5 One-Way ANOVA (Review) and Experimental Design

Samuels and Witmer Chapter 11 - all sections except 6.

The one-way analysis of variance (**ANOVA**) is a generalization of the two sample $t-$test to $k \geq 2$ groups. Assume that the populations of interest have the following (unknown) population means and standard deviations:

	population 1	population 2	$\cdots$	population k
mean	μ_1	μ_2	$\cdots$	μ_k
std dev	σ_1	σ_2	$\cdots$	σ_k

A usual interest in ANOVA is whether $\mu_1 = \mu_2 = \cdots = \mu_k$. If not, then we wish to know which means differ, and by how much. To answer these questions we select samples from each of the k populations, leading to the following data summary:

	sample 1	sample 2	$\cdots$	sample k
size	n_1	n_2	$\cdots$	n_k
mean	$\overline{Y}_1$	$\overline{Y}_2$	$\cdots$	$\overline{Y}_k$
std dev	s_1	s_2	$\cdots$	s_k

A little more notation is needed for the discussion. Let Y_{ij} denote the j^{th} observation in the i^{th} sample and define the total sample size $n^* = n_1 + n_2 + \cdots + n_k$. Finally, let $\overline{\overline{Y}}$ be the average response over all samples (combined), that is

$$\overline{\overline{Y}} = \frac{\sum_{ij} Y_{ij}}{n^*} = \frac{\sum_i n_i \overline{Y}_i}{n^*}.$$

Note that $\overline{\overline{Y}}$ is *not* the average of the sample means, unless the sample sizes n_i are equal.

An $F-$statistic is used to test $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ against $H_A :$ not H_0. The assumptions needed for the standard ANOVA $F-$test are analogous to the independent two-sample $t-$test assumptions: (1) Independent random samples from each population. (2) The population frequency curves are normal. (3) The populations have equal standard deviations, $\sigma_1 = \sigma_2 = \cdots = \sigma_k$.

The $F-$test is computed from the ANOVA table, which breaks the spread in the combined data set into two components, or **Sums of Squares** (SS). The **Within SS**, often called the **Residual SS** or the **Error SS**, is the portion of the total spread due to variability *within* samples:

$$\text{SS(Within)} = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2 = \textstyle\sum_{ij}(Y_{ij} - \overline{Y}_i)^2.$$

The **Between SS**, often called the Model SS, measures the spread between (actually among!) the sample means

$$\text{SS(Between)} = n_1(\overline{Y}_1 - \overline{\overline{Y}})^2 + n_2(\overline{Y}_2 - \overline{\overline{Y}})^2 + \cdots + n_k(\overline{Y}_k - \overline{\overline{Y}})^2 = \textstyle\sum_i n_i(\overline{Y}_i - \overline{\overline{Y}})^2,$$

weighted by the sample sizes. These two SS add to give

$$\text{SS(Total)} = \text{SS(Between)} + \text{SS(Within)} = \textstyle\sum_{ij}(Y_{ij} - \overline{\overline{Y}})^2.$$

Each SS has its own degrees of freedom (*df*). The *df*(Between) is the number of groups minus one, $k - 1$. The *df*(Within) is the total number of observations minus the number of groups: $(n_1-1)+(n_2-1)+\cdots(n_k-1) = n^*-k$. These two *df* add to give *df*(Total) $= (k-1)+(n^*-k) = n^*-1$.

The Sums of Squares and *df* are neatly arranged in a table, called the ANOVA table:

Source	*df*	SS	MS
Between Groups	$k-1$	$\sum_i n_i(\overline{Y}_i - \overline{\overline{Y}})^2$	
Within Groups	$n^* - k$	$\sum_i (n_i - 1)s_i^2$	
Total	$n^* - 1$	$\sum_{ij}(Y_{ij} - \overline{\overline{Y}})^2$.	

The ANOVA table often gives a **Mean Squares** (MS) column, left blank here. The Mean Square for each source of variation is the corresponding SS divided by its *df*. The Mean Squares can be easily interpreted.

The MS(Within)

$$\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \cdots + (n_k-1)s_k^2}{n^* - k} = s^2_{pooled}$$

is a weighted average of the sample variances. The MS(Within) is known as the pooled estimator of variance, and estimates the assumed common population variance. If all the sample sizes are equal, the MS(Within) is the average sample variance. The MS(Within) is identical to the **pooled variance estimator** in a two-sample problem when $k = 2$.

The MS(Between)

$$\frac{\sum_i n_i(\overline{Y}_i - \overline{\overline{Y}})^2}{k-1}$$

is a measure of variability among the sample means. This MS is a multiple of the sample variance of $\overline{Y}_1,\ \overline{Y}_2, ..., \overline{Y}_k$ when all the sample sizes are equal.

The MS(Total)

$$\frac{\sum_{ij}(Y_{ij} - \overline{\overline{Y}})^2}{n^* - 1}$$

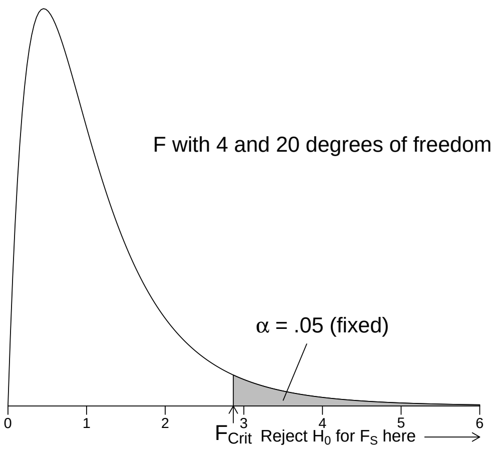
is the variance in the combined data set.

