Chapter 5

One Way ANOVA

5.1 Introduction

Definition 5.1. Models in which the response variable Y is quantitative, but all of the predictor variables are qualitative are called *analysis of variance* (ANOVA) models, *experimental design* models or *design of experiments* (DOE) models. Each combination of the levels of the predictors gives a different distribution for Y. A predictor variable W is often called a factor and a factor level a_i is one of the categories W can take.

Definition 5.2. A **lurking variable** is not one of the variables in the study, but may affect the relationships among the variables in the study. A **unit** is the experimental material assigned **treatments**, which are the conditions the investigator wants to study. The unit is *experimental* if it was randomly assigned to a treatment, and the unit is *observational* if it was not randomly assigned to a treatment.

Definition 5.3. In an **experiment**, the investigators use **randomization** to assign treatments to units. To assign p treatments to $n = n_1 + \cdots + n_p$ experimental units, draw a random permutation of $\{1, ..., n\}$. Assign the first n_1 units treatment 1, the next n_2 units treatment 2, ..., and the final n_p units treatment p.

Randomization allows one to do valid inference such as F tests of hypotheses and confidence intervals. Randomization also washes out the effects of lurking variables and makes the p treatment groups similar except for the treatment. The effects of lurking variables are present in observational stud-

ies defined in Definition 5.4.

Definition 5.4. In an **observational study**, investigators simply observe the response, and the treatment groups need to be p random samples from p populations (the levels) for valid inference.

Example 5.1. Consider using randomization to assign the following nine people (units) to three treatment groups.

Carroll, Collin, Crawford, Halverson, Lawes,
Stach, Wayman, Wenslow, Xumong

Balanced designs have the group sizes the same: $n_i \equiv m = n/p$. Label the units alphabetically so Carroll gets 1, ..., Xumong gets 9. The *R/Splus* function `sample` can be used to draw a random permutation. Then the first 3 numbers in the permutation correspond to group 1, the next 3 to group 2 and the final 3 to group 3. Using the output shown below, gives the following 3 groups.

group 1: Stach, Wayman, Xumong
group 2: Lawes, Carroll, Halverson
group 3: Collin, Wenslow, Crawford

```
> sample(9)
[1] 6 7 9 5 1 4 2 8 3
```

Often there is a table or computer file of units and related measurements, and it is desired to add the unit's group to the end of the table. The *regpack* function `rand` reports a random permutation and the quantity groups[i] = treatment group for the ith person on the list. Since persons 6, 7 and 9 are in group 1, groups[7] = 1. Since Carroll is person 1 and is in group 2, groups[1] = 2, et cetera.

```
> rand(9,3)
$perm
[1] 6 7 9 5 1 4 2 8 3

$groups
[1] 2 3 3 2 2 1 1 3 1
```

Definition 5.5. Replication means that for each treatment, the n_i response variables $Y_{i,1}, ..., Y_{i,n_i}$ are approximately iid random variables.

Example 5.2. a) If ten students work two types of paper mazes three times each, then there are 60 measurements that are not replicates. Each student should work the six mazes in random order since speed increases with practice. For the ith student, let Z_{i1} be the average time to complete the three mazes of type 1, let Z_{i2} be the average time for mazes of type 2 and let $D_i = Z_{i1} - Z_{i2}$. Then $D_1, ..., D_{10}$ are replicates.

b) Cobb (1998, p. 126) states that a student wanted to know if the shapes of sponge cells depends on the color (green or white). He measured hundreds of cells from one white sponge and hundreds of cells from one green sponge. There were only two units so $n_1 = 1$ and $n_2 = 1$. The student should have used a sample of n_1 green sponges and a sample of n_2 white sponges to get more replicates.

c) Replication depends on the goals of the study. Box, Hunter and Hunter (2005, p. 215-219) describes an experiment where the investigator times how long it takes him to bike up a hill. Since the investigator is only interested in his performance, each run up a hill is a replicate (the time for the ith run is a sample from all possible runs up the hill by the investigator). If the interest had been on the effect of eight treatment levels on student bicyclists, then replication would need $n = n_1 + \cdots + n_8$ student volunteers where n_i ride their bike up the hill under the conditions of treatment i.

5.2 Fixed Effects One Way ANOVA

Definition 5.6. Let $f_Z(z)$ be the pdf of Z. Then the family of pdfs $f_Y(y) = f_Z(y - \mu)$ indexed by the *location parameter* μ, $-\infty < \mu < \infty$, is the *location family* for the random variable $Y = \mu + Z$ with *standard pdf* $f_Z(z)$.

Definition 5.7. A *one way fixed effects ANOVA model* has a single qualitative predictor variable W with p categories a_1, ..., a_p. There are p different distributions for Y, one for each category a_i. The distribution of

$$Y|(W = a_i) \sim f_Z(y - \mu_i)$$

where the location family has second moments. Hence all p distributions come from the same location family with different location parameter μ_i and the same variance σ^2.

Definition 5.8. The *one way fixed effects normal ANOVA model* is the special case where

$$Y|(W = a_i) \sim N(\mu_i, \sigma^2).$$

Example 5.3. The pooled 2 sample t–test is a special case of a one way ANOVA model with $p = 2$. For example, one population could be ACT scores for men and the second population ACT scores for women. Then $W =$ *gender* and $Y =$ *score*.

Notation. It is convenient to relabel the response variable $Y_1, ..., Y_n$ as the vector $\boldsymbol{Y} = (Y_{11}, ..., Y_{1,n_1}, Y_{21}, ..., Y_{2,n_2}, ..., Y_{p1}, ..., Y_{p,n_p})^T$ where the Y_{ij} are independent and $Y_{i1}, ..., Y_{i,n_i}$ are iid. Here $j = 1, ..., n_i$ where n_i is the number of cases from the ith level where $i = 1, ..., p$. Thus $n_1 + \cdots + n_p = n$. Similarly use double subscripts on the errors. Then there will be many equivalent parameterizations of the one way fixed effects ANOVA model.

Definition 5.9. The *cell means model* is the parameterization of the one way fixed effects ANOVA model such that

$$Y_{ij} = \mu_i + e_{ij}$$

where Y_{ij} is the value of the response variable for the jth trial of the ith factor level. The μ_i are the unknown means and $E(Y_{ij}) = \mu_i$. The e_{ij} are iid from the location family with pdf $f_Z(z)$ and unknown variance $\sigma^2 = \text{VAR}(Y_{ij}) = \text{VAR}(e_{ij})$. For the normal cell means model, the e_{ij} are iid $N(0, \sigma^2)$ for $i = 1, ..., p$ and $j = 1, ..., n_i$.

The cell means model is a linear model (without intercept) of the form $\boldsymbol{Y} = \boldsymbol{X}_c\boldsymbol{\beta}_c + \boldsymbol{e} =$

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1,n_1} \\ Y_{21} \\ \vdots \\ Y_{2,n_2} \\ \vdots \\ Y_{p,1} \\ \vdots \\ Y_{p,n_p} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} + \begin{bmatrix} e_{11} \\ \vdots \\ e_{1,n_1} \\ e_{21} \\ \vdots \\ e_{2,n_2} \\ \vdots \\ e_{p,1} \\ \vdots \\ e_{p,n_p} \end{bmatrix}. \quad (5.1)$$

Notation. Let $Y_{i0} = \sum_{j=1}^{n_i} Y_{ij}$ and let

$$\hat{\mu}_i = \overline{Y}_{i0} = Y_{i0}/n_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}. \tag{5.2}$$

Hence the "dot notation" means sum over the subscript corresponding to the 0, eg j. Similarly, $Y_{00} = \sum_{i=1}^{p}\sum_{j=1}^{n_i} Y_{ij}$ is the sum of all of the Y_{ij}.

Notice that the indicator variables used in the cell means model (5.1) are $x_h^k = 1$ if the hth case has $W = a_k$, and $x_h^k = 0$, otherwise, for $k = 1, ..., p$ and $h = 1, ..., n$. So Y_{ij} has $x_h^k = 1$ only if $i = k$ and $j = 1, ..., n_i$. Here $\boldsymbol{x}^k$ is the kth column of $\boldsymbol{X}_c$. The model can use p indicator variables for the factor instead of $p-1$ indicator variables because the model does not contain an intercept. Also notice that

$$E(\boldsymbol{Y}) = \boldsymbol{X}_c\boldsymbol{\beta}_c = (\mu_1, ..., \mu_1, \mu_2, ..., \mu_2, ..., \mu_p, ..., \mu_p)^T,$$

$(\boldsymbol{X}_c^T\boldsymbol{X}_c) = \mathrm{diag}(\mathrm{n_1}, ..., \mathrm{n_p})$ and $\boldsymbol{X}_c^T\boldsymbol{Y} = (Y_{10}, ..., Y_{10}, Y_{20}, ..., Y_{20}, ..., Y_{p0}, ..., Y_{p0})^T$. Hence $(\boldsymbol{X}_c^T\boldsymbol{X}_c)^{-1} = \mathrm{diag}(1/\mathrm{n_1}, ..., 1/\mathrm{n_p})$ and the OLS estimator

$$\hat{\boldsymbol{\beta}}_c = (\boldsymbol{X}_c^T\boldsymbol{X}_c)^{-1}\boldsymbol{X}_c^T\boldsymbol{Y} = (\overline{Y}_{10}, ..., \overline{Y}_{p0})^T = (\hat{\mu}_1, ..., \hat{\mu}_p)^T.$$

Thus $\hat{\boldsymbol{Y}} = \boldsymbol{X}_c\hat{\boldsymbol{\beta}}_c = (\overline{Y}_{10}, ..., \overline{Y}_{10}, ..., \overline{Y}_{p0}, ..., \overline{Y}_{p0})^T$. Hence the ijth fitted value is

$$\hat{Y}_{ij} = \overline{Y}_{i0} = \hat{\mu}_i \tag{5.3}$$

and the ijth residual is

$$r_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \hat{\mu}_i. \tag{5.4}$$

Since the cell means model is a linear model, there is an associated response plot and residual plot. However, many of the interpretations of the OLS quantities for ANOVA models differ from the interpretations for MLR models. First, for MLR models, the conditional distribution $Y|\boldsymbol{x}$ makes sense even if $\boldsymbol{x}$ is not one of the observed $\boldsymbol{x}_i$ provided that $\boldsymbol{x}$ is not far from the $\boldsymbol{x}_i$. This fact makes MLR very powerful. For MLR, at least one of the variables in $\boldsymbol{x}$ is a continuous predictor. For the one way fixed effects ANOVA model, the p distributions $Y|\boldsymbol{x}_i$ make sense where $\boldsymbol{x}_i^T$ is a row of $\boldsymbol{X}_c$.

Also, the OLS MLR ANOVA F test for the cell means model tests $H_0 : \boldsymbol{\beta} = 0 \equiv H_0 : \mu_1 = \cdots = \mu_p = 0$, while the one way fixed effects ANOVA F test given after Definition 5.13 tests $H_0 : \mu_1 = \cdots = \mu_p$.

Definition 5.10. Consider the one way fixed effects ANOVA model. The *response plot* is a plot of $\hat{Y}_{ij} \equiv \hat{\mu}_i$ versus Y_{ij} and the *residual plot* is a plot of $\hat{Y}_{ij} \equiv \hat{\mu}_i$ versus r_{ij}.

