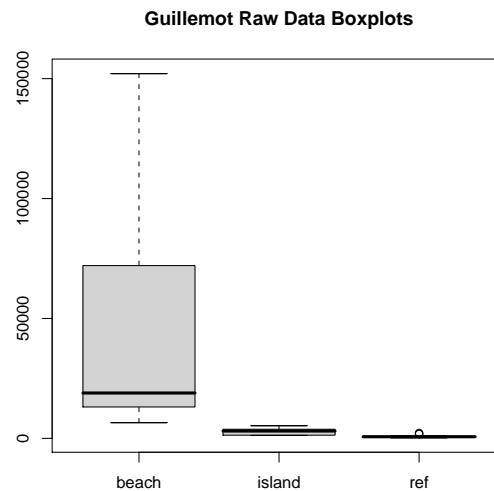


Figure 1: Boxplots of PCB concentrations in Guillemot birds at three different sites. The boxplots indicate highly skewed PCB concentration distributions at the three sites.

2 Single Factor ANOVA

An experiment with a single factor is called a single factor ANOVA or a *one-way ANOVA*. In the setting of the two-sample t -test, hypothesis testing was introduced to compare the means of two populations. In a one-way ANOVA, the classical statistical inference is to test if the means across the different levels of the factor are equal or not. Because we are typically dealing with factors with three or more levels, the two-sample t -test will not suffice. The following example will illustrate the ideas:

Example. A study was done on the liver PCB concentration in Black Guillemot birds in Canada at three different sites: A reference site, a nearby island, and the beach. Figure 1 shows a boxplot of the PCB concentrations for the birds at the three sites. The shape of the boxplots indicate strongly skewed distributions to the right which is quite common for data on concentrations of toxins. Because the data is strongly skewed, the log-transformed data was performed and the resulting boxplots are shown in Figure 2. The boxplots in Figure 2 no longer show any strong skewness. The goal of the study will be to compare the mean log-PCB lipid concentration in the birds at the three sites. The statistical inference begins with testing to see if there is any difference on average between the log(PCB) concentrations at the three sites. The boxplot in Figure 2 makes it quite clear that the average log(PCB) values differ between the three sites. We shall show how to formally test this.

Note that this data is *observational*, not experimental. That is, the data was collected by observing the birds at the three sites. In an experiment, the experimenter would assign units at randomly to the different treatments (i.e. levels of the factor in this case). Because the birds are observed at each site and not placed there, this is observational, not experimental.

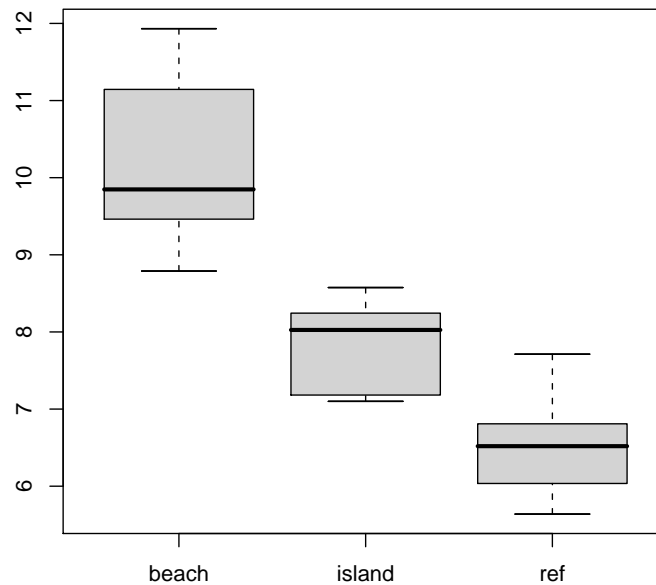


Figure 2: Boxplots of $\log(\text{PCB})$ concentrations in Gullmot birds at three different sites. The logarithm of the PCB concentration is not as skewed as the raw data.

2.1 Fixed or Random Effects?

In the Black Guillemot example above, the three sites are fixed and interest lies in comparing the $\log(\text{PCB})$ values between the three sites. Because the levels of the site factor are fixed, the model to be considered will be a *fixed effects* model. In other examples, the levels of the factor may be regarded as *random* in which case we would have a random effects model. To illustrate, suppose there were 100's of islands and we wanted to test if the mean $\log(\text{PCB})$ concentrations differ in birds among the different islands. It may be infeasible to sample birds at every island. Instead we could obtain a random sample of islands. Because the sites selected are random, this would be an example of a random effects model. The birds are clustered within the sites. Thus, data of this sort are often referred to as *clustered* data. For a one-way ANOVA, the same statistical formulas for a fixed effect and a random effects model can be the same, although the interpretation is different.

In the fixed effects example with the three fixed sites, the classic ANOVA statistical analysis is used to test the null hypothesis that the mean $\log(\text{PCB})$ levels are the same for birds at the three specific sites. In the random effects setting, if three islands are selected at random from a large number of islands for sampling, the null hypothesis is that there is no variability in the mean $\log(\text{PCB})$ at *all* the islands, not just the sampled islands. In other words, the average $\log(\text{PCB})$'s at all the islands are equal. Thus, in a random effects model, the statistical analysis allows us to make inference regarding all the islands, not just the islands selected for the sample.

In more complicated models (e.g. two-factor ANOVA), the statistical analysis can differ depending on whether or not the levels are fixed or random.

In many applications, an experiment may have both fixed and random effects. These models are known as *mixed effects models*. One of the more common examples of this occurs when repeated measurements are obtained on subjects in a study. For example, suppose a new drug is to be tested and the study will have a placebo arm as a control. A random sample of subjects are enrolled in the study. Each subject is randomized to either the drug or the placebo treatment. Each subject's response is measured repeatedly over time, say each week for 6 weeks. One factor is treatment (drug or placebo). Because the two treatments are fixed, the treatment factor is fixed. The subjects represent another factor. Because the subjects presumably represent a random sample from a population, this factor is considered random. This type of design is known as a *repeated measures design* or a *longitudinal design*. Because repeated measures on an individual subject are correlated, the statistical analysis tends to be more complicated.