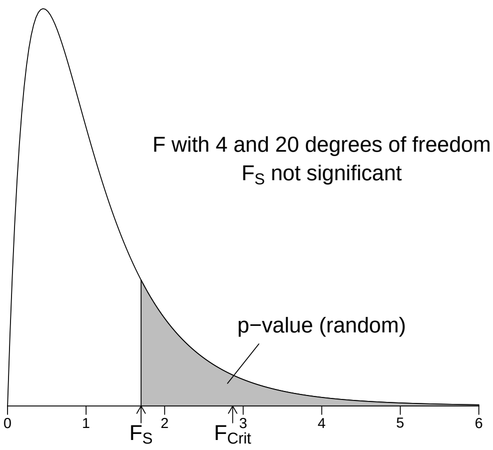
The decision on whether to reject $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ is based on the ratio of the MS(Between) and the MS(Within):

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

Large values of F_s indicate large variability among the sample means $\overline{Y}_1,\ \overline{Y}_2, ..., \overline{Y}_k$ relative to the spread of the data within samples. That is, large values of F_s suggest that H_0 is false.

Formally, for a size α test, reject H_0 if $F_s \geq F_{crit}$, where F_{crit} is the upper-α percentile from an F distribution with numerator degrees of freedom $k-1$ and denominator degrees of freedom $n^* - k$ (i.e. the *df* for the numerators and denominators in the $F-$ratio.). An F distribution table is given on pages 687-696 of SW. The p-value for the test is the area under the $F-$ probability curve to the right of F_s. **Stata** summarizes the ANOVA $F-$test with a p-value. In **Stata**, use the `anova` or `oneway` commands to perform 1-way ANOVA. The data should be in the form of a variable containing the response Y_{ij} and a grouping variable. For $k = 2$, the test is equivalent to the pooled two-sample $t-$test.



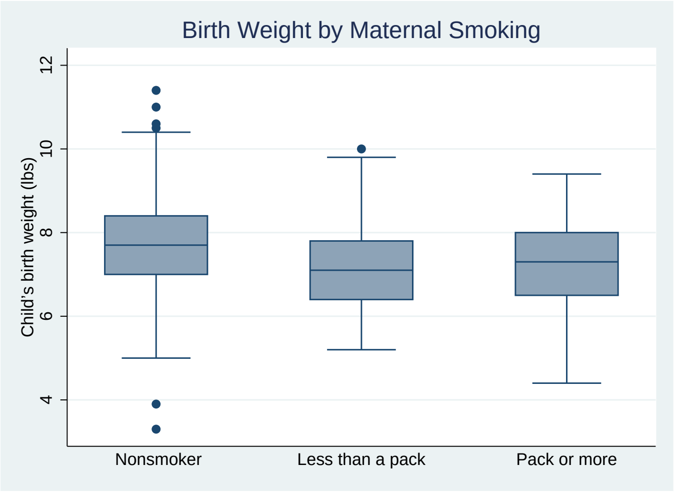


Example from the Child Health and Development Study (CHDS)

We consider data from the birth records of 680 live-born white male infants. The infants were born to mothers who reported for pre-natal care to three clinics of the Kaiser hospitals in northern California. As an initial analysis, we will examine whether maternal smoking has an effect on the birth weights of these children. To answer this question, we define 3 groups based on mother's smoking history: (1) mother does not currently smoke or never smoked (2) mother smoked less than one pack of cigarettes a day during pregnancy (3) mother smoked at least one pack of cigarettes a day during pregnancy.

Let μ_i = pop mean birth weight (in lbs) for children in group i, $(i = 1, 2, 3)$. We wish to test $H_0 : \mu_1 = \mu_2 = \mu_3$ against $H_A :$ not H_0.

The side-by-side boxplots of the data show roughly the same spread among groups and little evidence of skew:



There is no strong evidence against normality here. Furthermore the sample standard deviations are close (see the following output). We may formally test the equality of variances across the three groups (remember - the F-test is not valid if its assumptions are not met) using **Stata**'s `robvar` command. In this example we obtain a set of three *robust* tests for the hypothesis $H_0 : \sigma_1 = \sigma_2 = \sigma_3$ where σ_i is the population standard deviation of weight in group i, $i = 1, 2, 3$. What *robust* means in this context is that the test still works reasonably well if assumptions are not quite met. The classical test of this hypothesis is Bartlett's test, and that test is well known to be extraordinarily sensitive to the assumption of normality of all the distributions. There are two ways a test may not work well when assumptions are violated - the level may not be correct, or the power may be poor. For Bartlett's test, the problem is the level may not be accurate, which in this case means that you may see a small p-value that does not reflect unequal variances but instead reflects non-normality. A test with this property is known as *liberal* because it rejects H_0 too often (relative to the nominal α).

Stata output follows; we do not reject that the variances are equal across the three groups at any reasonable significance level using any of the three test statistics:

```
. robvar( weight),by(ms_gp_txt)
                |   Summary of Child's birth weight
                |              (lbs)
       ms_gp_txt |        Mean   Std. Dev.       Freq.
-----------------+------------------------------------
       Nonsmoker |   7.7328084   1.0523406         381
Less than a pack |   7.2213018   1.0777604         169
    Pack or more |   7.2661539   1.0909461         130
-----------------+------------------------------------
           Total |   7.5164706   1.0923455         680
W0  = .82007944   df(2, 677)     Pr > F = .44083367
W50 = .75912861   df(2, 677)     Pr > F = .46847213
W10 = .77842523   df(2, 677)     Pr > F = .45953896
```

There are multiple ways to get the ANOVA table here, the most common being the command `anova weight ms_gp` or the more specialized `oneway weight ms_gp`. Bartlett's test for equal variances is given when using the latter. In the following output, I also gave the `,b` option to get Bonferroni multiple comparisons, discussed after the Fisher's Method (next section).