The points in the response plot scatter about the identity line and the points in the residual plot scatter about the $r = 0$ line, but the scatter need not be in an evenly populated band. A *dot plot* of $Z_1, ..., Z_m$ consists of an axis and m points each corresponding to the value of Z_i. The response plot consists of p dot plots, one for each value of $\hat{\mu}_i$. The dot plot corresponding to $\hat{\mu}_i$ is the dot plot of $Y_{i1}, ..., Y_{i,n_i}$. The p dot plots should have roughly the same amount of spread, and each $\hat{\mu}_i$ corresponds to level a_i. If a new level a_f corresponding to $\boldsymbol{x}_f$ was of interest, hopefully the points in the response plot corresponding to a_f would form a dot plot at $\hat{\mu}_f$ similar in spread to the other dot plots, but it may not be possible to predict the value of $\hat{\mu}_f$. Similarly, the residual plot consists of p dot plots, and the plot corresponding to $\hat{\mu}_i$ is the dot plot of $r_{i1}, ..., r_{i,n_i}$.

Assume that each $n_i \geq 10$. Under the assumption that the Y_{ij} are from the same location scale family with different parameters μ_i, each of the p dot plots should have roughly the same shape and spread. This assumption is easier to judge with the residual plot. If the response plot looks like the residual plot, then a horizontal line fits the p dot plots about as well as the identity line, and there is not much difference in the μ_i. If the identity line is clearly superior to any horizontal line, then at least some of the means differ.

Definition 5.11. An **outlier** corresponds to a case that is far from the bulk of the data. Look for a large vertical distance of the plotted point from the identity line or the $r = 0$ line.

Rule of thumb 5.1. Mentally add 2 lines parallel to the identity line and 2 lines parallel to the $r = 0$ line that cover most of the cases. Then a case is an outlier if it is well beyond these 2 lines.

This rule often fails for large outliers since often the identity line goes through or near a large outlier so its residual is near zero. A response that is far from the bulk of the data in the response plot is a "large outlier" (large in magnitude). Look for a large gap between the bulk of the data and the large outlier.

Suppose there is a dot plot of n_j cases corresponding to level a_j that is far from the bulk of the data. This dot plot is probably not a cluster of "bad outliers" if $n_j \geq 4$ and $n \leq 50$. If $n_j = 1$, such a case may be a large outlier.

Rule of thumb 5.2. Often an outlier is very good, but more often an outlier is due to a measurement error and is very bad.

The assumption of the Y_{ij} coming from the same location scale family with different location parameters μ_i and the same constant variance σ^2 is a big assumption and often does not hold. Another way to check this assumption is to make a box plot of the Y_{ij} for each i. The box in the box plot corresponds to the lower, middle and upper quartiles of the Y_{ij}. The middle quartile is just the sample median of the data m_{ij}: at least half of the $Y_{ij} \geq m_{ij}$ and at least half of the $Y_{ij} \leq m_{ij}$. The p boxes should be roughly the same length and the median should occur in roughly the same position (eg in the center of each box). The "whiskers" in each plot should also be roughly similar. Histograms for each of the p samples could also be made. All of the histograms should look similar in shape.

Example 5.4. Kuehl (1994, p. 128) gives data for counts of hermit crabs on 25 different transects in each of six different coastline habitats. Let Z be the count. Then the response variable $Y = \log_{10}(Z + 1/6)$. Although the counts Z varied greatly, each habitat had several counts of 0 and often there were several counts of 1, 2 or 3. Hence Y is not a continuous variable. The cell means model was fit with $n_i = 25$ for $i = 1, ..., 6$. Each of the six habitats was a level. Figure 5.1a and b shows the response plot and residual plot. There are 6 dot plots in each plot. Because several of the smallest values in each plot are identical, it does not always look like the identity line is passing through the six sample means $\overline{Y}_{i0}$ for $i = 1, ..., 6$. In particular, examine the dot plot for the smallest mean (look at the 25 dots furthest to the left that fall on the vertical line FIT ≈ 0.36). Random noise (jitter) has been added to the response and residuals in Figure 5.1c and d. Now it is easier to compare the six dot plots. They seem to have roughly the same spread.

The plots contain a great deal of information. The response plot can be used to explain the model, check that the sample from each population (treatment) has roughly the same shape and spread, and to see which populations have similar means. Since the response plot closely resembles the residual plot in Figure 5.1, there may not be much difference in the six populations. Linearity seems reasonable since the samples scatter about the identity line. The residual plot makes the comparison of "similar shape" and "spread" easier.

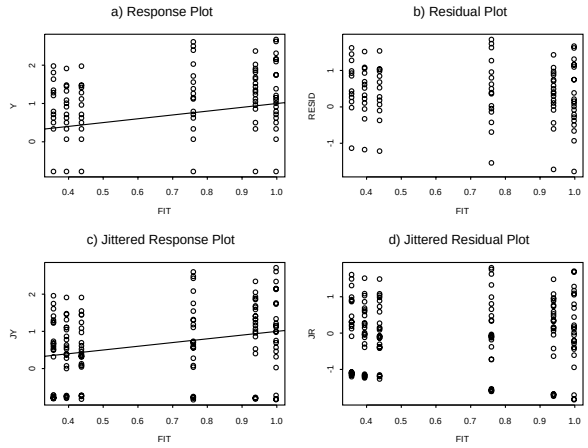

Figure 5.1: Plots for Crab Data

Definition 5.12. a) The *total sum of squares*

$$SSTO = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{00})^2.$$

b) The *treatment sum of squares*

$$SSTR = \sum_{i=1}^{p} n_i (\overline{Y}_{i0} - \overline{Y}_{00})^2.$$

c) The residual sum of squares or *error sum of squares*

$$SSE = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{io})^2.$$

Definition 5.13. Associated with each SS in Definition 5.12 is a degrees of freedom (df) and a mean square = SS/df. For SSTO, df = $n-1$ and $MSTO = SSTO/(n-1)$. For SSTR, df = $p-1$ and $MSTR = SSTR/(p-1)$. For SSE, $df = n-p$ and $MSE = SSE/(n-p)$.

Let $S_i^2 = \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i0})^2/(n_i - 1)$ be the sample variance of the ith group. Then the MSE is a weighted sum of the S_i^2:

$$\hat{\sigma}^2 = MSE = \frac{1}{n-p} \sum_{i=1}^{p} \sum_{j=1}^{n_i} r_{ij}^2 = \frac{1}{n-p} \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i0})^2 =$$

$$\frac{1}{n-p}\sum_{i=1}^{p}(n_i-1)S_i^2 = S_{pool}^2$$

where S_{pool}^2 is known as the pooled variance estimator.

The ANOVA table is the same as that for MLR, except that SSTR replaces the regression sum of squares. The MSE is again an estimator of σ^2. The ANOVA F test tests whether all p means μ_i are equal. Shown below is an ANOVA table given in symbols. Sometimes "Treatment" is replaced by "Between treatments," "Between Groups," "Model," "Factor" or "Groups." Sometimes "Error" is replaced by "Residual," or "Within Groups." Sometimes "p-value" is replaced by "P", "$Pr(> F)$" or "PR > F."

Summary Analysis of Variance Table

Source	df	SS	MS	F	p-value
Treatment	p-1	SSTR	MSTR	Fo=MSTR/MSE	for Ho:
Error	n-p	SSE	MSE		$\mu_1 = \cdots = \mu_p$

Be able to perform the 4 step fixed effects one way ANOVA F test of hypotheses:
i) State the hypotheses Ho: $\mu_1 = \mu_2 = \cdots = \mu_p$ and Ha: not Ho.
ii) Find the test statistic $F_o = MSTR/MSE$ or obtain it from output.
iii) Find the p–value from output or use the F–table: p–value =

$$P(F_{p-1,n-p} > F_o).$$

iv) State whether you reject Ho or fail to reject Ho. If the p–value $< \delta$, reject Ho and conclude that the mean response depends on the level of the factor. Otherwise fail to reject Ho and conclude that the mean response does not depend on the level of the factor. Give a nontechnical sentence.

Rule of thumb 5.3. If

$$\max(S_1, ..., S_p) \leq 2\min(S_1, ..., S_p),$$

then the one way ANOVA F test results will be approximately correct if the response and residual plots suggest that the remaining one way ANOVA model assumptions are reasonable. See Moore (1999, p. 512).

Remark 5.1. If the units are a representative sample of some population of interest, then randomization of units into groups makes the assumption

that $Y_{i1}, ..., Y_{i,n_i}$ are iid hold to a useful approximation. Random sampling from populations also induces the iid assumption. Linearity can be checked with the response plot, and similar shape and spread of the location families can be checked with both the response and residual plots. Also check that outliers are not present. If the p dot plots in the response plot are approximately symmetric, then the sample sizes n_i can be smaller than if the dot plots are skewed.

Remark 5.2. When the assumption that the p groups come from the same location family with finite variance σ^2 is violated, the one way ANOVA F test may not make much sense because unequal means may not imply the superiority of one category over another. Suppose Y is the time in minutes until relief from a headache and that $Y_{1j} \sim N(60, 1)$ while $Y_{2j} \sim N(65, \sigma^2)$. If $\sigma^2 = 1$, then the type 1 medicine gives headache relief 5 minutes faster, on average, and is superior, all other things being equal. But if $\sigma^2 = 100$, then many patients taking medicine 2 experience much faster pain relief than those taking medicine 1, and many experience much longer time until pain relief. In this situation, predictor variables that would identify which medicine is faster for a given patient would be very useful.

```
fat1  fat2  fat3  fat4 One way Anova for Fat1 Fat2 Fat3 Fat4
64    78    75    55   Source     DF SS     MS     F     P
72    91    93    66   treatment  3  1636.5 545.5  5.41  0.0069
68    97    78    49   error      20 2018.0 100.9
77    82    71    64
56    85    63    70
95    77    76    68
```

Example 5.5. The output above represents grams of fat (minus 100 grams) absorbed by doughnuts using 4 types of fat. See Snedecor and Cochran (1967, p. 259). Let μ_i denote the mean amount of fat$i$ absorbed by doughnuts, $i = 1, 2, 3$ and 4. a) Find $\hat{\mu}_1$. b) Perform a 4 step Anova F test.

Solution: a) $\hat{\boldsymbol{\beta}}_{1c} = \hat{\mu}_1 = \overline{Y}_{10} = Y_{10}/n_1 = \sum_{j=1}^{n_1} Y_{1j}/n_1 =$
$(64 + 72 + 68 + 77 + 56 + 95)/6 = 432/6 = 72.$

b) i) $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ H_a: not H_0

ii) F = 5.41

iii) pvalue = 0.0069

iv) Reject H_0, the mean amount of fat absorbed by doughnuts depends on the type of fat.

Definition 5.14. A **contrast** $C = \sum_{i=1}^{p} k_i \mu_i$ where $\sum_{i=1}^{p} k_i = 0$. The estimated contrast is $\hat{C} = \sum_{i=1}^{p} k_i \overline{Y}_{i0}$.

If the null hypothesis of the fixed effects one way ANOVA test is not true, then not all of the means μ_i are equal. Researchers will often have hypotheses, before examining the data, that they desire to test. Often such a hypothesis can be put in the form of a contrast. For example, the contrast $C = \mu_i - \mu_j$ is used to compare the means of the ith and jth groups while the contrast $\mu_1 - (\mu_2 + \cdots + \mu_p)/(p-1)$ is used to compare the last $p-1$ groups with the 1st group. This contrast is useful when the 1st group corresponds to a standard or control treatment while the remaining groups correspond to new treatments.