The preferred method of estimating parameters and performing hypothesis testing in random and mixed effect models is *maximum likelihood*. The basic principle behind maximum likelihood estimation is to propose a model and then determine the values of the parameters that make the observed data most likely to have occurred.

2.2 The Single Factor ANOVA Model

Returning to the black Guillemot example, let μ_1, μ_2, μ_3 denote the mean log(PCB) concentration of Black Guillemot birds at the three sites respectively (reference, island, beach). For the two-sample *t*-test, the null hypothesis is generally taken to be that the two means are equal. In the one-way ANOVA with 3 levels (e.g. three sites), the null hypothesis is:

$$H_0 : \mu_1 = \mu_2 = \mu_3. \quad (1)$$

The alternative hypothesis is that the means are not all equal.

A common way to model the data is as follows: let y_{ij} denote the response of the j th bird in the i th group. Then we can write

$$y_{ij} = \mu_i + \epsilon_{ij}, \quad (2)$$

where μ_i is the mean at the i th site (a fixed effect for site in this example) for $i = 1, 2, 3$. For illustration, y_{16} is the log(PCB) value for the sixth bird at the reference site. The random error ϵ_{ij} has a mean of zero and is often assumed to have a normal distribution with variance σ^2 . The normality assumption should be checked in practice because the classic ANOVA statistical inference assumes normality, at least approximately. If there is a strong deviation from normality, then the problem needs to be addressed. Two common techniques for addressing non-normality are: (1) transform the response variable (e.g. a logarithm transformation which has been done in the bird example) or (2) use a nonparametric test that does not require the assumption of normality, see Section 2.6.

Another common model violation is that the error variance σ^2 may differ at the different levels of the factor. Simple graphical summaries of the data can usually indicate if this is a problem or not. For example, in Figure 1, the boxplots indicate that the variability of the PCB levels differ at the three sites. Often an appropriate transformation may help alleviate the unequal variance problem. In fact, the logarithm transformation is often called a variance stabilizing transformation. The unequal variance problem is not as severe for the log-transformed PCB data as indicated in the boxplots of Figure 2. However, these boxplots indicate that the variances for the log-transformed data at the three sites may still differ. We shall ignore this problem for now.

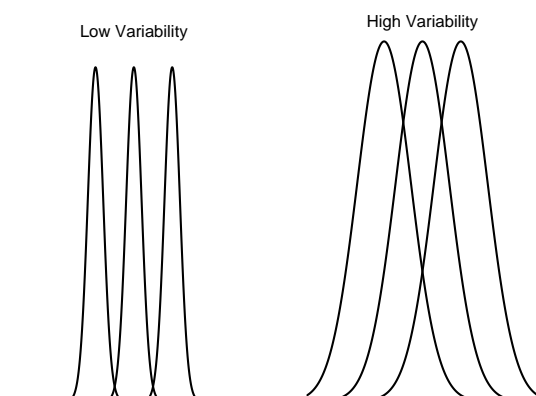


Figure 3: Left and right panels show normal distributions at three levels of the factor. The means at the three levels are the same for the right and left panel. However, the variability within each factor is much less for the left panel than for the right panel. In order to determine if population means differ from sample data, we need to analyze the *variance*, hence ANOVA

It is useful to note that the one-way ANOVA model in (2) can be parameterized in an equivalent manner as in the following model:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad (3)$$

where μ is the overall mean response and α_i is a fixed effect representing the average deviation of the responses at the i th level of the factor from the overall mean. Thus α_1 would represent how much the average log(PCB) level for birds at the reference site deviates from the overall average for all birds at all three sites. The null hypothesis (1) can be equivalently expressed as

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0. \quad (4)$$

2.3 F-Test in One-Way ANOVA

Now we look at how to use the data to test the one-way ANOVA hypothesis (1), (or (4)). It may seem odd that we call a test of equality of factor level means an *analysis of variance*. Figure 3 illustrates why we call the procedure analysis of *variance*. The left and right panels of Figure 3 show three normal densities for three levels of a factor. The variability in the left panel distributions is much less than the variability shown in the right panel. However, the factor level means in the left panel ($\mu_1 = 20, \mu_2 = 30, \mu_3 = 40$) are the same as in the right panel. If data were obtained from the three populations in the left panel, it would be likely to find a statistically significant difference because the three distributions do not overlap much. However, it would be more difficult to detect a statistically significant difference in the means in the right panel because the three distributions overlap considerably. They overlap a lot due to the high variance. Thus, in order to test for a difference in the means, we must analyze the variance. Hence the name ANOVA.

To set up the ANOVA hypothesis test procedure, we need to partition the overall variability in the response variable into two components: a component due to the variability within each factor level and the variability *between* the factor level means.

Let \bar{y}_i denote the mean of the response at level i . Let \bar{y} denote the overall mean of all the data. Look at the deviations of each individual response from the overall mean:

$$(y_{ij} - \bar{y}).$$

The total variability is given by the following *total sum of squares*

$$\text{Total Sum of Squares: } SS(\text{total}) = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 \quad (5)$$

We can rewrite the individual deviation from the mean above as:

$$(y_{ij} - \bar{y}) = (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i).$$

Squaring both sides of this equation and summing over all observations gives the *analysis of variance identity*:

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2, \quad (6)$$

where k equals the number of levels of the factor ($k = 3$ in the Black Guillemot example) and n_i equals the number of replications (or the sample size) at the i th factor level. The sums of squares in the above equation have names:

Total Sum of Squares: $\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$

Between Sum of Squares: $\sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2$

Within Sum of Squares: $\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$.