The ANOVA table is:

```
. oneway weight ms_gp_txt,b
                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      40.7012466      2   20.3506233     17.90     0.0000
Within groups        769.4943     677   1.13662378
------------------------------------------------------------------------
   Total            810.195546    679   1.19321877
Bartlett's test for equal variances:  chi2(2) =   0.3055  Prob>chi2 = 0.858
           Comparison of Child's birth weight (lbs) by ms_gp_txt
                               (Bonferroni)
Row Mean-|
Col Mean |    Nonsmok   Less tha
---------+----------------------
Less tha |   -.511507
         |      0.000
         |
Pack or  |   -.466655    .044852
         |      0.000      1.000
```

The p-value for the $F-$test is less than .0001. We would reject H_0 at any of the usual test levels (i.e. .05 or .01), concluding that the population mean birth weights differ *in some way* across smoking status groups. The data (boxplots) suggest that the mean birth weights are higher for children born to mothers that did not smoke during pregnancy, but that is not a legal conclusion based upon the F-test alone.

The **Stata** commands to obtain this analysis are:

```
infile id head ...et cetera... pheight using c:/chds.txt
generate ms_gp = 1 if msmoke == 0
replace  ms_gp = 2 if msmoke >= 1 & msmoke < 20
replace  ms_gp = 3 if msmoke >= 20
gene ms_gp_txt = " Nonsmoker" if ms_gp ==1
replace ms_gp_txt = "Less than a pack" if ms_gp ==2
replace ms_gp_txt = "Pack or more" if ms_gp ==3
graph box weight, over(ms_gp_txt)
robvar weight, by (ms_gp_txt)
oneway weight ms_gp_txt,b
```

Multiple Comparison Methods: Fisher's Method

The ANOVA $F-$test checks whether all the population means are equal. **Multiple comparisons** are often used as a follow-up to a significant ANOVA $F-$test to determine which population means are different. I will discuss Fisher, Bonferroni, and Tukey methods for comparing all pairs of means. Fisher's and Tukey's approaches are implemented in **Stata** using **Stata**'s `prcomp` command. This command is not automatically installed in **Stata** 8 or 9. You will have to search for "pairwise comparisons" under **Help** > **Search...** and click on the blue `sg101` link. Click on `[Click here to install]` (your computer must be connected to the internet to do this) and you will then have access to this command.

Fisher's Least significant difference method **(LSD or FSD)** is a two-step process:

1. Carry out the ANOVA $F-$test of $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ at the α level. If H_0 is not rejected, stop and conclude that there is insufficient evidence to claim differences among population means. If H_0 is rejected, go to step 2.

2. Compare each pair of means using a pooled two sample $t-$test at the α level. Use s_{pooled} from the ANOVA table and $df = df$(Residual). Using this denominator is different from just doing all the possible pair-wise t-tests.

To see where the name LSD originated, consider the $t-$test of $H_0 : \mu_i = \mu_j$ (i.e. populations i and j have same mean). The $t-$statistic is

$$t_s = \frac{\overline{Y}_i - \overline{Y}_j}{s_{pooled}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}.$$

You reject H_0 if $|t_s| \geq t_{crit}$, or equivalently, if

$$|\overline{Y}_i - \overline{Y}_j| \geq t_{crit}\ s_{pooled}\ \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}.$$

The minimum absolute difference between $\overline{Y}_i$ and $\overline{Y}_j$ needed to reject H_0 is the LSD, the quantity on the right hand side of this inequality.

Stata gives all possible comparisons between pairs of populations means. The error level (i.e. α) can be set to an arbitrary value using the `level()` subcommand, with 0.05 being the standard. Looking at the CI's in the **Stata** output, we conclude that the mean birth weights for children born to non-smoking mothers (group 1) is significantly different from the mean birth weights for each of the other two groups (2 and 3), since confidence intervals do not contain 0. The **Stata** command `prcomp weight ms_gp` produced the output (it needs group defined numerically); the default output includes CIs for differences in means. Alternatively, one obtains the p-values for testing the hypotheses that the population means are equal using the `test` subcommand. This is illustrated in the section on Tukey's method. Examining the output from the prcomp command, we see the FSD method is called the **t method** by **Stata**.

```
. prcomp weight ms_gp
                   Pairwise Comparisons of Means
Response variable (Y): weight      Child's birth weight (lbs)
Group variable (X):    ms_gp       Maternal Smoking Group
   Group variable (X): ms_gp        Response variable (Y): weight
-------------------------------    -------------------------------
      Level                             n        Mean        S.E.
------------------------------------------------------------------
          1                           381    7.732808     .053913
          2                           169    7.221302    .0829046
          3                           130    7.266154    .0956823
------------------------------------------------------------------

Individual confidence level: 95%    (t method)
Homogeneous error SD = 1.066126, degrees of freedom = 677

                                                                 95%
 Level(X)    Mean(Y)   Level(X)    Mean(Y)      Diff Mean    Confidence Limits
-------------------------------------------------------------------------------
        2   7.221302          1   7.732808     -.5115066   -.7049746  -.3180387
        3   7.266154          1   7.732808     -.4666546   -.6792774  -.2540318
                              2   7.221302      .0448521   -.1993527   .2890568
-------------------------------------------------------------------------------
```

Discussion of the FSD Method

With k groups, there are $c = \binom{k}{2} = \frac{k(k-1)}{2}$ pairs of means to compare in the second step of the FSD method. Each comparison is done at the α level, where for a generic comparison of the i^{th} and j^{th} populations

$$\alpha = \text{probability of rejecting } H_0 : \mu_i = \mu_j \text{ when } H_0 \text{ is true.}$$

This probability is called the **comparison error rate** or the **individual error rate**.