Assume that the normal cell means model is a useful approximation to the data. Then the $\overline{Y}_{i0} \sim N(\mu_i, \sigma^2/n_i)$ are independent, and

$$\hat{C} = \sum_{i=1}^{p} k_i \overline{Y}_{i0} \sim N\left(C, \sigma^2 \sum_{i=1}^{p} \frac{k_i^2}{n_i}\right).$$

Hence the standard error

$$SE(\hat{C}) = \sqrt{MSE \sum_{i=1}^{p} \frac{k_i^2}{n_i}}.$$

The degrees of freedom is equal to the MSE degrees of freedom $= n - p$.

Consider a family of null hypotheses for contrasts $\{Ho : \sum_{i=1}^{p} k_i \mu_i = 0$ where $\sum_{i=1}^{p} k_i = 0$ and the k_i may satisfy other constraints$\}$. Let δ_S denote the probability of a type I error for a single test from the family where a type I error is a false rejection. The **family level** δ_F is an upper bound on the (usually unknown) size δ_T. Know how to interpret $\delta_F \approx \delta_T =$
P(of making at least one type I error among the family of contrasts).

Two important families of contrasts are the family of all possible contrasts and the family of pairwise differences $C_{ij} = \mu_i - \mu_j$ where $i \neq j$. The Scheffé multiple comparisons procedure has a δ_F for the family of all possible contrasts while the Tukey multiple comparisons procedure has a δ_F for the family of all $\binom{p}{2}$ pairwise contrasts.

To interpret output for multiple comparisons procedures, the underlined means or blocks of letters besides groups of means indicate that the group of means are not significantly different.

Example 5.6. The output below uses data from SAS Institute (1985, p. 126-129). The mean nitrogen content of clover depends on the strain of clover (3dok1, 3dok5, 3dok7, compos, 3dok4, 3dok13). Recall that means μ_1 and μ_2 are significantly different if you can conclude that $\mu_1 \neq \mu_2$ while μ_1 and μ_2 are not significantly different if there is not enough evidence to conclude that $\mu_1 \neq \mu_2$ (perhaps because the means are approximately equal or perhaps because the sample sizes are not large enough).

Notice that the strain of clover 3dok1 appears to have the highest mean nitrogen content. There are 4 pairs of means that are not significantly different. The letter B suggests 3dok5 and 3dok7, the letter C suggests 3dok7 and compos, the letter D suggests compos and 3dok4, while the letter E suggests 3dok4 and 3dok13 are not significantly different.

```
Means with the same letter are not significantly different.
     Waller Grouping          Mean      N    strain
                    A        28.820      5    3dok1

                    B        23.980      5    3dok5
                    B
                C   B        19.920      5    3dok7
                C
                C   D        18.700      5    compos
                    D
                E   D        14.640      5    3dok4
                E
                E            13.260      5    3dok13
```

Definition 5.15. Graphical Anova for the one way model uses the residuals as a reference set instead of a t, F or normal distribution. The scaled treatment deviations or scaled effect $c(\overline{Y}_{i0} - \overline{Y}_{00}) = c(\hat{\mu}_i - \overline{Y}_{00})$ are scaled to have the same variability as the residuals. A dot plot of the scaled deviations is placed above the dot plot of the residuals. Assume that $n_i \equiv m = n/p$ for $i = 1, ..., p$. For small $n \leq 40$, suppose the distance between two scaled deviations (A and B, say) is greater than the range of the residuals $= \max(r_{ij}) - \min(r_{ij})$. Then declare μ_A and μ_B to be significantly

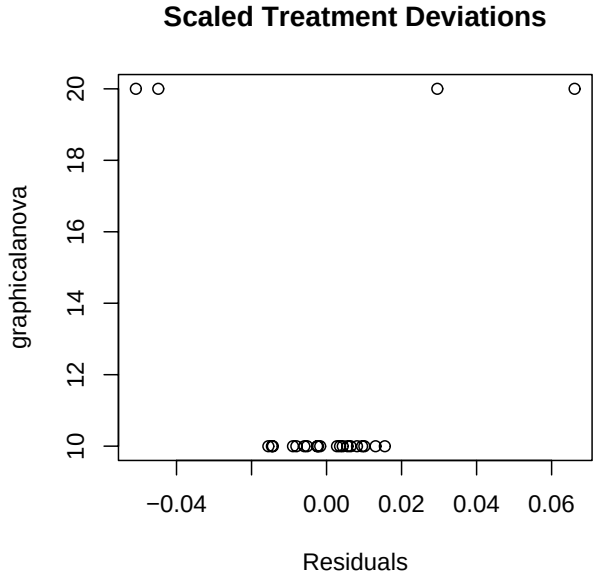

Figure 5.2: Graphical Anova

different. If the distance is less than the range, do not declare μ_A and μ_B to be significantly different. Scaled deviations that lie outside the range of the residuals are significant (so significantly different from the overall mean).

For $n \geq 100$, let $r_{(1)} \leq r_{(2)} \leq \cdots \leq r_{(n)}$ be the order statistics of the residuals. Then instead of the range, use $r_{(\lceil 0.975n \rceil)} - r_{(\lceil 0.025n \rceil)}$ as the distance where $\lceil x \rceil$ is the smallest integer $\geq x$, eg $\lceil 7.7 \rceil = 8$. So effects outside of the interval $(r_{(\lceil 0.025n \rceil)}, r_{(\lceil 0.975n \rceil)})$ are significant. See Box, Hunter and Hunter (2005, p. 136, 166). A derivation of the scaling constant $c = \sqrt{(n-p)/(p-1)}$ is given in Section 5.6.

```
ganova(x,y)
sdev    0.02955502  0.06611268 -0.05080048 -0.04486722
Treatments "A"      "B"        "C"          "D"
```

Example 5.7. Cobb (1998, p. 160) describes a one way Anova design used to study the amount of calcium in the blood. For many animals, the body's ability to use calcium depends on the level of certain hormones in the blood. The response was 1/(level of plasma calcium). The four groups

were A: Female controls, B: Male controls, C: Females given hormone and D: Males given hormone. There were 10 birds of each gender, and five from each gender were given the hormone. The output above uses the `regpack` function *ganova* to produce Figure 5.2.

In Figure 5.2, the top dot plot has the scaled treatment deviations. From left to right, these correspond to C, D, A and B since the output shows that the deviation corresponding to C is the smallest with value -0.050. Since the deviations corresponding to C and D are much closer than the range of the residuals, the C and D effects yielded similar mean response values. A and B appear to be significantly different from C and D. The distance between the scaled A and B treatment deviations is about the same as the distance between the smallest and largest residuals, so there is only marginal evidence that the A and B effects are significantly different.

Since all 4 scaled deviations lie outside of the range of the residuals, all effects A, B, C and D appear to be significant.

5.3 Random Effects One Way ANOVA

Definition 5.16. For the **random effects one way Anova**, the levels of the factor are a random sample of levels from some population of levels Λ_F. The cell means model for the random effects one way Anova is $Y_{ij} = \mu_i + e_{ij}$ for $i = 1, ..., p$ and $j = 1, ..., n_i$. The μ_i are randomly selected from some population Λ with mean μ and variance σ_μ^2, where $i \in \Lambda_F$ is equivalent to $\mu_i \in \Lambda$. The e_{ij} and μ_i are independent, and the e_{ij} are iid from a location family with pdf f, mean 0 and variance σ^2. The $Y_{ij}|\mu_i \sim f(y - \mu_i)$, the location family with location parameter μ_i and variance σ^2. Unconditionally, $E(Y_{ij}) = \mu$ and $V(Y_{ij}) = \sigma_\mu^2 + \sigma^2$.

For the random effects model, the μ_i are independent random variables with $E(\mu_i) = \mu$ and $V(\mu_i) = \sigma_\mu^2$. The cell means model for fixed effects one way Anova is very similar to that for the random effects model, but the μ_i are fixed constants rather than random variables.

Definition 5.17. For the *normal random effects one way Anova* model, $\Lambda \sim N(\mu, \sigma_\mu^2)$. Thus the μ_i are independent $N(\mu, \sigma_\mu^2)$ random variables. The e_{ij} are iid $N(0, \sigma^2)$ and the e_{ij} and μ_i are independent. For this model, $Y_{ij}|\mu_i \sim N(\mu_i, \sigma^2)$ for $i = 1, ..., p$. Note that the conditional variance σ^2 is the same for each $\mu_i \in \Lambda$. Unconditionally, $Y_{ij} \sim N(\mu, \sigma_\mu^2 + \sigma^2)$.

The fixed effects one way Anova tested $Ho : \mu_1 = \cdots = \mu_p$. For the random effects one way Anova, interest is in whether $\mu_i \equiv \mu$ for every μ_i in Λ where the population Λ is not necessarily finite. Note that if $\sigma_\mu^2 = 0$, then $\mu_i \equiv \mu$ for all $\mu_i \in \Lambda$. In the sample of p levels, the μ_i will differ if $\sigma_\mu^2 > 0$.

Be able to perform the 4 step random effects one way ANOVA F test of hypotheses:
i) $Ho : \sigma_\mu^2 = 0 \;\; Ha : \;\; \sigma_\mu^2 > 0$
ii) $F_o = MSTR/MSE$ is usually obtained from output.
iii) The p-value $= P(F_{p-1,n-p} > Fo)$ is usually obtained from output.
iv) If p–value $< \delta$ reject Ho, conclude that $\sigma_\mu^2 > 0$ and that the mean response depends on the level of the factor. Otherwise, fail to reject Ho, conclude that $\sigma_\mu^2 = 0$ and that the mean response does not depend on the level of the factor.

The ANOVA tables for the fixed and random effects one way Anova models are exactly the same, and the two F tests are very similar. The main difference is that the conclusions for the random effects model can be generalized to the entire population of levels. For the fixed effects model, the conclusions only hold for the p fixed levels. If $Ho : \sigma_\mu^2 = 0$ is true and the random effect model holds, then the Y_{ij} are iid with pdf $f(y - \mu)$. So the F statistic for the random effects test has an approximate $F_{p-1,n-p}$ distribution if the n_i are large by the results for the fixed effects one way Anova test. For both tests, the output p-value is an estimate of the population p-value.

Source	df	SS	MS	F	P
brand	5	854.53	170.906	238.71	0.0000
error	42	30.07	0.716		

Example 5.8. Data is from Kutner, Nachtsheim, Neter and Li (2005, problem 25.7). A researcher is interested in the amount of sodium in beer. She selects 6 brands of beer at random from 127 brands and the response is the average sodium content measured from 8 cans of each brand.

a) State whether this is a random or fixed effects one way Anova. Explain briefly.

b) Using the output above, perform the appropriate 4 step Anova F test.

Solution: a) Random effects since the beer brands were selected at random from a population of brands.

b) i) $H_0 : \sigma_\mu^2 = 0$ $H_a : \sigma_\mu^2 > 0$

ii) $F_0 = 238.71$

iii) pvalue = 0.0

iv) Reject H_0, so $\sigma_\mu^2 > 0$ and the mean amount of sodium depends on the beer brand.