Thus, the ANOVA identity is:

$$SS(\text{Total}) = SS(\text{Between}) + SS(\text{Within}). \quad (7)$$

The logic behind the ANOVA procedure is as follows: if the means are equal across the levels of the factor, then the within group variance and the between group variances should be equal. Therefore, all we need to do in order to test the null hypothesis (1) is to compare the within and between sum of squares. However, first, these sums of squares have to be normalized to make them comparable. We can compare the within and between group variances by dividing the within and between sum of squares by the number of terms that make up the respective sums (i.e. adjust the sum of squares by their corresponding degrees of freedom). The results are called the *Mean Squares* (MS):

$$\text{MS}(\text{Between}): = \frac{SS(\text{Between})}{(k - 1)},$$

and

$$\text{MS}(\text{Within}): = \frac{SS(\text{Within})}{(n - k)},$$

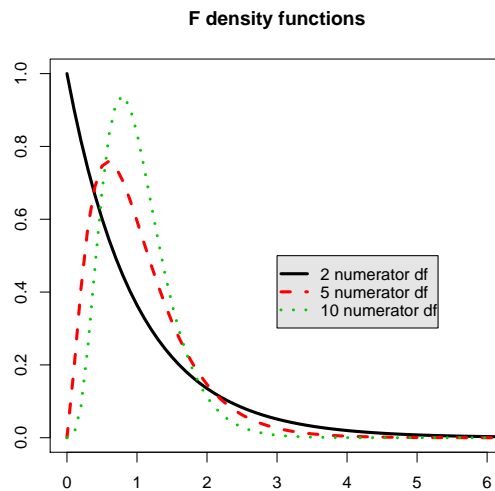


Figure 4: F density functions for numerator degrees of freedom 2, 5, and 10. The denominator degrees of freedom in each curve is 100.

where $n = n_1 + n_2 + \cdots + n_k$ denotes the overall sample size.

Note, another commonly used term for MS(Within) is the mean squared error (MSE):

$$\text{MS(Within)} = \text{MSE} = \text{MS(Error)}.$$

If the null hypothesis is true, then the sample means at each factor level should be approximately equal in which case the between mean square will yield an unbiased estimate of the error variance. Regardless of whether or not the null hypothesis is true, the within mean square is an unbiased estimate of the error variance. The test statistic for testing equality of means across the levels of the factor is called the F -test and is formed by taking the ratio of the mean squares:

$$F = \frac{\text{MS(Between)}}{\text{MS(Within)}}.$$

For the two-sample t -test, the t -distribution was our reference distribution for deciding whether or not an observed difference was too big to have happened by chance alone. In the ANOVA setting, the reference distribution is the F -distribution which is a probability distribution that takes positive values only and is skewed to the right. If H_0 is true, then the F ratio will vary according to an F probability distribution and will take the value 1 on average. However, if H_0 is false, then MS(Between) tends to be larger than the MS(Within), causing the F test statistic to become large. Thus, *we shall reject H_0 of equal factor level means when F is too big.* The question arises: what constitutes “too big.” To answer this question, we compare the F test statistic with percentiles of the F probability distribution. The F probability distribution, because it is defined as a ratio of two mean squares, is defined in terms of a numerator and a denominator degrees of freedom.

F numerator degrees of freedom = $k - 1$,

and

F denominator degrees of freedom = $n - k$.

Figure 4 shows a plot of three F density functions with numerator degrees of freedom equal to 2, 5, and 10 respectively. The denominator degrees of freedom in each case is 100. The curves are each skewed to the right.

The F -distribution has a probability density (not given here) and the percentiles of the F -distribution can be used to compute cut-off or critical values for deciding whether or not to reject H_0 or for computing p -values. The SAS software can also be used to compute needed F -values. Most common statistics textbooks also have F -tables.

The Black Guillemot bird data will now be used to illustrate the F -testing procedure using SAS. Here is the SAS program with the data:

```

/*****
Study on PCB concentration in black guillemot nestling liver samples
obtained at several different locations around Saglek Bay, Canada.
The sites at which data is collected are:
Reference, Islands, Beach
The responses are in units of ng/g for wet weight and lipid.
Source: Environmental Sciences Group, Royal Military College
*****/
options ls=76;
data pcb;
input site $ wetwt lipid;
logwetwt=log(wetwt);
loglipid=log(lipid);
datalines;
ref 11 325
ref 15 281
ref 20 412
ref 23 621
ref 25 424
ref 25 768
ref 26 904
ref 29 678
ref 33 908
ref 61 2234
ref 70 1819
island 34 1315
island 39 1315
island 49 1212
island 80 1999
island 116 3108
island 139 3011
island 141 3804
island 156 4450

```

```

island 170 3779
island 186 5301
beach 295 6564
beach 329 10460
beach 357 10650
beach 488 17230
beach 606 15540
beach 872 18920
beach 1600 30970
beach 1940 51960
beach 3070 92130
beach 5180 139190
beach 6480 152100
;
proc sort;
    by site;
proc plot;
    plot loglipid*site/ vpos=20;
proc anova;
    class site;
    model loglipid=site;
    means site/tukey;
run;

```

The output from this SAS program follows:

The ANOVA Procedure

Class Level Information

Class	Levels	Values
site	3	beach island ref

Number of Observations Read	32
Number of Observations Used	32

The ANOVA Procedure

Dependent Variable: loglipid

Sum of

Source	DF	Squares	Mean Square	F Value	Pr > F
Model	2	78.67164417	39.33582209	59.84	<.0001
Error	29	19.06336582	0.65735744		
Corrected Total	31	97.73500999			

R-Square	Coeff Var	Root MSE	loglipid Mean
0.804948	9.847244	0.810776	8.233530

Source	DF	Anova SS	Mean Square	F Value	Pr > F
site	2	78.67164417	39.33582209	59.84	<.0001