The individual error rate is not the only error rate that is important in multiple comparisons. The **family error rate** (FER), or the **experimentwise error rate**, is defined to be the probability of at least one false rejection of a true hypothesis $H_0 : \mu_i = \mu_j$ over all comparisons. When many comparisons are made, you *may* have a large probability of making one or more false rejections of true null hypotheses. In particular, when all c comparisons of two population means are performed, each at the α level, then $\alpha \leq FER \leq c\alpha$.

For example, in the birth weight problem where $k = 3$, there are $c = .5 * 3 * 2 = 3$ possible comparisons of two groups. If each comparison is carried out at the 5% level, then $.05 \leq FER \leq .15$. At the second step of the FSD method, you *could* have up to a 15% chance of claiming one or more pairs of population means are different if no differences existed between population means.

The first step of the FSD method is the ANOVA "screening" test. The multiple comparisons are carried out *only if* the $F-$test suggests that not all population means are equal. This screening test tends to deflate the FER for the two-step FSD procedure. However, the FSD method is commonly criticized for being extremely liberal (too many false rejections of true null hypotheses) when some, but not many, differences exist - especially when the number of comparisons is large. This conclusion is fairly intuitive. When you do a large number of tests, each, say, at the 5% level, then sampling variation alone will suggest differences in 5% of the comparisons where the H_0 is true. The number of false rejections could be enormous with a large number of comparisons. For example, chance variation alone would account for an average of 50 significant differences in 1000 comparisons each at the 5% level.

The Bonferroni Multiple Comparison Method

The Bonferroni method goes directly after the preceding relationship, $\alpha \leq FER \leq c\alpha$. To keep the FER below level α, do the individual tests at level $\frac{\alpha}{c}$, or equivalently multiply each of the reported p-values by c. This is, in practice, extremely conservative but it does guarantee the FER is below α. If you have, for instance, $c = 3$ comparisons to make, and a reported p-value from a t-test is .02 then the Bonferroni p-value is $3(.02) = .06$ and the difference would not be judged significant. With more comparisons it becomes extremely hard for the Bonferroni method to *find* anything. The FSD method tends to have a too-high FER, the Bonferroni method a too-low FER. Very often they agree.

Earlier (p. 69) we looked at the ANOVA output following the `oneway weight group,b` command. Examining that output we see p-values of 0 for testing $H_0 : \mu_1 = \mu_2$ and $H_0 : \mu_1 = \mu_3$, and a p-value of 1 for testing $H_0 : \mu_2 = \mu_3$ using the Bonferroni method. The Bonferroni tests see group 1 differing from both 2 and 3, and no difference between 2 and 3, in complete agreement with FSD.

Tukey's Multiple Comparison Method

One commonly used alternative to FSD and Bonferroni is **Tukey's** honest significant difference method **(HSD)**. Unlike FSD (but similar to Bonferroni), Tukey's method allows you to prespecify the FER, at the cost of making the individual comparisons more conservative than in FSD (but less conservative than Bonferroni).

To implement Tukey's method with a FER of α, reject $H_0 : \mu_i = \mu_j$ when

$$|\overline{Y}_i - \overline{Y}_j| \geq \frac{q_{crit}}{\sqrt{2}} s_{pooled} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}},$$

where q_{crit} is the α level critical value of the studentized range distribution (tables not in SW). The right hand side of this equation is called the HSD. For the birth weight data, the groupings based on the Tukey and Fisher methods are identical. We obtain Tukey's groupings via the **Stata** command `prcomp weight group, tukey test`. The differences with an asterisk next to them are significant (the |numerator| is larger than the denominator):

```
. prcomp weight ms_gp,tukey test
                   Pairwise Comparisons of Means
Response variable (Y): weight      Child's birth weight (lbs)
Group variable (X):    ms_gp       Maternal Smoking Group
   Group variable (X): ms_gp        Response variable (Y): weight
-------------------------------    -------------------------------
      Level                           n         Mean          S.E.
------------------------------------------------------------------
          1                        381     7.732808        .053913
          2                        169     7.221302       .0829046
          3                        130     7.266154       .0956823
------------------------------------------------------------------
Simultaneous significance level: 5%    (Tukey wsd method)
Homogeneous error SD = 1.066126, degrees of freedom = 677
          (Row Mean - Column Mean) / (Critical Diff)
Mean(Y) |   7.7328    7.2213 Level(X)
        |        1         2
--------+-------------------
  7.2213|  -.51151*
       2|   .23145
        |
  7.2662|  -.46665*   .04485
       3|   .25436    .29214
        |
```