Remark 5.3. The response and residual plots for the random effects models are interpreted in the same way as for the fixed effects model, except that the dot plots are from a random sample of p levels instead of from p fixed levels.

5.4 Response Transformations for Experimental Design

A model for an experimental design is $Y_i = E(Y_i) + e_i$ for $i = 1, ..., n$ where the error $e_i = Y_i - E(Y_i)$ and $E(Y_i) \equiv E(Y_i|\boldsymbol{x}_i)$ is the expected value of the response Y_i for a given vector of predictors $\boldsymbol{x}_i$. Many models can be fit with least squares (OLS or LS) and are linear models of the form

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$$

for $i = 1, \ldots, n$. Often $x_{i,1} \equiv 1$ for all i. In matrix notation, these n equations become

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e},$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ design matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors. If the fitted values are $\hat{Y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$, then $Y_i = \hat{Y}_i + r_i$ where the residuals $r_i = Y_i - \hat{Y}_i$.

The applicability of an experimental design model can be expanded by allowing response transformations. An important class of *response transformation models* adds an additional unknown transformation parameter λ_o, such that

$$Y_i = t_{\lambda_o}(Z_i) \equiv Z_i^{(\lambda_o)} = E(Y_i) + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i.$$

If λ_o was known, then $Y_i = t_{\lambda_o}(Z_i)$ would follow the linear model for the experimental design.

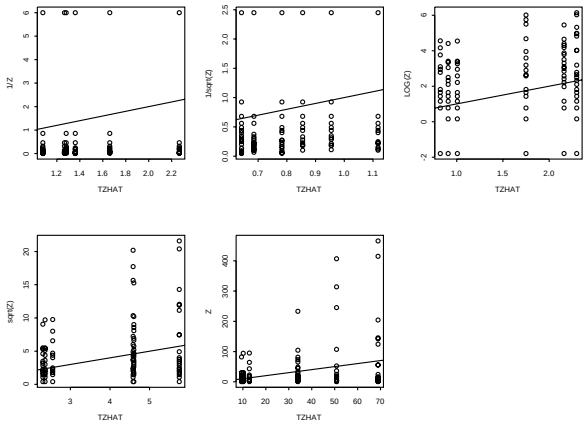

Figure 5.3: Transformation Plots for Crab Data

Definition 5.18. Assume that **all** of the values of the "response" Z_i are **positive**. A *power transformation* has the form $Y = t_\lambda(Z) = Z^\lambda$ for $\lambda \neq 0$ and $Y = t_0(Z) = \log(Z)$ for $\lambda = 0$ where $\lambda \in \Lambda_L = \{-1, -1/2, 0, 1/2, 1\}$.

A graphical method for response transformations computes the fitted values $\hat{W}_i$ from the experimental design model using $W_i = t_\lambda(Z_i)$ as the "response." Then a plot of the $\hat{W}$ versus W is made for each of the five values of $\lambda \in \Lambda_L$. The plotted points follow the identity line in a (roughly) evenly populated band if the experimental design model is reasonable for $(\hat{W}, W)$. If more than one value of $\lambda \in \Lambda_L$ gives a linear plot, consult subject matter experts and use the simplest or most reasonable transformation. Note that Λ_L has 5 models, and the graphical method selects the model with the best response plot. After selecting the transformation, the usual checks should be made. In particular, the transformation plot is also the response plot, and a residual plot should be made.

Definition 5.19. A *transformation plot* is a plot of $(\hat{W}, W)$ with the identity line added as a visual aid.

In the following example, the plots show $t_\lambda(Z)$ on the vertical axis. The label "TZHAT" of the horizontal axis are the fitted values that result from using $t_\lambda(Z)$ as the "response" in the software.

For one way Anova models with $n_i \equiv m \geq 5$, look for a transformation

plot that satisfies the following conditions. i) The p dot plots scatter about the identity line with similar shape and spread. ii) Dot plots with more skew are worse than dot plots with less skew or dot plots that are approximately symmetric. iii) Spread that increases or decreases with TZHAT is bad.

Example 5.4, continued. Following Kuehl (1994, p. 128), let C be the count of crabs and let the "response" $Z = C + 1/6$. Figure 5.3 shows the five *transformation plots*. The transformation $\log(Z)$ results in dot plots that have roughly the same shape and spread. The transformations $1/Z$ and $1/\sqrt{Z}$ do not handle the 0 counts well, and the dot plots fail to cover the identity line. The transformations $\sqrt{Z}$ and Z have variance that increases with the mean.

Remark 5.4. The graphical method for response transformations can be used for design models that are linear models, not just one way Anova models. The method is nearly identical to that of Chapter 3, but Λ_L only has 5 values. The **log rule** states that if all of the $Z_i > 0$ and if $\frac{\max(Z_i)}{\min(Z_i)} \geq 10$, then the response transformation $Y = \log(Z)$ will often work.

5.5 Summary

1) The **fixed effects one way Anova** model has one qualitative explanatory variable called a **factor** and a quantitative response variable Y_{ij}. The factor variable has p levels, $E(Y_{ij}) = \mu_i$ and $V(Y_{ij}) = \sigma^2$ for $i = 1, ..., p$ and $j = 1, ..., n_i$. **Experimental units** are randomly assigned to the treatment levels.

2) Let $n = n_1 + \cdots + n_p$. In an **experiment**, the investigators use randomization to randomly assign n units to treatments. Draw a random permutation of $\{1, ..., n\}$. Assign the first n_1 units to treatment 1, the next n_2 units to treatment 2, ..., and the final n_p units to treatment p. Use $n_i \equiv m = n/p$ if possible. Randomization washes out the effect of lurking variables.

3) The 4 step fixed effects one way Anova F test has steps
i) Ho: $\mu_1 = \mu_2 = \cdots = \mu_p$ and Ha: not Ho.
ii) Fo = MSTR/MSE is usually given by output.
iii) The p-value = $P(F_{p-1,n-p} > Fo)$ is usually given by output.
iv) If the p–value $< \delta$, reject Ho and conclude that the mean response depends on the level of the factor. Otherwise fail to reject Ho and conclude that the

mean response does not depend on the level of the factor. Give a nontechnical sentence.

Summary Analysis of Variance Table

Source	df	SS	MS	F	p-value
Treatment	p-1	SSTR	MSTR	Fo=MSTR/MSE	for Ho:
Error	n-p	SSE	MSE		$\mu_1 = \cdots = \mu_p$

4) Shown is an ANOVA table given in symbols. Sometimes "Treatment" is replaced by "Between treatments," "Between Groups," "Model," "Factor" or "Groups." Sometimes "Error" is replaced by "Residual," or "Within Groups." Sometimes "p-value" is replaced by "P", "$Pr(> F)$" or "PR > F."

5) Boxplots and dot plots for each level are useful for this test. A *dot plot* of $Z_1, ..., Z_m$ consists of an axis and m points each corresponding to the value of Z_i. If all of the boxplots or dot plots are about the same, then probably the Anova F test will fail to reject Ho. If Ho is true, then $Y_{ij} = \mu + e_{ij}$ where the e_{ij} are iid with 0 mean and constant variance σ^2. Then $\hat{\mu} = \overline{Y}_{00}$ and the factor doesn't help predict Y_{ij}.

6) Let $f_Z(z)$ be the pdf of Z. Then the family of pdfs $f_Y(y) = f_Z(y - \mu)$ indexed by the *location parameter* μ, $-\infty < \mu < \infty$, is the *location family* for the random variable $Y = \mu + Z$ with *standard pdf* $f_Z(y)$. A one way fixed effects ANOVA model has a single qualitative predictor variable W with p categories a_1, ..., a_p. There are p different distributions for Y, one for each category a_i. The distribution of

$$Y|(W = a_i) \sim f_Z(y - \mu_i)$$

where the location family has second moments. Hence all p distributions come from the same location family with different location parameter μ_i and the same variance σ^2. The one way fixed effects normal ANOVA model is the special case where $Y|(W = a_i) \sim N(\mu_i, \sigma^2)$.

7) The *response plot* is a plot of $\hat{Y}$ versus Y. For the one way Anova model, the response plot is a plot of $\hat{Y}_{ij} = \hat{\mu}_i$ versus Y_{ij}. Often the identity line with unit slope and zero intercept is added as a visual aid. Vertical deviations from the identity line are the residuals $r_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \hat{\mu}_i$. The plot will consist of p dot plots that scatter about the identity line with similar shape and spread if the fixed effects one way ANOVA model is appropriate. The ith dot plot is a dot plot of $Y_{i,1}, ..., Y_{i,n_i}$. Assume that each $n_i \geq 10$. If

the response plot looks like the residual plot, then a horizontal line fits the p dot plots about as well as the identity line, and there is not much difference in the μ_i. If the identity line is clearly superior to any horizontal line, then at least some of the means differ.

8) The *residual plot* is a plot of $\hat{Y}$ versus residual $r = Y - \hat{Y}$. The plot will consist of p dot plots that scatter about the $r = 0$ line with similar shape and spread if the fixed effects one way ANOVA model is appropriate. The ith dot plot is a dot plot of $r_{i,1}, ..., r_{i,n_i}$. Assume that each $n_i \geq 10$. Under the assumption that the Y_{ij} are from the same location scale family with different parameters μ_i, each of the p dot plots should have roughly the same shape and spread. This assumption is easier to judge with the residual plot than with the response plot.

9) Rule of thumb: If $\max(S_1, ..., S_p) \leq 2 \min(S_1, ..., S_p)$, then the one way ANOVA F test results will be approximately correct if the response and residual plots suggest that the remaining one way ANOVA model assumptions are reasonable.

10) In an **experiment**, the investigators assign units to treatments. In an **observational study**, investigators simply observe the response, and the treatment groups need to be p random samples from p populations (the levels). The effects of lurking variables are present in observational studies.

11) If a qualitative variable has c levels, represent it with $c - 1$ or c indicator variables. Given a qualitative variable, know how to represent the data with indicator variables.

12) The **cell means model** for the fixed effects one way Anova is $Y_{ij} = \mu_i + e_{ij}$ where Y_{ij} is the value of the response variable for the jth trial of the ith factor level for $i = 1, ..., p$ and $j = 1, ..., n_i$. The μ_i are the unknown means and $E(Y_{ij}) = \mu_i$. The e_{ij} are iid from the location family with pdf $f_Z(z)$, zero mean and unknown variance $\sigma^2 = V(Y_{ij}) = V(e_{ij})$. For the normal cell means model, the e_{ij} are iid $N(0, \sigma^2)$. The estimator $\hat{\mu}_i = \overline{Y}_{i0} = \sum_{j=1}^{n_i} Y_{ij}/n_i = \hat{Y}_{ij}$. The ith residual is $r_{ij} = Y_{ij} - \overline{Y}_{i0}$, and $\overline{Y}_{00}$ is the sample mean of all of the Y_{ij} and $n = \sum_{i=1}^{p} n_i$. The total sum of squares SSTO $= \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{00})^2$, the treatment sum of squares SSTR $= \sum_{i=1}^{p} n_i (\overline{Y}_{i0} - \overline{Y}_{00})^2$, and the error sum of squares SSE $= \sum_{i=1}^{p} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i0})^2$. The MSE is an estimator of σ^2. The Anova table is the same as that for multiple linear regression, except that SSTR replaces the regression sum of squares and that SSTO, SSTR and SSE have $n - 1$, $p - 1$ and $n - p$ degrees of freedom.