Note that the F -test statistic is

$$F = \frac{\text{MS(Between)}}{\text{MSE}} = \frac{39.3358}{0.65736} = 59.84.$$

If the null hypothesis was true (i.e. the mean log-PCB level was the same for the birds at the three sites), the F -test statistic would follow an F -distribution on 2 numerator degrees of freedom (3 sites -1) and 29 denominator degrees of freedom (total sample size minus 3 for the three sites). How likely would it be to observe an F test statistic as big as the observed value of 59.84 or bigger if the null hypothesis were true? The answer is the p -value which SAS gives as $p < 0.0001$. We can compute the p -value directly in SAS using the “prob f ” function as in the following SAS code:

```
data f;
pval = 1-probf(59.84,2,29);
proc print;
run;
```

The first argument of prob f is the F value. The 2nd and 3rd arguments are the numerator and denominator degrees of freedom. Note that the p -value is area under the upper-right tail of the F -density. Hence, to get the p -value, we need to compute “1-probf(59.84,2,29)” in SAS. The value SAS gives is extremely small, $p = 0.000000000050946!$ Thus, there is very strong statistical evidence that the mean log(PCB) values differ among the three sites. If the means were equal at the three sites, then it is very very unlikely to have observed such big differences between the three sites. Thus, we reject the null hypothesis and conclude that the mean log-PCB levels for birds at the three sites differ.

2.4 Multiple Comparisons

The F -test allowed us to conclude that the log(PCB) levels in the livers of the black Guillemot differed depending on site, but the F -test does not tell us where the differences lie. Do the mean levels differ at all

three sites, or are the PCB levels equal on average at two of the sites with one site more extreme? Often, the next step in an ANOVA after the F -test is to determine where the differences lie.

One way to approach this problem is to compute a confidence interval for the difference in the means for all possible pairs of means using the methods of the last chapter. However, this is not efficient and it can also lead to erroneous results.

Recall for the two-sample inference procedures, we pooled the data to estimate the common variance (provided the common variance assumption seemed plausible). In the ANOVA setting, we can use the MSE as a pooled (across all factor levels) estimate of the error variance:

$$\hat{\sigma}^2 = \text{MSE}.$$

Recall that if a significance level of $\alpha = 0.05$ is used in hypothesis testing, there is a probability of 0.05 committing a type I error when rejecting the null hypothesis. If our goal is to compare all pairs of means in a one-way ANOVA then there are

$$\binom{k}{2} = \frac{k(k-1)}{2}$$

possible pairwise comparisons. If the same α level is used for all comparisons, then the overall probability of rejecting at least one true null hypothesis becomes greater than α . In other words, the probability of making a type I error becomes inflated unless we make an adjustment for the *multiple comparisons*. For an analogy, suppose there is a 1% chance you get a speeding ticket on any given day that you drive to work and you drive to work every day for a year. Is the probability of getting a speeding ticket at least once during the year equal to 0.01? The answer is no! The probability of getting at least one speeding ticket will be greater than 0.01.

There are several methods for correcting for multiple comparisons in ANOVA and the two most popular methods are the **Bonferroni** and **Tukey** methods.

Bonferroni Method for Multiple Comparisons. We shall illustrate the Bonferroni method for confidence intervals first. Specify a confidence level $1 - \alpha$ for all pairwise comparisons. Then for each pair of means μ_i and μ_j , $i, j = 1, \dots, k$, $i \neq j$, the confidence interval for their difference is

$$(\bar{y}_i - \bar{y}_j) \pm t_{n-k, 1-\alpha/(2g)} \sqrt{\text{MSE}} \sqrt{1/n_i + 1/n_j}, \quad (8)$$

where

$$g = k(k-1)/2$$

is equal to the number of pairwise comparisons. Thus, to make the Bonferroni correction for multiple comparisons, all one has to do is modify the t -distribution percentile from $\alpha/2$ to $\alpha/(2g)$. Using this procedure, we can compare all pairs of means and maintain an overall level of confidence of $1 - \alpha/2$ for the entire family of comparisons. The theoretical justification for the Bonferroni procedure is based on a simple probability inequality (not given here).

Instead of forming confidence intervals for the pairwise differences, one can instead do two-tailed t -tests with significance level $\alpha/(2g)$ for each test. However, recall that the results of the two-tailed t -test are equivalent to the confidence interval approach – if zero is in the confidence interval for the difference, then we would fail to reject the null hypothesis $H_0 : \mu_i - \mu_j = 0$.

The Bonferroni method is one of the most popular multiple comparison corrections and is easy to implement. However, it is also somewhat conservative. That is, if we use the Bonferroni method and specify a

95% confidence level for the family of all pairwise comparisons, the actual confidence level will typically be higher than 95%.

Tukey Procedure for Multiple Comparisons. Another very popular method for correcting for the multiple comparisons is to use *Tukey's Studentized Range Distribution*. This method is also known as the *Honestly Significant Difference* (HSD) procedure. In the SAS program given above, a multiple comparison procedure using Tukey's method was specified. The Tukey procedure derives from the Studentized range distribution and the confidence intervals for individual pairwise differences is

$$(\bar{y}_i - \bar{y}_j) \pm q_{\alpha, n-k, k} \frac{s}{\sqrt{2}} \sqrt{1/n_i + 1/n_j}, \quad (9)$$

where $q_{\alpha, n-k, k}$ is a percentile from the studentized range distribution. In the SAS output below

$$q_{0.05, 29, 3} = 3.49263.$$

If one is interested in looking at all possible pairwise comparisons, then the Tukey procedure is generally preferred over the Bonferroni method because Tukey's procedure will lead to narrower confidence intervals using the same level of confidence as the Bonferroni procedure and hence, Tukey's procedure provides a sharper estimate of the mean differences.

On the other hand, if only a subset of all possible pairwise comparisons are of interest at the outset, i.e. $g < k(k-1)/2$ in (8), then the Bonferroni method may lead to narrower intervals than the Tukey procedure.