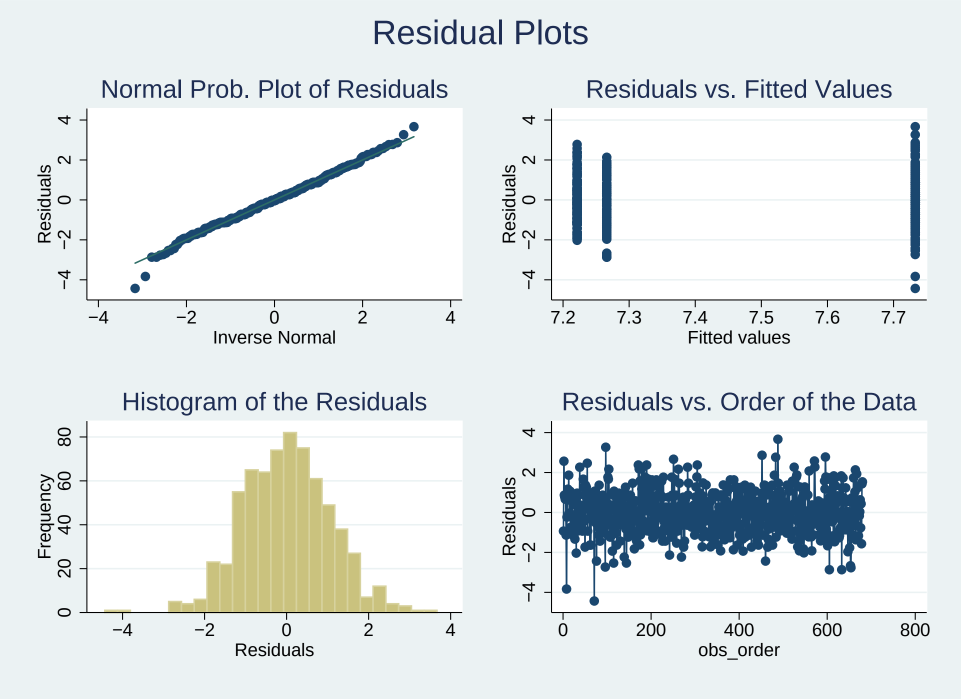
Stata does not provide, as built-in commands or options, very many multiple comparison procedures. The one-way ANOVA problem we have been looking at is relatively simple, and the Tukey method appears as something of an afterthought for it. For more complicated multi-factor models, about all **Stata** offers is Bonferroni and a two other methods (Holm and Sidak) that adjust the p-value similarly using slightly different principles to control FER, but less conservatively than Bonferroni. The help file on mtest has details. FSD is always available, since that amounts to no adjustment. In response to questions on the www about doing multiple comparisons, **Stata** has pointed out how easy it is to program whatever you want in do files (probably the right answer for experts). Some packages like **SAS** offer a larger number of options. What **Stata** offers is adequate for many areas of research, but for some others it will be necessary to go beyond the built-in offerings of **Stata** (a reviewer on your paper will let you know!)

Checking Assumptions in ANOVA Problems

The classical ANOVA assumes that the populations have normal frequency curves and the populations have equal variances (or spreads). You can test the normality assumption using multiple Wilk-Shapiro tests (i.e. one for each sample). In addition, you can save (to the worksheet) the **centered** data values, which are the observations minus the mean for the group from which each observation comes. These centered values, or **residuals**, should behave as a single sample from a normal population. A boxplot and normal quantile test of the residuals gives an overall assessment of normality. The commands `predict residuals, resid` and then `swilk residuals` indicates that, although not significant at the 5% level, normality may be suspect:

```
. swilk residuals
                   Shapiro-Wilk W test for normal data
    Variable |    Obs       W           V         z     Prob>z
-------------+--------------------------------------------------
   residuals |    680    0.99580      1.866     1.520  0.06425
```

Mathematically, this is just a specialized regression problem and we can construct the same diagnostic plots we have been doing for regression. Cook's D is not worth doing in this case, though.



There are several alternative procedures that can be used when either the normality or equal variance assumption are not satisfied. Welch's ANOVA method (available in JMP-In, not directly available in Stata) is appropriate for normal populations with unequal variances. The test is a generalization of Satterthwaite's two-sample test discussed last semester. Most statisticians probably would use weighted least squares or transformations to deal with the unequal variance problem (we will discuss this if time permits this semester). The Wilcoxon or Kruskal-Wallis non-parametric ANOVA is appropriate with non-normal populations with similar spreads.

For the birth weight data, recall that formal tests of equal variances are not significant (p-values $> .4$). Thus, there is insufficient evidence that the population variances differ. Given that the distributions are fairly symmetric, with no extreme values, the standard ANOVA appears to be the method of choice. As an illustration of an alternative method, though, the summary from the Kruskal-Wallis approach follows, leading to the same conclusion as the standard ANOVA. One weakness of **Stata** is that it does not directly provide for non-parametric multiple comparisons. One could do all the pair-wise Mann-Whitney two-sample tests and use a Bonferroni adjustment (the `ranksum` command implements this two sample version of the Kruskal-Wallis test). The Bonferroni adjustment just multiplies all the p-values by 3 (the number of comparisons). If you do this, you find the same conclusions as with the normal-theory procedures: Group 1 differs from the other two, and groups 2 and 3 are not significantly different. Recall from last semester the Kruskal-Wallis and the Mann-Whitney amount to little more than one-way ANOVA and two-sample t-tests, respectively, on ranks in the combined samples (this controls for outliers).

```
. kwallis weight,by(ms_gp_txt)
Test: Equality of populations (Kruskal-Wallis test)
  +-------------------------------------+
  |        ms_gp_txt | Obs |  Rank Sum |
  |------------------+-----+-----------|
  |        Nonsmoker | 381 | 144979.00 |
  | Less than a pack | 169 |  47591.00 |
  |     Pack or more | 130 |  38970.00 |
  +-------------------------------------+
chi-squared =    36.594 with 2 d.f.
probability =     0.0001
chi-squared with ties =    36.637 with 2 d.f.
probability = 0.0001
```

Basics of Experimental Design

This section describes an experimental design to compare the effectiveness of four insecticides to eradicate beetles. The primary interest is determining which treatment is most effective, in the sense of providing the lowest typical survival time.