13) Let $Y_{i0} = \sum_{j=1}^{n_i} Y_{ij}$ and let

$$\hat{\mu}_i = \overline{Y}_{i0} = Y_{i0}/n_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}.$$

Hence the "dot notation" means sum over the subscript corresponding to the 0, eg j. Similarly, $Y_{00} = \sum_{i=1}^{p}\sum_{j=1}^{n_i} Y_{ij}$ is the sum of all of the Y_{ij}. Be able to find $\hat{\mu}_i$ from data.

14) If the p treatment groups have the same pdf (so $\mu_i \equiv \mu$ in the location family) with finite variance σ^2, and if the one way ANOVA F test statistic is computed from all $\frac{n!}{n_1! \cdots n_p!}$ ways of assigning n_i of the response variables to treatment i, then the histogram of the F test statistic is approximately $F_{p-1,n-p}$ for large n_i.

15) For the one way Anova, the fitted values $\hat{Y}_{ij} = \overline{Y}_{i0}$ and the residuals $r_{ij} = Y_{ij} - \hat{Y}_{ij}$.

16) **Know** that for the **random effects one way Anova**, the levels of the factor are a random sample of levels from some population of levels Λ_F. Assume the μ_i are iid with mean μ and variance σ^2_μ. The cell means model for the random effects one way Anova is $Y_{ij} = \mu_i + e_{ij}$ for $i = 1, ..., p$ and $j = 1, ..., n_i$. The sample size $n = n_1 + \cdots + n_p$ and often $n_i \equiv m$ so $n = pm$. The μ_i and e_{ij} are independent. The e_{ij} have mean 0 and variance σ^2. The $Y_{ij}|\mu_i \sim f(y - \mu_i)$, a location family with variance σ^2 while $e_{ij} \sim f(y)$. In the test below, if $H_0 : \sigma^2_\mu = 0$ is true, then the Y_{ij} are iid with pdf $f(y - \mu)$, so the F statistic $\approx F_{p-1,n-p}$ if the n_i are large.

17) **Know** that the 4 step random effects one way Anova test is
i) H_0 $\sigma^2_\mu = 0$ $\;$ H_A $\sigma^2_\mu > 0$
ii) $F_0 = MSTR/MSE$ is usually obtained from output.
iii) The pvalue $= P(F_{p-1,n-p} > F_0)$ is usually obtained from output.
iv) If pvalue $< \delta$ reject Ho, conclude that $\sigma^2_\mu > 0$ and that the mean response depends on the level of the factor. Otherwise, fail to reject Ho, conclude that $\sigma^2_\mu = 0$ and that the mean response does not depend on the level of the factor.

18) Know how to tell whether the experiment is a fixed or random effects one way Anova. (Were the levels fixed or a random sample from a population of levels?)

19) The applicability of a DOE (design of experiments) model can be expanded by allowing response transformations. An important class of *response*

transformation models is

$$Y = t_{\lambda_o}(Z) = E(Y) + e = \boldsymbol{x}^T\boldsymbol{\beta} + e$$

where the subscripts (eg Y_{ij}) have been suppressed. If λ_o was known, then $Y = t_{\lambda_o}(Z)$ would follow the DOE model. Assume that **all** of the values of the "response" Z are **positive**. A **power transformation** has the form $Y = t_\lambda(Z) = Z^\lambda$ for $\lambda \neq 0$ and $Y = t_0(Z) = \log(Z)$ for $\lambda = 0$ where $\lambda \in \Lambda_L = \{-1, -1/2, 0, 1/2, 1\}$.

20) A graphical method for response transformations computes the fitted values $\hat{W}$ from the DOE model using $W = t_\lambda(Z)$ as the "response" for each of the five values of $\lambda \in \Lambda_L$. Let $\hat{T} = \hat{W}$ = TZHAT and plot TZHAT vs $t_\lambda(Z)$ for $\lambda \in \{-1, -1/2, 0, 1/2, 1\}$. These plots are called **transformation plots**. The residual or error degrees of freedom used to compute the MSE should not be too small. Choose the transformation $Y = t_{\lambda^*}(Z)$ that has the best plot. Consider the one way Anova model with $n_i > 4$ for $i = 1, ..., p$. i) The dot plots should spread about the identity line with similar shape and spread. ii) Dot plots that are approximately symmetric are better than skewed dot plots. iii) Spread that increases or decreases with TZHAT (the shape of the plotted points is similar to a right or left opening megaphone) is bad.

21) The transformation plot for the selected transformation is also the response plot for that model (eg for the model that uses $Y = \log(Z)$ as the response). Make all of the usual checks on the DOE model (residual and response plots) after selecting the response transformation.

22) The **log rule** says try $Y = \log(Z)$ if $\max(Z)/\min(Z) > 10$ where $Z > 0$ and the subscripts have been suppressed (so $Z \equiv Z_{ij}$ for the one way Anova model).

23) A contrast $C = \sum_{i=1}^p k_i\mu_i$ where $\sum_{i=1}^p k_i = 0$. The estimated contrast is $\hat{C} = \sum_{i=1}^p k_i\overline{Y}_{i0}$.

24) Consider a family of null hypotheses for contrasts $\{Ho : \sum_{i=1}^p k_i\mu_i = 0$ where $\sum_{i=1}^p k_i = 0$ and the k_i may satisfy other constraints $\}$. Let δ_S denote the probability of a type I error for a single test from the family. The **family level** δ_F is an upper bound on the (usually unknown) size δ_T. Know how to interpret $\delta_F \approx \delta_T$ = P(of making at least one type I error among the family of contrasts) where a type I error is a false rejection.

25) Two important families of contrasts are the family of all possible contrasts and the family of pairwise differences $C_{ij} = \mu_i - \mu_j$ where $i \neq j$.

The Scheffé multiple comparisons procedure has a δ_F for the family of all possible contrasts while the Tukey multiple comparisons procedure has a δ_F for the family of all $\binom{p}{2}$ pairwise contrasts.

26) **Know** how to interpret output for multiple comparisons procedures. Underlined means or blocks of letters besides groups of means indicates that the group of means are not significantly different.

27) **Graphical Anova** for the **one way Anova** model makes a dot plot of scaled treatment deviations (effects) above a dot plot of the residuals. For small $n \leq 40$, suppose the distance between two scaled deviations (A and B, say) is greater than the range of the residuals $= \max(r_{ij}) - \min(r_{ij})$. Then declare μ_A and μ_B to be significantly different. If the distance is less than the range, do not declare μ_A and μ_B to be significantly different. Assume the $n_i \equiv m$ for $i = 1, ..., p$. Then the ith scaled deviation is $c(\overline{Y}_{i0} - \overline{Y}_{00}) = c\hat{\alpha}_i = \tilde{\alpha}_i$ where $c = \sqrt{df_e/df_{treat}} = \sqrt{\dfrac{n-p}{p-1}}$.

28) The analysis of the response, not that of the residuals, is of primary importance. The response plot can be used to analyze the response in the background of the fitted model. For linear models such as experimental designs, the estimated mean function is the identity line and should be added as a visual aid.

29) Assume that the residual degrees of freedom are large enough for testing. Then the response and residual plots contain much information. Linearity and constant variance may be reasonable if the p dot plots have roughly the same shape and spread, and the dot plots scatter about the identity line. The p dot plots of the residuals should have similar shape and spread, and the dot plots scatter about the $r = 0$ line. It is easier to check linearity with the response plot and constant variance with the residual plot. Curvature is often easier to see in a residual plot, but the response plot can be used to check whether the curvature is monotone or not. The response plot is more effective for determining whether the signal to noise ratio is strong or weak, and for detecting outliers or influential cases.

5.6 Complements

Often the data does not consist of samples from p populations, but consists of a group of $n = mp$ units where m units are randomly assigned to each of the p treatments. Then the ANOVA models can still be used to compare

treatments, but statistical inference to a larger population can not be made. Of course a nonstatistical generalization to larger populations can be made. The nonstatistical generalization from the group of units to a larger population is most compelling if several experiments are done with similar results. For example, generalizing the results of an experiment for psychology students to the population of all of the university students is less compelling than the following generalization. Suppose one experiment is done for psychology students, one for engineers and one for English majors. If all three experiments give similar results, then generalize the results to the population of all of the university's students.

Four good tests on the design and analysis of experiments are Box, Hunter and Hunter (2005), Cobb (1998), Kuehl (1994) and Ledolter and Swersey (2007). Also see Dean and Voss (2000), Kirk (1982), Montgomery (2005) and Oehlert (2000).

A **randomization test** has H_0: *the different treatments have no effect.* This null hypothesis is also true if all p pdfs $Y|(W = a_i) \sim f_Z(y - \mu)$ are the same. An impractical randomization test uses all $M = \frac{n!}{n_1! \cdots n_p!}$ ways of assigning n_i of the Y_{ij} to treatment i for $i = 1, ..., p$. Let F_0 be the usual F statistic. The F statistic is computed for each of the M permutations and H_0 is rejected if the proportion of the M F statistics that are larger than F_0 is less than δ. The distribution of the M F statistics is approximately $F_{p-1,n-p}$ for large n when H_0 is true. The power of the randomization test is also similar to that of the usual F test. See Hoeffding (1952). These results suggest that the usual F test is semiparametric: the pvalue is approximately correct if n is large and if all p pdfs $Y|(W = a_i) \sim f_Z(y - \mu)$ are the same.

Let $[x]$ be the integer part of x, eg $[7.7] = 7$. Olive (2009c) shows that practical randomization tests that use a random sample of $\max(1000, [n \log(n)])$ permutations have level and power similar to the tests that use all M possible permutations. See Ernst (2009) and the *regpack* function `rand1way` for `R` code.

All of the parameterizations of the one way fixed effects ANOVA model yield the same predicted values, residuals and ANOVA F test, but the interpretations of the parameters differ. The cell means model is a linear model (without intercept) of the form $\boldsymbol{Y} = \boldsymbol{X}_c \boldsymbol{\beta}_c + \boldsymbol{e} =$ that can be fit using OLS. The OLS MLR output gives the correct fitted values and residuals but an incorrect Anova table. An equivalent linear model (with intercept) with correct OLS MLR Anova table as well as residuals and fitted values can be

formed by replacing any column of the cell means model by a column of ones **1**. Removing the last column of the cell means model and making the first column **1** gives the model $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} + e$ given in matrix form by (5.5).

It can be shown that the OLS estimators corresponding to (5.5) are $\hat{\beta}_0 = \overline{Y}_{p0} = \hat{\mu}_p$, and $\hat{\beta}_i = \overline{Y}_{i0} - \overline{Y}_{p0} = \hat{\mu}_i - \hat{\mu}_p$ for $i = 1, ..., p-1$. The cell means model has $\hat{\beta}_i = \hat{\mu}_i = \overline{Y}_{i0}$.