Caution: The transitive law of equality is not applicable to the pairwise comparison procedure. For instance, suppose $k = 3$. Then it is possible to obtain a result (using Bonferroni or Tukey) where zero is in the confidence interval for $\mu_1 - \mu_2$ and also in $\mu_2 - \mu_3$ but zero is not in the interval for $\mu_1 - \mu_3$. At first this seems paradoxical because it seems to be saying that $\mu_1 = \mu_2$ and $\mu_2 = \mu_3$ but $\mu_1 \neq \mu_3$ which violates the transitive law of equality. However, if zero is in the confidence interval for the mean difference then we do not necessarily claim the means are equal ($\mu_1 = \mu_2$ say). Instead we claim that there is no statistically significant difference between the means.

In the SAS program above, the Tukey multiple comparison procedure is specified by the option "/tukey" in the means statement. The output from this option is shown below:

```

The ANOVA Procedure
Tukey's Studentized Range (HSD) Test for loglipid
```

NOTE: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	29
Error Mean Square	0.657357
Critical Value of Studentized Range	3.49263

Comparisons significant at the 0.05 level are indicated by ***.

site Comparison	Difference	Simultaneous 95%		
	Between Means	Confidence Limits		
beach - island	2.4134	1.5385	3.2883	***
beach - ref	3.7322	2.8784	4.5860	***
island - beach	-2.4134	-3.2883	-1.5385	***
island - ref	1.3188	0.4439	2.1937	***
ref - beach	-3.7322	-4.5860	-2.8784	***
ref - island	-1.3188	-2.1937	-0.4439	***

We can summarize the results of the Tukey procedure as follows: With 95% confidence we estimate that the log(PCB) levels in black Guillemot at the beach site are 1.54 to 3.29ng/g units higher on average than for birds on the island and 2.88 to 4.59ng/g units higher on average compared to birds at the reference site. The log(PCB) levels are 0.44 to 2.19ng/g units higher on average for the birds on the island compared to birds at the the reference site. Thus, birds on the beach have the highest log(PCB) levels on average and the birds at the reference site have the lowest log(PCB) levels on average.

2.5 Power and Sample Size

Before undertaking a one-factor experiment (or any experiment!), it is important to determine the required sample size so that the F -test will have adequate power in detecting a difference between the means. The same principles that existed for sample size determination and power analysis in the t -tests hold in the ANOVA setting. Those principals are:

- Higher power requires larger sample sizes.
- Larger sample sizes lead to a more powerful F -test.
- As the effect size grows (i.e. the difference in means one wishes to detect), the power increases for a fixed sample size or the required sample size decreases for a given power.
- As the error variance decreases, the power increases for a fixed sample size. As the error variance decreases, the required sample size for a given power decreases.
- As the level of significance α decreases, the power decreases. Or, to maintain a given power, the required sample size increases as α decreases. In other words, if we want to decrease the chances of committing a type I error, then either the power of the test will decrease or a larger sample size will be required.

In the one-way ANOVA setting, matters are a bit more complicated because the power analysis and sample size computations require specifying an effect size. The question arises as to how to do this if there are k means being compared. Here are two common specifications:

(1) specify ϕ

$$\phi = \sqrt{\frac{n \sum (\mu_i - \mu)^2}{ks^2}},$$

where $\mu = (\mu_1 + \dots + \mu_k)/k$.

(2) Specify the smallest difference δ that we would like to detect between the two most different means and let

$$\phi = \sqrt{\frac{n\delta^2}{2ks^2}}.$$

As before, there are many statistical software programs that will do these computations. Older statistics textbooks have power and sample size tables as well for one-way ANOVA.

Note: If the experimenter is limited in terms of sample size and the power computations indicate that the resulting F -test and/or the power for the multiple comparisons will be low, then there are a couple possible solutions:

- Use more homogeneous experimental units to decrease the error variance and hence make the test more powerful.
- Look at the possibility of using a different experimental design such as a *repeated measures* design which is a generalization of the paired t -test. This will factor out subject to subject variation and may make the test procedure more powerful for finding differences between treatments of interest.

2.6 The Kruskal-Wallis Nonparametric Test

If the normality assumption is violated in the two-sample t -test setting, one can use a nonparametric test instead, such as the Wilcoxon rank-sum test. The analogue of the Wilcoxon rank-sum test in the one-way ANOVA setting is the *Kruskal-Wallis* test which based on the same principle. That is, pool all the data across all k levels together and then rank the data. Then analyze the data in terms of their ranks instead of their raw values.

Recall that the lipid PCB concentration distribution in the Guillemot bird example was skewed and the normality assumption was violated. Also, even after the logarithm transformation, the equal variance assumption was dubious based on the boxplots in Figure 2. Instead of analyzing the log-transformed data, we could instead perform the nonparametric Kruskal-Wallis test. In SAS, if we add the following commands to the SAS program for the Guillemot bird example above, then SAS will perform the Kruskal-Wallis nonparametric test:

```
proc npar1way wilcoxon anova median;
  class site;
  var lipid;
run;
```

The output from running this is given below:

Wilcoxon Scores (Rank Sums) for Variable lipid

site	N	Classified by Variable site			
		Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
beach	11	297.0	181.50	25.201856	27.000000
island	10	158.0	165.00	24.594494	15.800000
ref	11	73.0	181.50	25.201856	6.636364

Average scores were used for ties.

Kruskal-Wallis Test

Chi-Square	26.0031
DF	2
Pr > Chi-Square	<.0001

The results of the Kruskal-Wallis test are to reject the null hypothesis that the PCB lipid distributions are the same for the Guillemots at the three locations since the p -value < 0.0001 . This is the same conclusion that was reached using the parametric test based on the normality assumption for the log-transformed data. One can also perform multiple comparisons based on the mean ranks but details are not given here.

The bootstrap and permutation based tests are also available for performing nonparametric comparisons.