In a **completely randomized design** (CRD), the scientist might select a sample of genetically identical beetles for the experiment, and then randomly assign a predetermined number of beetles to the treatment groups (insecticides). The sample sizes for the groups need not be equal. A power analysis is often conducted to determine sample sizes for the treatments. For simplicity, assume that 48 beetles will be used in the experiment, with 12 beetles assigned to each group.

After assigning the beetles to the four groups, the insecticide is applied (uniformly to all experimental units or beetles), and the individual survival times recorded. A natural analysis of the data would be to compare the survival times using a one-way ANOVA.

There are several important controls that should be built into this experiment. The same strain of beetles should be used to ensure that the four treatment groups are alike as possible, so that differences in survival times are attributable to the insecticides, and not due to genetic differences among beetles. Other factors that may influence the survival time, say the concentration of the insecticide or the age of the beetles, would be held constant, or fixed by the experimenter, if possible. Thus, the same concentration would be used with the four insecticides.

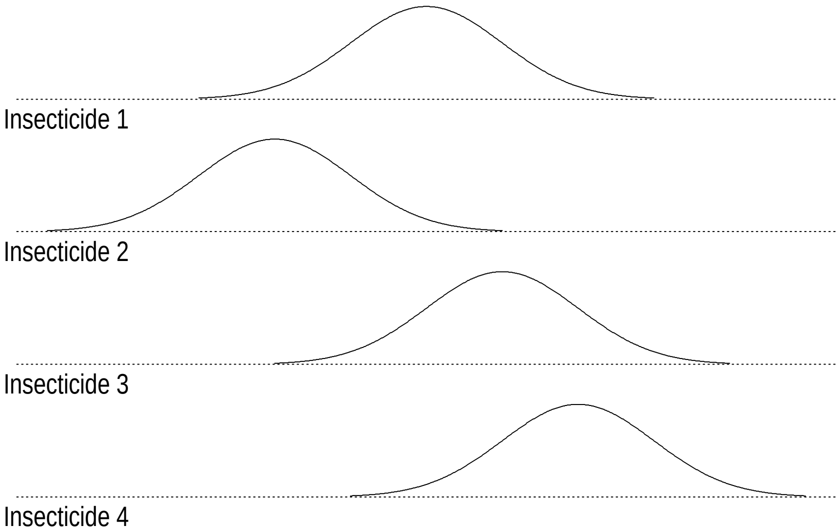
In complex experiments, there are always potential influences that are not realized or thought to be unimportant that you do not or can not control. The **randomization** of beetles to groups ensures that there is no systematic dependence of the observed treatment differences on the uncontrolled influences. This is extremely important in studies where genetic and environmental influences can not be easily controlled (as in humans, more so than in bugs or mice). The randomization of beetles to insecticides tends to diffuse or greatly reduce the effect of the uncontrolled influences on the comparison of insecticides, in the sense that these effects become part of the uncontrolled or error variation of the experiment.

Suppose y_{ij} is the response for the j^{th} experimental unit in the i^{th} treatment group, where $i = 1, 2, ..., I$. The statistical model for a completely randomized **one-factor design** that leads to a one-way ANOVA is given by:

$$y_{ij} = \mu_i + e_{ij},$$

where μ_i is the (unknown) population mean for all potential responses to the i^{th} treatment, and e_{ij} is the residual or deviation of the response from the population mean. The responses within and across treatments are assumed to be independent, normal random variables with constant variance.

For the insecticide experiment, y_{ij} is the survival time for the j^{th} beetle given the i^{th} insecticide, where $i = 1, 2, 3, 4$ and $j = 1, 2, .., 12$. The random selection of beetles coupled with the randomization of beetles to groups ensures the independence assumptions. The assumed population distributions of responses for the $I = 4$ insecticides can be represented as follows:



Let $\mu = \frac{1}{I}\sum_i \mu_i$ be the grand mean, or average of the population means. Let $\alpha_i = \mu_i - \mu$ be the i^{th} **treatment group effect**. The treatment effects add to zero, $\alpha_1 + \alpha_2 + \cdots + \alpha_I = 0$, and measure the difference between the treatment population means and the grand mean. Given this notation, the one-way ANOVA model is

$$y_{ij} = \mu + \alpha_i + e_{ij}.$$

The model specifies that the

$$\text{Response} = \text{Grand Mean} + \text{Treatment Effect} + \text{Residual}.$$

An hypothesis of interest is whether the population means are equal: $H_0 : \mu_1 = \cdots = \mu_I$, which is equivalent to the hypothesis of no treatment effects: $H_0 : \alpha_1 = \cdots = \alpha_I = 0$. If H_0 is true, then the one-way model is

$$y_{ij} = \mu + e_{ij},$$

where μ is the common population mean. We know how to test H_0 and do multiple comparisons of the treatments, so I will skip this material.

Most epidemiological studies are **observational studies** where the groups to be compared ideally consist of individuals that are similar on all characteristics that influence the response, except for the feature that defines the groups. In a designed experiment, the groups to be compared are defined by treatments randomly assigned to individuals. If, in an observational study we can not define the groups to be homogeneous on important factors that might influence the response, then we should adjust for these factors in the analysis. I will discuss this more completely in the next 2 weeks. In the analysis we just did on smoking and birth weight, we were not able to randomize with respect to several factors that might influence the response, and will need to adjust for them.

Paired Experiments and Randomized Block Experiment

A **randomized block design** is often used instead of a completely randomized design in studies where there is extraneous variation among the experimental units that may influence the response. A significant amount of the extraneous variation may be removed from the comparison of treatments by partitioning the experimental units into fairly **homogeneous subgroups** or **blocks**.