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1,n_1} \\ Y_{21} \\ \vdots \\ Y_{2,n_2} \\ \vdots \\ Y_{p,1} \\ \vdots \\ Y_{p,n_p} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 1 \\ 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} e_{11} \\ \vdots \\ e_{1,n_1} \\ e_{21} \\ \vdots \\ e_{2,n_2} \\ \vdots \\ e_{p,1} \\ \vdots \\ e_{p,n_p} \end{bmatrix}. \qquad (5.5)$$

Wilcox (2005) gives an excellent discussion of the problems that outliers and skewness can cause for the one and two sample t–intervals, the t–test, tests for comparing 2 groups and the ANOVA F test. Wilcox (2005) replaces ordinary population means by truncated population means and uses trimmed means to create analogs of one way ANOVA and multiple comparisons.

Graphical Anova uses scaled treatment effects = scaled treatment deviations $\tilde{d}_i = cd_i = c(\overline{Y}_{i0} - \overline{Y}_{00})$ for $i = 1, ..., p$. Following Box, Hunter and Hunter (2005, p. 166), suppose $n_i \equiv m = n/p$ for $i = 1, ..., n$. If Ho $\mu_1 = \cdots = \mu_p$ is true, want the sample variance of the scaled deviations to be approximately equal to the sample variance of the residuals. So want

$$1 \approx \frac{\frac{1}{p}\sum_{i=1}^{p} c^2 d_i^2}{\frac{1}{n}\sum_{i=1}^{n} r_i^2} = F_0 = \frac{MSTR}{MSE} = \frac{SSTR/(p-1)}{SSE/(n-p)} = \frac{\sum_{i=1}^{p} m d_i^2/(p-1)}{\sum_{i=1}^{n} r_i^2/(n-p)}$$

since $SSTR = \sum_{i=1}^{p} m(\overline{Y}_{i0} - \overline{Y}_{00})^2 = \sum_{i=1}^{p} md_i^2$. So

$$F_0 = \frac{\sum_{i=1}^{p} c^2 \frac{n}{p} d_i^2}{\sum_{i=1}^{n} r_i^2} = \frac{\sum_{i=1}^{p} \frac{m(n-p)}{p-1} d_i^2}{\sum_{i=1}^{n} r_i^2}.$$

Equating numerators gives

$$c^2 = \frac{mp}{n} \quad \frac{(n-p)}{(p-1)} = \frac{(n-p)}{(p-1)}$$

since $mp/n = 1$. Thus $c = \sqrt{(n-p)/(p-1)}$.

For Graphical Anova, see Box, Hunter and Hunter (2005, p. 136, 150, 164, 166) and Hoaglin, Mosteller, and Tukey (1991). The R package `granova`, available from (http://streaming.stat.iastate.edu/CRAN/) and authored by R.M. Pruzek and J.E. Helmreich, may be useful.

The *modified power transformation family*

$$Y_i = t_\lambda(Z_i) \equiv Z_i^{(\lambda)} = \frac{Z_i^\lambda - 1}{\lambda}$$

for $\lambda \neq 0$ and $t_0(Z_i) = \log(Z_i)$ for $\lambda = 0$ where $\lambda \in \Lambda_L$.

Box and Cox (1964) give a numerical method for selecting the response transformation for the modified power transformations. Although the method gives a point estimator $\hat{\lambda}_o$, often an interval of "reasonable values" is generated (either graphically or using a profile likelihood to make a confidence interval), and $\hat{\lambda} \in \Lambda_L$ is used if it is also in the interval.

There are several reasons to use a coarse grid Λ_L of powers. First, several of the powers correspond to simple transformations such as the log, square root, and reciprocal. These powers are easier to interpret than $\lambda = .28$, for example. Secondly, if the estimator $\hat{\lambda}_n$ can only take values in Λ_L, then sometimes $\hat{\lambda}_n$ will converge in probability to $\lambda^* \in \Lambda_L$. Thirdly, Tukey (1957) showed that neighboring modified power transformations are often very similar, so restricting the possible powers to a coarse grid is reasonable.

The graphical method for response transformations is due to Olive (2004) and Olive and Hawkins (2009a). A variant of the method would plot the residual plot or both the response and the residual plot for each of the five values of λ. Residual plots are also useful, but they do not distinguish between nonlinear monotone relationships and nonmonotone relationships. See Fox (1991, p. 55). Alternative methods are given by Cook and Olive (2002) and Box, Hunter and Hunter (2005, p. 321).

An alternative to one way ANOVA is to use FWLS (see Chapter 4) on the cell means model with $\sigma^2\boldsymbol{V} = diag(\sigma_1^2, ..., \sigma_p^2)$ where σ_i^2 is the variance of the ith group for $i = 1, ..., p$. Then $\hat{\boldsymbol{V}} = diag(S_1^2, ..., S_p^2)$ where $S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i0})^2$ is the sample variance of the Y_{ij}. Hence the estimated weights for FWLS are $\hat{w}_{ij} \equiv \hat{w}_i = 1/S_i^2$. Then the FWLS cell means model has $Y = \boldsymbol{X}_c\boldsymbol{\beta}_c + \boldsymbol{e}$ as in (5.1) except $\mathrm{Cov}(\boldsymbol{e}) = diag(\sigma_1^2, ..., \sigma_p^2)$.

Hence $\boldsymbol{Z} = \boldsymbol{U}_c\boldsymbol{\beta}_c + \boldsymbol{\epsilon}$. Then $\boldsymbol{U}_c^T\boldsymbol{U}_c = diag(n_1\hat{w}_1, ..., n_p\hat{w}_p)$, $(\boldsymbol{U}_c^T\boldsymbol{U}_c)^{-1} = diag(S_1^2/n_1, ..., S_p^2/n_p) = (\boldsymbol{X}\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}^T)^{-1}$, and $\boldsymbol{U}_c^T\boldsymbol{Z} = (\hat{w}_1 Y_{10}, ..., \hat{w}_p Y_{p0})^T$. Thus

$$\hat{\boldsymbol{\beta}}_{FWLS} = (\overline{Y}_{10}, ..., \overline{Y}_{p0})^T = \hat{\boldsymbol{\beta}}_c.$$

That is, the FWLS estimator equals the one way ANOVA estimator of $\boldsymbol{\beta}$ based on OLS applied to the cell means model. The ANOVA F test generalizes the pooled t test in that the two tests are equivalent for $p = 2$. The FWLS procedure is also known as the Welch one way ANOVA and generalizes the Welch t test. The Welch t test is thought to be much better than the pooled t test. See Brown and Forsythe (1974ab), Kirk (1982, p. 100, 101, 121, 122),Welch (1947, 1951) and Problem 5.11.

In matrix form $\boldsymbol{Z} = \boldsymbol{U}_c\boldsymbol{\beta}_c + \boldsymbol{\epsilon}$ becomes

$$\begin{bmatrix} \sqrt{\hat{w}_1}Y_{1,1} \\ \vdots \\ \sqrt{\hat{w}_1}Y_{1,n_1} \\ \sqrt{\hat{w}_2}Y_{21} \\ \vdots \\ \sqrt{\hat{w}_2}Y_{2,n_2} \\ \vdots \\ \sqrt{\hat{w}_p}Y_{p,1} \\ \vdots \\ \sqrt{\hat{w}_p}Y_{p,n_p} \end{bmatrix} = \begin{bmatrix} \sqrt{\hat{w}_1} & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ \sqrt{\hat{w}_1} & 0 & 0 & \dots & 0 \\ 0 & \sqrt{\hat{w}_2} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & \sqrt{\hat{w}_2} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & \sqrt{\hat{w}_p} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & \sqrt{\hat{w}_p} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1,n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2,n_2} \\ \vdots \\ \epsilon_{p,1} \\ \vdots \\ \epsilon_{p,n_p} \end{bmatrix}. \quad (5.6)$$

Four tests for $Ho : \mu_1 = \cdots = \mu_p$ can be used if Rule of Thumb 5.1: $\max(S_1, ..., S_p) \leq 2\min(S_1, ..., S_p)$ fails. Let $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$, and let $Y_{(1)} \leq Y_{(2)} \cdots \leq Y_{(n)}$ be the order statistics. Then the rank transformation of the response is $\boldsymbol{Z} = rank(\boldsymbol{Y})$ where $Z_i = j$ if $Y_i = Y_{(j)}$ is the jth order statistic. For example, if $\boldsymbol{Y} = (7.7, 4.9, 33.3, 6.6)^T$, then $\boldsymbol{Z} = (3, 1, 4, 2)^T$. The first test performs the one way ANOVA F test with $\boldsymbol{Z}$ replacing $\boldsymbol{Y}$. See Montgomery

(1984, p. 117-118). Two of the next three tests are described in Brown and Forsythe (1974b). Let $\lceil x \rceil$ be the smallest integer $\geq x$, eg $\lceil 7.7 \rceil = 8$. Then the Welch (1951) ANOVA F test uses test statistic

$$F_W = \frac{\sum_{i=1}^{p} w_i(\overline{Y}_{i0} - \tilde{Y}_{00})^2/(p-1)}{1 + \frac{2(p-2)}{p^2-1}\sum_{i=1}^{p}(1-\frac{w_i}{u})^2/(n_i-1)}$$

where $w_i = n_i/S_i^2, u = \sum_{i=1}^{p} w_i$ and $\tilde{Y}_{00} = \sum_{i=1}^{p} w_i \overline{Y}_{i0}/u$. Then the test statistic is compared to an F_{p-1,d_W} distribution where $d_W = \lceil f \rceil$ and

$$1/f = \frac{3}{p^2-1}\sum_{i=1}^{p}(1-\frac{w_i}{u})^2/(n_i-1).$$

For the modified Welch (1947) test, the test statistic is compared to an $F_{p-1,d_{MW}}$ distribution where $d_{MW} = \lceil f \rceil$ and

$$f = \frac{\sum_{i=1}^{p}(S_i^2/n_i)^2}{\sum_{i=1}^{p}\frac{1}{n_i-1}(S_i^2/n_i)^2} = \frac{\sum_{i=1}^{p}(1/w_i)^2}{\sum_{i=1}^{p}\frac{1}{n_i-1}(1/w_i)^2}.$$

Some software uses f instead of d_W or d_{MW}, and variants on the denominator degrees of freedom d_W or d_{MW} are common.

The modified ANOVA F test uses test statistic

$$F_M = \frac{\sum_{i=1}^{p} n_i(\overline{Y}_{i0} - \overline{Y}_{00})^2}{\sum_{i=1}^{p}(1-\frac{n_i}{n})S_i^2}$$

The test statistic is compared to an F_{p-1,d_M} distribution where $d_M = \lceil f \rceil$ and

$$1/f = \sum_{i=1}^{p} c_i^2/(n_i-1)$$

where

$$c_i = (1-\frac{n_i}{n})S_i^2/[\sum_{i=1}^{p}(1-\frac{n_i}{n})S_i^2].$$

The `regpack` function *anovasim* can be used to compare the five tests.