3 Two-Factor ANOVA

We have just shown how to perform an F -test in a one-way or single factor analysis of variance. Many experiments have more than one factor of interest. That is, the experimenter would like to study how a dependent variable depends on more than one factor. It is usually much more efficient to run a single experiment with multiple factors instead of running several experiments with only one-factor each. Running a factorial experiment (using more than one factor) allow the experimenter to study *interactions* between factors which we shall explain later.

Example. A study was conducted where the weights of hearts of patients with (previously) normal mitral valves prior to infective endocarditis were measured. The patients were classified based on race (black and white) and on sex (male and female). Thus, this is a two-factor study with race as one factor and sex as the other factor. In other words, this is a 2×2 factorial experiment. The response variable y is heart weight.

We can model the data from this type of experiment as follows:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, \quad (10)$$

where

- y_{ijk} is the response for the k th subject at level i of the the first factor (race) and level j of the second factor (sex).
- μ is the overall mean heart weight for the entire population.

- α_i is a parameter representing the effect for the i th level of the race factor. Thus, α_1 represents the mean effect for blacks and α_2 represents the mean effect for whites.
- β_j is parameter representing the effect for sex.
- γ_{ij} is the parameter representing the interaction effect between race and sex.
- e_{ijk} is the random error.

By convention, we have

$$\sum_{i=1}^r \alpha_i = 0, \quad \sum_{j=1}^c \beta_j = 0,$$

and

$$\sum_{i=1}^r \gamma_{ij} = 0 \text{ for all } j,$$

$$\sum_{j=1}^c \gamma_{ij} = 0 \text{ for all } i.$$

The classical statistical analysis for two-way ANOVA is done using F -tests which require the same assumptions as before: normally distributed error and independent observations.

The two-way analysis of variance procedure is more complicated than the one-way ANOVA. In the two-way ANOVA, one typically performs the following F -tests:

- Step 0:** (Optional) An overall F -test to test the null hypothesis that all the parameters in (10) are zero in which case the mean response is the same at all treatments. If we fail to reject this null hypothesis, then we can stop the analysis and conclude there are no differences in mean response among all the treatments.
- Step 1:** Test if there is a significant interaction. The null hypothesis is that all the γ_{ij} are zero in (10) and the alternative hypothesis is that the interaction terms are not all zero. If we reject this null hypothesis and conclude that there is a significant interaction, then we skip steps 2 and 3 below and generally proceed to compare treatment means using a multiple comparison procedure such as Bonferroni or Tukey. If there does not appear to be a significant interaction, then it makes sense to compare factor level means in steps 2 and 3 below.
- Step 2:** Perform an F test for the first factor. The null hypothesis is that all the α_i 's are zero. The alternative hypothesis is that they are not all zero. In the heart example, if we reject this hypothesis, then we would conclude that the average heart weights differ for the two races.
- Step 3:** Perform an F test for the second factor. The null hypothesis is that all the β_j 's are zero versus the alternative that the β_j 's are not all zero. In the heart example, rejection of this null hypothesis would lead us to conclude that average heart weights differ for males and females.

Below is a SAS program which generates the two-factor ANOVA F -test output. This program uses PROC GLM (general linear model). Note that this experiment is unbalanced because we do not have equal sample sizes at the four treatments (black males, black females, white males, white females). In such cases, one should use PROC GLM instead of PROC ANOVA. The F -test statistics corresponding to the Type III sum of squares in the SAS output are used to test the hypotheses above.

```

/*****

```

```

Unbalanced design

```

Fernicola and Roberts (Am J Cardiol. 1994) gave the data shown below for the heart weight (grams) of a convenience sample of patients with previously normal mitral valves prior to infective endocarditis. The patients were classified by race (white and black) and by gender (male, female).

```

DATA:

```

```

    sex:  1=male, 2=female

```

```

    race: 1=white 2=black

```

```

*****/

```

```

proc format;

```

```

    value sexfmt 1='Male' 2='Female';

```

```

    value racefmt 1='White' 2='Black';

```

```

run;

```

```

data heart;

```

```

input  sex race wt;

```

```

label sex = 'sex'

```

```

       race = 'race';

```

```

format sex sexfmt. race racefmt.;

```

```

datalines;

```

```

1 1 280

```

```

1 1 420

```

```

1 1 340

```

```

1 1 370

```

```

1 1 390

```

```

1 1 440

```

```

1 1 510

```

```

1 1 300

```

```

1 1 480

```

```

1 2 335

```

```

1 2 715

```

```

1 2 440

```

```

1 2 620

```

```

1 2 520

```

```

1 2 600

```

```

1 2 540

```

```

1 2 345

```

```

1 2 380

```

```

1 2 270

```

```

1 2 485

```

```

1 2 500

```

```

1 2 485

```

```

1 2 620

```

```

1 2 360

```

```

2 1 205

```

```
2 1 280
2 1 195
2 1 380
2 1 200
2 2 210
2 2 390
2 2 360
2 2 320
2 2 350
;
proc glm;
  title 'Two-Factor ANOVA';
  title2 'With an Interaction';
  class sex race;
  model wt=sex race sex*race;
  run;
proc glm;
  title 'Two-Factor ANOVA with no Interaction';
  class sex race;
  model wt = sex race;
  run;
proc glm;
  class sex race;
  model wt= sex|race;
  lsmeans sex|race;
output out=a p=pred r=resid;
run;

/* Next several lines will generate a plot of cell means useful to access
   interactions. Note: This can be done via the analyst as well */
proc means nway noprint;
class sex race;
var wt;
output out=means mean=;
proc plot;
  plot wt*sex=race;
  plot wt*race=sex;
run;
quit;
```

The output from running this program is shown below:

Two-Factor ANOVA with no Interaction

The GLM Procedure

Dependent Variable: wt

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	235823.4150	78607.8050	7.43	0.0007
Error	30	317415.5556	10580.5185		
Corrected Total	33	553238.9706			