For example, suppose you are interested in comparing the effectiveness of four antibiotics for a bacterial infection. The recovery time after administering an antibiotic may be influenced by the patients general health, the extent of their infection, or their age. Randomly allocating experimental subjects to the treatments (and then comparing them using a one-way ANOVA) may produce one treatment having a "favorable" sample of patients with features that naturally lead to a speedy recovery. Additionally, if the characteristics that affect the recovery time are spread across treatments, then the variation within samples due to these uncontrolled features can dominate the effects of the treatment, leading to an inconclusive result.

A better way to design this experiment would be to **block** the subjects into groups of four patients who are alike as possible on factors other than the treatment that influence the recovery time. The four treatments are then randomly assigned to the patients (one per patient) within a block, and the recovery time measured. The blocking of patients usually produces a more sensitive comparison of treatments than does a completely randomized design because the variation in recovery times due to the blocks is eliminated from the comparison of treatments.

A randomized block design is a **paired experiment** when two treatments are compared. The usual analysis for a paired experiment is a parametric or non-parametric paired comparison. In certain experiments, each experimental unit receives each treatment. The experimental units are "natural" blocks for the analysis.

Example: Comparison of Treatments to Relieve Itching

Ten male volunteers between 20 and 30 years old were used as a study group to compare seven treatments (5 drugs, a placebo, and no drug) to relieve itching. Each subject was given a different treatment on seven study days. The time ordering of the treatments was randomized across days. Except on the no-drug day, the subjects were given the treatment intravenously, and then itching was induced on their forearms using an effective itch stimulus called cowage. The subjects recorded the duration of itching, in seconds. The data are given in the table below. From left to right the drugs are: papaverine, morphine, aminophylline, pentobarbitol, tripelenamine.

Patient	Nodrug	Placebo	Papv	Morp	Amino	Pento	Tripel
1	174	263	105	199	141	108	141
2	224	213	103	143	168	341	184
3	260	231	145	113	78	159	125
4	255	291	103	225	164	135	227
5	165	168	144	176	127	239	194
6	237	121	94	144	114	136	155
7	191	137	35	87	96	140	121
8	100	102	133	120	222	134	129
9	115	89	83	100	165	185	79
10	189	433	237	173	168	188	317

The volunteers in the study were treated as blocks in the analysis. At best, the volunteers might be considered a representative sample of males between the ages of 20 and 30. This limits the extent of inferences from the experiment. The scientists can not, without sound medical justification, extrapolate the results to children or to senior citizens.

The Analysis of a Randomized Block Design

Assume that you designed a randomized block experiment with I blocks and J treatments, where each treatment occurs once in each block. Let y_{ij} be the response for the j^{th} treatment within the i^{th} block. The model for the experiment is

$$y_{ij} = \mu_{ij} + e_{ij},$$

where μ_{ij} is the population mean response for the j^{th} treatment in the i^{th} block and e_{ij} is the deviation of the response from the mean. The population means are assumed to satisfy the additive model

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$

where μ is a grand mean, α_i is the effect for the i^{th} block, and β_j is the effect for the j^{th} treatment. The responses are assumed to be independent across blocks, normally distributed and with constant variance. The randomized block model does not require the observations within a block to be independent, but does assume that the correlation between responses within a block is identical for each pair of treatments. This is a reasonable working assumption in many analyses. In this case you really need to be sure the order in which treatments are administered to subjects is randomized in order to assume equal correlation.

The model is sometimes written as

$$\text{Response} = \text{Grand Mean} + \text{Treatment Effect} + \text{Block Effect} + \text{Residual}.$$

Given the data, let $\bar{y}_{i\cdot}$ be the i^{th} block sample mean (the average of the responses in the i^{th} block), $\bar{y}_{\cdot j}$ be the j^{th} treatment sample mean (the average of the responses on the j^{th} treatment), and $\bar{y}_{\cdot\cdot}$ be the average response of all IJ observations in the experiment.

An ANOVA table for the randomized block experiment partitions the Model SS into SS for Blocks and Treatments.

Source	*df*	SS	MS
Blocks	$I-1$	$J\sum_i(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2$	
Treats	$J-1$	$I\sum_j(\bar{y}_{\cdot j}-\bar{y}_{\cdot\cdot})^2$	
Error	$(I-1)(J-1)$	$\sum_{ij}(y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y}_{\cdot\cdot})^2$	
Total	$IJ-1$	$\sum_{ij}(y_{ij}-\bar{y}_{\cdot\cdot})^2.$	

A primary interest is testing whether the treatment effects are zero: $H_0 : \beta_1 = \cdots = \beta_J = 0$. The treatment effects are zero if the population mean responses are identical for each treatment. A formal test of no treatment effects is based on the p-value from the F-statistic $F_{obs} =$ MS Treat/MS Error. The p-value is evaluated in the usual way (i.e. as an upper tail area from an F-distribution with $J-1$ and $(I-1)(J-1)$ df.) This H_0 is rejected when the treatment averages $\bar{y}_{\cdot j}$ vary significantly relative to the error variation.

A test for no block effects ($H_0 : \alpha_1 = \cdots = \alpha_I = 0$) is often a secondary interest, because, if the experiment is designed well, the blocks will be, by construction, noticeably different. There are no block effects if the block population means are identical. A formal test of no block effects is based on the p-value from the F-statistic $F_{obs} =$ MS Blocks/MS Error. This H_0 is rejected when the block averages $\bar{y}_{i\cdot}$ vary significantly relative to the error variation.