5.7 Problems

Problems with an asterisk * are especially important.

```
Output for Problem 5.1.
  A     B    C    D     E
 9.8   9.8  8.5  7.9   7.6  Analysis of Variance for Time
10.3  12.3  9.6  6.9  10.6  Source DF SS    MS     F     P
13.6  11.1  9.5  6.6   5.6  Design  4 44.88 11.22  5.82  0.002
10.5  10.6  7.4  7.6  10.1  Error  25 48.17  1.93
 8.6  11.6  7.6  8.9  10.5  Total  29 93.05
11.1  10.9  9.9  9.1   8.6
```

5.1. In a psychology experiment on child development, the goal is to study how different designs of mobiles vary in their ability to capture the infants' attention. Thirty 3-month-old infants are randomly divided into five groups of six each. Each group was shown a mobile with one of five designs A, B, C, D or E. The time that each infant spent looking at the design is recorded in the output above along with the Anova table. Data is taken from McKenzie and Goldman (1999, p. 234). See the above output.

a) Find $\hat{\mu}_2 = \hat{\mu}_B$.

b) Perform a 4 step Anova F test.

```
Output for Problem 5.2.
Variable    MEAN    SAMPLE SIZE  GROUP STD DEV
NONE        10.650       4        2.0535
N1000       10.425       4        1.4863
N5000        5.600       4        1.2437
N10000       5.450       4        1.7711
TOTAL        8.0312     16        1.6666

One Way Analysis of Variance Table
Source         df       SS           MS           F     p-value
Treatments      2    100.647       33.549      12.08    0.0006
Error          28     33.328        2.777
Total          15    133.974
```

```
Bonferroni Comparison of Means
                    Homogeneous
Variable    Mean    Groups
----------------------------
NONE        10.650   I
N1000       10.425   I
N5000        5.600  .. I
N10000       5.450  .. I
```

5.2. Moore (2000, p. 526): Nematodes are microscopic worms. A botanist desires to learn how the presence of the nematodes affects tomato growth. She uses 16 pots each with a tomato seedling. Four pots get 0 nematodes, four get 1000, four get 5000, and four get 10000. These four groups are denoted by "none," "n1000," "n5000" and "n10000" respectively. The seedling growths were all recorded and the table on the previous page gives the one way ANOVA results.

a) What is $\hat{\mu}_{none}$?

b) Do a four step test for whether the four mean growths are equal. (So Ho: $\mu_{none} = \mu_{n1000} = \mu_{n5000} = \mu_{n10000}$.)

c) Examine the Bonferroni comparison of means. Which groups of means are not significantly different?

5.3. According to Cobb (1998, p. 9) when the famous statistician W. G. Cochran was starting his career, the experiment was to study rat nutrition with two diets: ordinary rat food and rat food with a supplement. It was thought that the diet with the supplement would be better. Cochran and his coworker grabbed rats out of a cage, one at a time, and Cochran assigned the smaller less healthy rats to the better diet because he felt sorry for them. The results were as expected for the rats chosen by Cochran's coworker, but the better diet looked bad for Cochran's rats.

a) What were the units?

b) Suppose rats were taken from the cage one at a time. How should the rats have been assigned to the two diets?

5.4. Use the output from the command below

```
> sample(11)
 [1]  7 10  9  8  1  6  3 11  2  4  5
```

to assign the following 11 people to three groups of size $n_1 = n_2 = 4$ and $n_3 = 3$.

Anver, Arachchi, Field, Haenggi, Hazaimeh,
Liu, Pant, Tosun, Yi, Zhang, Zhou

5.5. Sketch a good response plot if there are 4 levels with $\overline{Y}_{10} = 2$, $\overline{Y}_{20} = 4$, $\overline{Y}_{30} = 6$, $\overline{Y}_{40} = 7$, and $n_i = 5$.

```
output for problem 5.6
level     1          2          3          4          5
     15 percent 20 percent 25 percent 30 percent 35 percent
```

$\overline{y}_1$	$\overline{y}_5$	$\overline{y}_2$	$\overline{y}_3$	$\overline{y}_4$
9.8	10.8	15.4	17.6	21.6
___	___	___	___	

5.6. The tensile strength of a cotton nylon fiber used to make women's shirts is believed to be affected by the percentage of cotton in the fiber. The 5 levels of cotton percentage that are of interest are tabled above. Also shown is a (Tukey pairwise) comparison of means. Which groups of means are not significantly different? Data is from Montgomery (1984. p. 51, 66).

```
output for problem 5.7
Source       df     SS       MS        F         P
color         2     7.60     3.80      0.390   0.684
error        12    116.40    9.70
```

5.7. A researcher is interested in whether the color (red, blue or green) of a paper maze effects the time to complete the maze.

a) State whether this is a random or fixed effects one way Anova. Explain briefly.

b) Using the output above, perform the appropriate 4 step Anova F test.

```
A     B    C          Output for problem 5.8.
9.5   8.5  7.7      Analysis of Variance for Time
3.2   9.0  11.3     Source   DF   SS      MS     F      P
4.7   7.9  9.7      Design    2   49.168  24.584 4.4625 0.0356
7.5   5.0  11.5      Error   12   66.108   5.509
8.3   3.2  12.4
```

5.8. Ledolter and Swersey (2007, p. 49) describe a one way Anova design used to study the effectiveness of 3 product displays (A, B and C). Fifteen stores were used and each display was randomly assigned to 5 stores. The response Y was the sales volume for the week during which the display was present compared to the base sales for that store.

a) Find $\hat{\mu}_2 = \hat{\mu}_B$.

b) Perform a 4 step Anova F test.

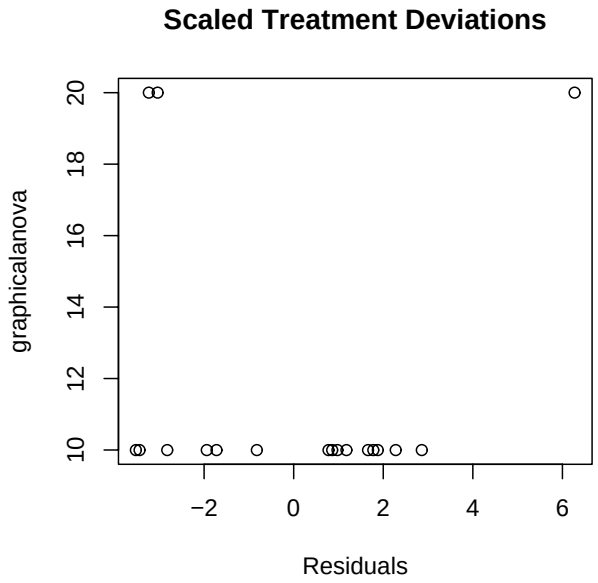

Figure 5.4: Graphical Anova for Problem 5.9

```
ganova(x,y)
smn    -3.233326 -3.037367  6.270694
Treatments "A"      "B"        "C"
```

5.9. Ledolter and Swersey (2007, p. 49) describe a one way Anova design used to study the effectiveness of 3 product displays (A, B and C). Fifteen stores were used and each display was randomly assigned to 5 stores. The response Y was the sales volume for the week during which the display was present compared to the base sales for that store. Figure 5.4 is the Graphical Anova plot found using the function `ganova`.

a) Which two displays (from A, B and C) yielded similar mean sales volume?

b) Which effect (from A, B and C) appears to be significant?

Source	df	SS	MS	F	P
treatment	3	89.19	29.73	15.68	0.0002
error	12	22.75	1.90		

5.10. A textile factory weaves fabric on a large number of looms. They would like to obtain a fabric of uniform strength. Four looms are selected at random and four samples of fabric are obtained from each loom. The strength of each fabric sample is measured. Data is from Montgomery (1984, p. 74-75).

a) State whether this is a random or fixed effects one way Anova. Explain briefly.

b) Using the output above, perform the appropriate 4 step Anova F test.

Problems using R/Splus.

Warning: Use the command *source("A:/regpack.txt")* **to download the programs**, and *source("A:/regdata.txt")* **to download the data. See Preface or Section 17.1.** Typing the name of the `regpack` function, eg *pcisim*, will display the code for the function. Use the `args` command, eg *args(pcisim)*, to display the needed arguments for the function.

5.11. The pooled t procedures are a special case of one way Anova with $p = 2$. Consider the pooled t CI for $\mu_1 - \mu_2$. Let $X_1, ..., X_{n_1}$ be iid with mean μ_1 and variance σ_1^2. Let $Y_1, ..., Y_{n_2}$ be iid with mean μ_2 and variance σ_2^2. Assume

that the two samples are independent (or that $n_1 + n_2$ units were randomly assigned to two groups) and that $n_i \to \infty$ for $i = 1, 2$ in such a way that $\hat{\rho} = \frac{n_1}{n_1+n_2} \to \rho \in (0, 1)$. Let $\theta = \sigma_2^2/\sigma_1^2$, and let the pooled sample variance $S_p^2 = [(n_1-1)S_1^2+(n_2-1)S_2^2]/[n_1+n_2-2]$ and $\tau^2 = (1-\rho+\rho\theta)/[\rho+(1-\rho)\theta]$. Then

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \xrightarrow{D} N(0, 1)$$

and

$$\frac{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \;=\; \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \xrightarrow{D} N(0, \tau^2).$$

Now let $\hat{\theta} = S_2^2/S_1^2$ and $\hat{\tau}^2 = (1 - \hat{\rho} + \hat{\rho}\ \hat{\theta})/(\hat{\rho} + (1 - \hat{\rho})\ \hat{\theta})$. Notice that $\hat{\tau} = 1$ if $\hat{\rho} = 1/2$, and $\hat{\tau} = 1$ if $\hat{\theta} = 1$. The usual large sample $(1 - \alpha)100\%$ pooled t CI for $(\mu_1 - \mu_2)$ is

$$\overline{X} - \overline{Y} \ \pm \ t_{n_1+n_2-2,1-\alpha/2} \ \ S_p \ \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{5.7}$$

is valid if $\tau = 1$. The large sample $(1 - \alpha)100\%$ modified pooled t CI for $(\mu_1 - \mu_2)$ is

$$\overline{X} - \overline{Y} \ \pm \ t_{n_1+n_2-4,1-\alpha/2} \ \hat{\tau} \ S_p \ \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}. \tag{5.8}$$

The large sample $(1 - \alpha)100\%$ Welch CI for $(\mu_1 - \mu_2)$ is

$$\overline{X} - \overline{Y} \ \pm \ t_{d,1-\alpha/2} \ \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \tag{5.9}$$

where $d = max(1, [d_0])$, and

$$d_0 = \frac{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})^2}{\frac{1}{n_1-1}(\frac{S_1^2}{n_1})^2 + \frac{1}{n_2-1}(\frac{S_2^2}{n_2})^2}.$$

Suppose $n_1/(n_1 + n_2) \to \rho$. It can be shown that if the CI length is multiplied by $\sqrt{n_1}$, then the scaled length of the pooled t CI converges in probability to $2z_{1-\alpha/2}\sqrt{\frac{\rho}{1-\rho}\sigma_1^2 + \sigma_2^2}$ while the scaled lengths of the modified pooled

t CI and Welch CI both converge in probability to $2z_{1-\alpha/2}\sqrt{\sigma_1^2 + \frac{\rho}{1-\rho}\sigma_2^2}$. The pooled t CI should have coverage that is too low if

$$\frac{\rho}{1-\rho}\sigma_1^2 + \sigma_2^2 \ < \ \sigma_1^2 + \frac{\rho}{1-\rho}\sigma_2^2.$$

See Olive (2009b, Example 9.23).

a) Download the function `pcisim`.

b) Type the command
`pcisim(n1=100,n2=200,var1=10,var2=1)` to simulate the CIs for $N(\mu_i, \sigma_i^2)$ data for $i = 1, 2$. The terms *pcov*, *mpcov* and *wcov* are the simulated coverages for the pooled, modified pooled and Welch 95% CIs. Record these quantities. Are they near 0.95?