R-Square	Coeff Var	Root MSE	wt Mean
0.426260	25.64940	102.8616	401.0294

Source	DF	Type I SS	Mean Square	F Value	Pr > F
sex	1	177800.0123	177800.0123	16.80	0.0003
race	1	57645.4327	57645.4327	5.45	0.0265
sex*race	1	377.9701	377.9701	0.04	0.8514

Source	DF	Type III SS	Mean Square	F Value	Pr > F
sex	1	150847.2009	150847.2009	14.26	0.0007
race	1	45859.5085	45859.5085	4.33	0.0460
sex*race	1	377.9701	377.9701	0.04	0.8514

The GLM Procedure
Least Squares Means

sex	wt LSMEAN
Female	289.000000
Male	436.611111

race	wt LSMEAN
Black	403.500000
White	322.111111

sex	race	wt LSMEAN
Female	Black	326.000000
Female	White	252.000000
Male	Black	481.000000
Male	White	392.222222

The p -value from the F -test for the interaction is $p = 0.8514$ indicating that there is not a significant interaction. Thus, it makes sense to test for main effects of race and sex.

The F -tests for race and sex yielded p -values of 0.0007 and 0.0460 respectively indicating that there is a statistically significant difference in mean heart weights between blacks and whites and also between males and females. The p -value for the sex factor (0.0460) is only slightly less than the 0.05 significance level that one often uses in practice.

At this point, one can construct confidence intervals for the difference in mean heart weights for blacks and whites and also the mean difference in heart weights for males and females. The SAS output above shows the estimated means.

4 Problems

1. The mean height of the *Sagittaria lancifolia* plant (which grows in the Florida Everglades) is to be compared at four different levels of phosphorus contamination in the soil. Twenty plants are monitored at each of the four soil conditions (for a total of 80 plants). Suppose the resulting data on the *Sagittaria lancifolia* plant heights is strongly skewed to the left. What is a reasonable way to analyze the data (circle the best answer):
 - a) Re-do the experiment using less phosphorus contamination in the soil.
 - b) Re-do the experiment using a test plant other than *Sagittaria lancifolia*.
 - c) Use the Kruskal-Wallis nonparametric ANOVA test.

- d) Since the data is skewed left, the normality assumption is reasonable and one can use the ANOVA F -test.
 - e) Compute the correlations between the plant heights at the different sites.
 - f) Do a paired comparison t -test.
2. Concentrations of nitrate were compared and contrasted among four major land covers: urban, forested, cropland, and pasture. Ten soil samples were obtained at each type of land cover. An ANOVA was performed to test if the mean nitrate concentrations differed among the four types of land cover. The data produced the following ANOVA table from SAS:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model		21.85			0.0113
Error		62.77			
Corrected Total					

- a) Fill in the values for the degrees of freedom (DF), total sum of squares, the Mean Squares, and the F -test statistic (F Value) in the above table.
- b) Would we have rejected the null hypothesis using $\alpha = 0.05$? (Circle One) YES or NO
- c) Would we have rejected the null hypothesis using $\alpha = 0.01$? (Circle One) YES or NO
- d) Based on the p -value, state the conclusion of the ANOVA F -test in the context of this problem.
- e) A Tukey multiple comparison procedure was conducted using 95% confidence by forming confidence intervals for the differences of the mean nitrate levels at the four sites. Letting U=urban, F=forested, C=cropland, and P=pasture, the Tukey results were summarized as follows:

C P U F,

where the underline indicates no significant difference. From this illustration, write a sentence or two describing the results of the Tukey multiple comparison procedure.

3. $\text{TNF}\alpha$, a cytokine, causes inflammation which worsens the complications due to acute pancreatitis (Source: Daniels, Biostatistics, 8th edition, page 401). An experiment was conducted on rats to determine if a bile-infusion of a $\text{TNF}\alpha$ antibody will ameliorate the effects of acute pancreatitis. The experiment had three treatment groups: 1. Sham group that received only a saline infusion. 2. Untreated Group receiving a bile infusion without treatment. 3. Treated Group receiving bile infusion with an anti- $\text{TNF}\alpha$ antibody. The response measured in this experiment is hematocrit (%) for surviving animals after 48 hours which measures the proportion, by volume, of the blood that consists of red blood cells. Use the SAS program "cytokine.sas" in the Appendix to do this problem.
 - a) Let μ_1, μ_2 and μ_3 denote the mean levels of hematocrit in rats in the three treatment groups indicated above. Write out the ANOVA null and alternative hypotheses in terms of μ_1, μ_2 and μ_3 to test if the means are equal or not.

- b) Compute the within and between mean squares using the appropriate degrees of freedoms and the sum of squares from the SAS ANOVA table.
 - c) Compute the F -test statistic from the mean squares you computed in the preceding problem.
 - d) What is the numerator and denominator degrees of freedom for the F -test statistic?
 - e) Based on the SAS output, what can one conclude from this experiment?
4. A study was done on the effects of thermal pollution on clams. Clams were collected at three sites: an intake site to a plant, a discharge site and a site near Interstate 55. The SAS program *clams.sas* in the Appendix has the data and will compute statistics regarding the heights of the clams (note that there is also data on width and length of the clams too). The goal of this problem is to determine if the clams differ in terms of heights at the three sites. Do the following parts:
- a) To test if the mean heights of the clams are equal at the three sites, define the appropriate parameters and state H_0 and H_a for this problem.
 - b) Run the SAS program and verify that the F -test statistic is the ratio of the appropriate mean squares from the ANOVA table.
 - c) If H_0 is true, what value should F take on average if the experiment were to be repeated over and over?
 - d) What are the numerator and denominator degrees of freedom for the F -test?
 - e) The SAS output gives the p -value from the F -test. Would we have rejected the null hypothesis using a significance level $\alpha = 0.05$? How about $\alpha = 0.10$?
 - f) Does it make sense to do multiple comparisons looking at differences in pairs of mean heights for this problem? Explain.
5. A study was done to examine blood coagulation times (in seconds) for animals randomly allocated to four different diets (Source: Box, Hunter, and Hunter, 1978). The data is in the SAS program *blood.sas* in the appendix. Run the SAS program to test if the mean coagulation times differ among the four diets. If a difference exists, use Tukey's procedure to compare the diets.