A Randomized Block Analysis of the Itching Data

The `anova` command is used to get the randomized block analysis. You will be shown the steps in Thursday's Lab, but I will mention a few important points.

- The data are comprised of three variables: **itchtime**, **person** (ranges from 1-10), and **treatment** (ranges from 1-7). A data file called itch.txt was created with these three variables to be read into **Stata**.

- In the **anova** table, persons play the role of Blocks in this analysis. Using the commands `infile itchtime person treatment using c:/itch.txt` and then `anova itchtime person treatment` we obtain the following output:

```
                           Number of obs =      70     R-squared     =  0.4832
                           Root MSE      = 55.6327     Adj R-squared =  0.3397
                  Source | Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |    156292.6    15  10419.5067       3.37     0.0005
                         |
                  person |  103279.714     9  11475.5238       3.71     0.0011
               treatment |  53012.8857     6  8835.48095       2.85     0.0173
                         |
                Residual |  167129.686    54  3094.99418
              -----------+----------------------------------------------------
                   Total |  323422.286    69   4687.2795
```

- The Model SS is the Sum of the **person** SS and **treatment** SS; check that they add up. The F-test on the Whole-Model test ANOVA checks for whether Treatments or Persons, or both, are significant, i.e. provides an overall test of all effects in the model.

- Next comes the SS for Persons and Treatments, and the corresponding F-statistics and p-values.
- It is possible in Minitab, JMP-IN and SAS (but not directly in **Stata**) to obtain Tukey multiple comparisons of the treatments. These are options in the analysis of the individual effects.
- In **Stata**, we obtain the results of testing differences in the treatments (averaged over persons) using Fisher's method from the `test` command. You will cover this in more detail in Thursday's lab. To obtain Bonferroni's adjusted p-values, simply multiply the p-value for each of Fisher's tests by the number of comparisons you are making; in the itching time example this is $\binom{7}{2} = 21$ paired comparisons. We obtain, for example the results of Fisher's method for comparing treatment 1 with treatment 2 (no drug versus placebo) and treatment 1 with treatment 3 (no drug versus papaverine) with the following commands:

```
test _b[treatment[1]] = _b[treatment[2]]
test _b[treatment[1]] = _b[treatment[3]]
```

We obtain the output:

```
 ( 1)  treatment[1] - treatment[2] = 0
       F(  1,    54) =    0.31
            Prob > F =    0.5814
 ( 1)  treatment[1] - treatment[3] = 0
       F(  1,    54) =    8.56
            Prob > F =    0.0050
```

We see that using Fisher's method, treatments 1 and 2 do not significantly differ, but treatments 1 and 3 do significantly differ at the 5% level. The corresponding Bonferroni p-values are $0.58(2) > 1$ and $0.005(2) = 0.01$ for only the *two* comparisons. They are $0.58(21) > 1$ and $0.005(21) = 0.105$ if *all* 21 paired comparisons are to be made. We would accept that there is no significant difference in mean itching time between either pairs of treatments when all 21 comparisons are to be made. The tabulated p-values resulting from Fisher's method are:

Treatment	1	2	3	4	5	6
2	0.58					
3	0.01	0.00				
4	0.09	0.03	0.23			
5	0.07	0.02	0.30	0.88		
6	0.56	0.26	0.02	0.26	0.20	
7	0.34	0.13	0.05	0.44	0.36	0.71

We have the following groupings:

```
3 5 4 7 6 1 2
-------
  ---------       Fisher's
      -------


-----------
  -----------     Bonferroni's [and Tukey's obtained in JMP-IN]
```

Looking at the means for each treatment averaged over persons, we see that each of the five drugs appears to have an effect, compared to the placebo and to no drug, which have similar means. Papaverine appears to be the most effective drug, whereas placebo is the least effective treatment. A formal F-test shows significant differences among the treatments (p-value=0.017), and among patients (p-value=0.001). The only significant pairwise difference in treatments is between papaverine and placebo using Bonferroni (or Tukey) adjustments.

This all looks more difficult than it needs to be in practice. The usual strategy to start grouping population means this way is first to get the ordering of sample means. Examining the following produces the order `3 5 4 7 6 1 2` above.

```
. mean itchtime,over(treatment) noheader
--------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
itchtime     |
           1 |        191   17.34871      156.3903    225.6097
           2 |      204.8   33.43278      138.1034    271.4966
           3 |      118.2   16.69983      84.88474    151.5153
           4 |        148   14.14763      119.7762    176.2238
           5 |      144.3   13.30585      117.7556    170.8444
           6 |      176.5   21.77422      133.0616    219.9384
           7 |      167.2   21.34521      124.6175    209.7825
--------------------------------------------------------------
```

The target p-value for the Fisher method probably will be .05, and for the Bonferroni method is obtained by simple calculation:

```
. disp .05/21
.00238095
```

We really want to avoid running all 21 tests, and we can skip most. Once the comparisons between 3 and 7 turn out not significant, it is unnecessary to compare 3 to 5 and 4. Once the comparison between 3 and 6 turns out significant, it is not necessary to compare 3 to 1 and 2. Careful examination of patterns can make this fairly quick.

For this particular problem, there are a few outliers and possible problems with the normality assumption. The data set is on the web site - do the residual analysis, and try transforming itchtime with something like the square root to handle outliers a little better. Boxplots can be very valuable here. Redo the comparisons to see if anything has changed in the transformed scale.

A final note: An analysis that ignored person (the blocking factor), i.e. a simple one-way ANOVA, would be incorrect here since it would be assuming all observations are independent. In fact, that analysis finds no differences because the MSE is too large when ignoring blocks (you still should not treat that p-value as valid).