5.12. From the end of Section 5.6, four tests for $Ho : \mu_1 = \cdots = \mu_k$ can be used if Rule of Thumb: $\max(S_1, ..., S_k) \leq 2\min(S_1, ..., S_k)$ fails. In R, get the function `anovasim`. When H_0 is true, the coverage = proportion of times the test rejects H_0 has a nominal value of 0.05. The terms faovcov is for the usual F test, modfcov is for a modified F test, wfcov is for the Welch test, mwfcov for the modified Welch test and rfcov for the rank test. The function generates 1000 data sets with $k = 4$, ni = $n_i = 20$, mi = μ_i and sdi = σ_i.

a) Get the coverages for the following command. Since the four population means and the four population standard deviations are equal, want the coverages to be near or less than 0.05. Are they? anovasim(m1 = 0, m2 = 0, m3 = 0, m4 = 0, sd1 = 1, sd2 = 1, sd3 = 1, sd4 = 1)

b) Get the coverages for the following command. The population means are equal, but the population standard deviations are not. Are the coverages near or less than 0.05? anovasim(m1 = 0, m2 = 0, m3 = 0, m4 = 0, sd1 = 1, sd2 = 2, sd3 = 3, sd4 = 4)

c) Now use the following command where H_0 is false: the four population means are not all equal. Want the coverages near 1. Are they?
anovasim(m1 = 1, m2 = 0, m3 = 0, m4 = 1)

d) Now use the following command where H_0 is false. Want the coverages near 1. Since the σ_i are not equal, the Anova F test is expected to perform poorly. Is the Anova F test the best?
anovasim(m4 = 1, s4 = 9)

5.13. This problem uses data from Kuehl (1994, p. 128).
a) Get *regdata* and *regpack* into R. Type the following commands. Then

simultaneously press the *Ctrl* and *c* keys. In *Word* use the menu command "Edit>Paste." Print out the figure.

```
y <- ycrab+1/6
aovtplt(crabhab,y)
```

b) From the figure, what response transformation should be used: $Y = 1/Z, Y = 1/\sqrt{Z}, Y = \log(Z), Y = \sqrt{Z}$, or $Y = Z$?

5.14. The following data set considers the number of warp breaks per loom, where the factor is tension (low, medium or high). The commands for this problem can be found at (www.math.siu.edu/olive/reghw.txt).

a) Type the following commands:

```
help(warpbreaks)
out <- aov(breaks ~ tension, data = warpbreaks)
out
summary(out)
plot(out$fit,out$residuals)
title("Residual Plot")
```

Highlight the ANOVA table by pressing the left mouse key and dragging the cursor over the ANOVA table. Then use the menu commands "Edit> Copy." Enter *Word* and use the menu commands "Edit>Paste."

b) To place the residual plot in *Word*, get into *R* and click on the plot, hit the *Ctrl* and *c* keys at the same time. Enter *Word* and use the menu commands "Edit>Paste."

c) Type the following commands:

```
warpbreaks[1,]
plot(out$fit,warpbreaks[,1])
abline(0,1)
title("Response Plot")
```

Click on the response plot, hit the *Ctrl* and *c* keys at the same time. Enter *Word* and use the menu commands "Edit>Paste."

5.15. Obtain the Box, Hunter and Hunter (2005, p. 134) blood coagulation data from (www.math.siu.edu/olive/regdata.txt) and the *R* program `ganova` from (www.math.siu.edu/olive/regpack.txt). The program does graphical Anova for the one way Anova model.

a) Enter the following commands and include the plot in *Word* by simultaneously pressing the *Ctrl* and *c* keys, then using the menu commands "Copy>Paste" in *Word*.

```
ganova(bloodx,bloody)
```

The scaled treatment deviations are on the top of the plot. As a rule of thumb, if all of the scaled treatment deviations are within the spread of the residuals, then population treatment means are not significantly different (they all give response near the grand mean). If some deviations are outside of the spread of the residuals, then not all of the population treatment means are equal. Box, Hunter and Hunter (2005, p. 137) state 'The graphical analysis discourages overreaction to high significance levels and avoids underreaction to "very nearly" significant differences.'

b) From the output, which two treatments means were approximately the same?

c) To perform a randomization F test in *R*, get the program `rand1way` from (www.math.siu.edu/olive/regpack.txt), and type the following commands. The output z$rdist is the randomization distribution, z$Fpval is the pvalue of the usual F test, and z$randpval is the pvalue of the randomized F test.

```
z<-rand1way(y=bloody,group=bloodx,B=1000)
hist(z$rdist)
z$Fpval
z$randpval
```

d) Include the histogram in *Word*.

One Way Anova in SAS

To get into *SAS*, often you click on a *SAS* icon, perhaps something like *The SAS System for ....* A window with a split screen will open. The top screen says *Log-(Untitled)* while the bottom screen says *Editor-Untitled1*. Press the spacebar and an asterisk appears: *Editor-Untitled1**.

For problem 5.16, consider saving your file as hw5d16.sas on your diskette (A: drive). (On the top menu of the editor, use the commands "File > Save as". A window will appear. Use the upper right arrow to locate "31/2 Floppy A" and then type the file name in the bottom box. Click on OK.) From the

top menu in SAS, use the "File> Open" command. A window will open. Use the arrow in the NE corner of the window to navigate to "31/2 Floppy(A:)". (As you click on the arrow, you should see My Documents, C: etc, then 31/2 Floppy(A:).) Double click on **hw5d16.sas**.

This point explains the SAS commands. The semicolon ";" is used to end SAS commands and the "options ls = 70;" command makes the output readable. (An "*" can be used to insert comments into the SAS program. Try putting an * before the options command and see what it does to the output.) The next step is to get the data into SAS. The command "data clover;" gives the name "clover" to the data set. The command "input strain $ nitrogen @ @;" says the first entry is variable strain and the $ means it is categorical, the second variable is nitrogen and the @@ means read 2 variables, then 2, ..., until the end of the data. The command "cards;" means that the data is entered below. Then the data in entered and the isolated semicolon indicates that the last case has been entered.

The commands "proc glm; class = strain; model nitrogen = strain;" tells SAS to perform one way Anova with nitrogen as the response variable and strain as the factor.

5.16. Cut and paste the SAS program from (www.math.siu.edu/olive/reghw.txt) for 5.16 into the SAS Editor.

To execute the program, use the top menu commands "Run>Submit". An output window will appear if successful.

(If you were not successful, look at the *log window* for hints on errors. A single typo can cause failure. Reopen your file in *Word* or *Notepad* and make corrections. Occasionally you can not find your error. Then find your instructor or wait a few hours and reenter the program.)

Data is from SAS Institute (1985, p. 126-129). See Example 5.6.

a) In SAS, use the menu commands "Edit>Select All" then "Edit>Copy." In *Word*, use the menu commands "Edit>Paste." Highlight the first page of output and use the menu commands "Edit>Cut." (SAS often creates too much output. These commands reduce the output from 4 pages to 3 pages.)

You may want to save your SAS output as the file HW5d16.doc on your disk.

b) Perform the 4 step test for Ho $\mu_1 = \mu_2 = \cdots = \mu_6$.

c) From the residual and response plots, does the assumption of equal

population standard deviations ($\sigma_i = \sigma$ for $i = 1, ..., 6$) seem reasonable?

One Way Anova in ARC

5.17. To get in `ARC`, you need to find the `ARC` icon. Suppose the `ARC` icon is in a *math progs* folder. Move the cursor to the math progs folder, click the right mouse button twice, move the cursor to `ARC`, double click, move the cursor to `ARC`, double click. These menu commands will be written "math progs > ARC > ARC." To quit `ARC`, move cursor to the **x** in the northeast corner and click.

This Cook and Weisberg (1999, p. 289) data set contains IQ scores on 27 pairs of identical twins, one raised by foster parents IQf and the other by biological parents IQb. C gives the social class of the biological parents: $C = 1$ for upper class, 2 for middle class and 3 for lower class. Hence the Anova test is for whether mean IQ depends on class.

a) Activate *twins.lsp* dataset with the menu commands
"File > Load > Data > ARCG > twins.lsp".

b) Use the menu commands "Twins>Make factors", select C and click on *OK*. The line "{F}C Factor 27 Factor–first level dropped" should appear on the screen.

c) Use the menu commands "Twins>Description" to see a description of the data.

d) Enter the menu commands "Graph&Fit>Fit linear LS" and select {F}C as the term and IQb as the response. Highlight the output by pressing the left mouse key and dragging the cursor over the output. Then use the menu commands "Edit> Copy." Enter *Word* and use the menu commands "Edit>Paste."

e) Enter the menu commands "Graph&Fit>Boxplot of" and enter IQb in the *selection box* and C in the *Condition on* box. Click on *OK*. When the boxplots appear, click on the *Show Anova* box. Click on the plot, hit the *Ctrl* and *c* keys at the same time. Enter *Word* and use the menu commands "Edit>Paste." Include the output in *Word*. Notice that the regression and Anova F statistic and p-value are the same.

f) Residual plot: Enter the menu commands "Graph&Fit>Plot of," select "L1:Fit-Values" for the "H" box and "L1:Residuals" for the "V" box, and click on "OK." Click on the plot, hit the *Ctrl* and *c* keys at the same time. Enter *Word* and use the menu commands "Edit>Paste."

g) Response plot: Enter the menu commands "Graph&Fit>Plot of," se-

lect "L1:Fit-Values" for the "H" box and "IQb" for the "V" box, and click on "OK." When the plot appears, move the OLS slider bar to 1 to add the identity line. Click on the plot, hit the *Ctrl* and *c* keys at the same time. Enter *Word* and use the menu commands "Edit>Paste."

h) Perform the 4 step test for Ho $\mu_1 = \mu_2 = \mu_3$.

One Way Anova in Minitab

5.18. a) In Minitab, use the menu command "File>Open Worksheet" and double click on *Baby.mtw*. A window will appear. Click on "OK."

This McKenzie and Goldman (1999, p. T-234) data set has 30 three month old infants randomized into five groups of 6 each. Each infant is shown a mobile of one of five multicolored designs, and the goal of the study is to see if the infant attention span varies with type of design of mobile. The times that each infant spent watching the mobile are recorded.

b) Choose "Stat>Basic Statistics>Display Descriptive Statistics," select "C1 Time" as the "Variable," click the "By variable" option and press *Tab*. Select "C2 Design" as the "By variable."

c) From the window in b), click on "Graphs" the "Boxplots of data" option, and "OK" twice. Click on the plot and then click on the *printer* icon to get a plot of the boxplots.

d) Select "Stat>ANOVA>One-way," select "C1-time" as the response and "C2-Design" as the factor. Click on "Store residuals" and click on "Store fits." Then click on "OK." Click on the output and then click on the *printer* icon.

e) To make a residual plot, select "Graph>Plot." Select "Resi1" for "Y" and "Fits1" for "X" and click on "OK." Click on the plot and then click on the *printer* icon to get the residual plot.

f) To make a response plot, select "Graph>Plot." Select "C1 Time" for "Y" and "Fits1" for "X" and click on "OK." Click on the plot and then click on the *printer* icon to get the response plot.

g) Do the 4 step test for Ho $\mu_1 = \mu_2 = \cdots = \mu_5$.

To get out of Minitab, move your cursor to the "x" in the NE corner of the screen. When asked whether to save changes, click on "no."