5 Appendix

```

/*****
TNFalpha, a cytokine, causes inflammation which worsens the complications
due to acute pancreatitis (Source: Daniels, Biostatistics, 8th edition, page 401).
An experiment was conducted on rats to determine if a bile-infusion of a TNFalpha
will ameliorate the effects of acute pancreatitis. The experiment had three
treatment groups:
1. Sham group that received only a saline infusion.
2. Untreated Group receiving a bile infusion without treatment.
3. Treated Group receiving bile infusion with an anti-TNFalpha antibody.
*****/
options ls=76;
data tnfa;
input group $ hemacrit;

```

```
datalines;
sham 38
sham 40
sham 32
sham 36
sham 40
sham 40
sham 38
sham 40
sham 38
sham 40
untreated 56
untreated 60
untreated 50
untreated 50
untreated 50
treated 40
treated 42
treated 38
treated 46
treated 36
treated 35
treated 40
treated 40
treated 55
treated 35
treated 36
treated 40
treated 40
treated 35
treated 45
;
proc sort;
    by group;
proc plot;
    plot hemacrit*group / vpos=20;
    run;
proc glm;
    class group;
    model hemacrit=group;
    means group;
    means group/tukey;
run;
```

clams.sas

```

/*****
Data from WSU grad student John Brooker (1997) from Biological Sciences investigating
the effect of thermol pollution on growth of Corbicula Fluminea (asiatic clam).
Clams were collected from 3 sites:
1=intake site, 2= discharge site, and 3=site near Interstate 55
*****/
options ls=80;
data clams;
input obs site length width height;
datalines;
1      1  7.20  6.10  4.45
2      1  7.50  5.90  4.65
3      1  6.89  5.45  4.00
4      1  6.95  5.76  4.02
5      1  6.73  5.36  3.90
6      1  7.25  5.84  4.40
7      1  7.20  5.83  4.19
8      1  6.85  5.75  3.95
9      1  7.52  6.27  4.60
10     1  7.01  5.65  4.20
11     1  6.65  5.55  4.10
12     1  7.55  6.25  4.72
13     1  7.14  5.65  4.26
14     1  7.45  6.05  4.85
15     1  7.24  5.73  4.29
16     1  7.75  6.35  4.85
17     1  6.85  6.05  4.50
18     1  6.50  5.30  3.73
19     1  6.64  5.36  3.99
20     1  7.19  5.85  4.05
21     1  7.15  6.30  4.55
22     1  7.21  6.12  4.37
23     1  7.15  6.20  4.36
24     1  7.30  6.15  4.65
25     1  6.35  5.25  3.75
26     2  7.25  6.25  4.65
27     2  7.23  5.99  4.20
28     2  6.85  5.61  4.01
29     2  7.07  5.91  4.31
30     2  6.55  5.30  3.95
31     2  7.43  6.10  4.60
32     2  7.30  5.95  4.29
33     2  6.90  5.80  4.33
34     2  7.10  5.81  4.26
35     2  6.95  5.65  4.31
36     2  7.39  6.04  4.50
37     2  6.54  5.89  3.65

```

38	2	6.39	5.00	3.72
39	2	6.08	4.80	3.51
40	2	6.30	5.05	3.69
41	2	6.35	5.10	3.73
42	2	7.34	6.45	4.55
43	2	6.70	5.51	3.89
44	2	7.08	5.81	4.34
45	2	7.09	5.95	4.39
46	2	7.40	6.25	4.85
47	2	6.00	4.75	3.37
48	2	6.94	5.63	4.09
49	2	5.95	4.75	3.20
50	3	7.60	6.45	4.56
51	3	6.15	5.05	3.50
52	3	7.00	5.80	4.30
53	3	6.81	5.61	4.22
54	3	7.10	5.75	4.10
55	3	6.85	5.55	3.89
56	3	6.68	5.50	3.90
57	3	5.51	4.52	2.70
58	3	6.85	5.53	4.00
59	3	7.10	5.80	4.45
60	3	6.81	5.45	3.51
61	3	7.30	6.00	4.31
62	3	7.05	6.25	4.71
63	3	6.75	5.65	4.00
64	3	6.75	5.57	4.06
65	3	7.35	6.21	4.29
66	3	6.22	5.11	3.35
67	3	6.80	5.81	4.50
68	3	6.29	4.95	3.69
69	3	7.55	5.93	4.55
70	3	7.45	6.19	4.70
71	3	6.70	5.55	4.00
72	3	7.51	6.20	4.74
73	3	6.95	5.69	4.29
74	3	7.50	6.20	4.65

```
;  
run;  
proc print;  
proc glm;  
  class site;  
  model width=site;  
  means site/bon;  
run;
```

Blood.sas

```
* anova.sas;
/* Blood coagulation time (in seconds) for blood drawn from 24
   animals randomly allocated to four different diets.
   (Source: Box, Hunter, and Hunter, 1978)
   *****/
options ls=76;
data blood;
input diet $ time;
cards;
A 62
A 60
A 63
A 59
B 63
B 67
B 71
B 64
B 65
B 66
C 68
C 66
C 71
C 67
C 68
C 68
D 56
D 62
D 60
D 61
D 63
D 64
D 63
D 59
;
proc print;
  title 'Blood coagulation time for four different diets';
;
proc plot;
  plot time*diet / vpos=20;
;
proc glm;
  class diet;
  model time = diet;
  means diet;
  means diet / Tukey;
run;
```