6 One-Way Analysis of Variance

SW 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 samples 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:

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

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

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

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

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

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)

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

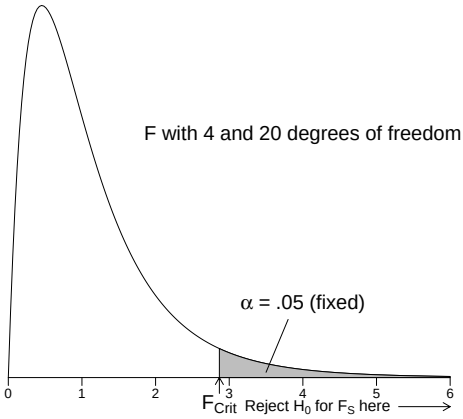
is the variance in the combined data set.

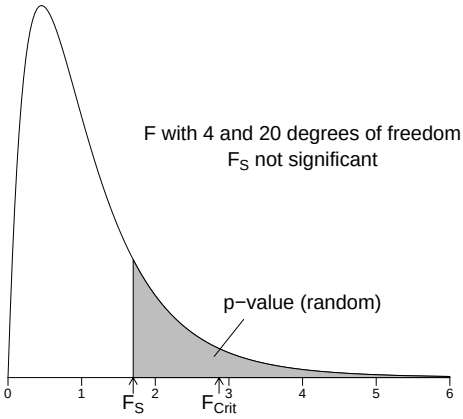
The decision on whether to reject $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ is based on the ratio of the MS(Between) and the MS(Within):

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

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.) The p-value for the test is the area under the $F-$ probability curve to the right of F_s:





For $k = 2$ the ANOVA $F-$ test is equivalent to the pooled two-sample $t-$test.

Minitab summarizes the ANOVA $F-$test with a p-value. The data can be either UNSTACKED or STACKED, but for multiple comparisons discussed later the data must be STACKED. To carry out the analysis, follow the sequence: STAT > ANOVA > ONE-WAY for STACKED data or ONE-WAY (unstacked) for UNSTACKED data. With STACKED data, you need to specify the **response variable** (i.e. the column containing the measurements to be analyzed) and the **factor** (i.e. the column with subscripts that identify the samples) in the dialog box. As with a two-sample analysis, high quality side-by-side boxplots and dotplots can be generated from the ANOVA dialog box. The command line syntax for ANOVA can be obtained from the on-line help, if you are interested.

Example: Comparison of Fats

During cooking, doughnuts absorb fat in various amounts. A scientist wished to learn whether the amount absorbed depends on the type of fat. For each of 4 fats, 6 batches of 24 doughnuts were prepared. The data are grams of fat absorbed per batch (minus 100).

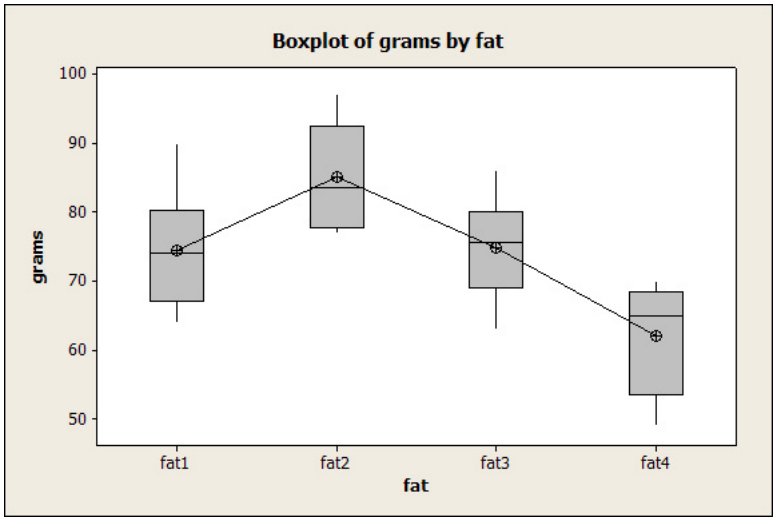
Let

$$\mu_i = \text{pop mean grams of fat } i \text{ absorbed per batch of 24 doughnuts (-100).}$$

The scientist wishes to test $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ against $H_A :$ not H_0. There is no strong evidence against normality here. Furthermore the sample standard deviations (see output below) are close. The standard ANOVA appears to be appropriate here.

The p-value for the $F-$test is .001. The scientist would reject H_0 at any of the usual test levels (i.e. .05 or .01). The data suggest that the population mean absorption rates differ across fats *in some way*. The $F-$test does not say *how* they differ.

```
Row  fat1  fat2  fat3  fat4
  1    64    78    75    55
  2    72    91    86    66
  3    68    97    78    49
  4    77    82    71    64
  5    90    85    63    70
  6    76    77    76    68
```




```
One-way ANOVA: grams versus fat

Source  DF      SS     MS     F      P
fat      3  1595.5  531.8  7.95  0.001
Error   20  1338.3   66.9
Total   23  2933.8

S = 8.180   R-Sq = 54.38%   R-Sq(adj) = 47.54%

                          Individual 95% CIs For Mean Based on
                          Pooled StDev
Level  N    Mean  StDev  -----+---------+---------+---------+----
fat1   6  74.500  9.028              (------*-----)
fat2   6  85.000  7.772                     (------*------)
fat3   6  74.833  7.627              (------*------)
fat4   6  62.000  8.222  (------*------)
                          -----+---------+---------+---------+----
                              60        70        80        90

Pooled StDev = 8.180

Fisher 95% Individual Confidence Intervals     <<<<<<<<<<< WILL EXPLAIN SOON
All Pairwise Comparisons among Levels of fat

Simultaneous confidence level = 80.83%

fat = fat1 subtracted from:

fat     Lower   Center   Upper  ------+---------+---------+---------+---
fat2    0.648   10.500  20.352                  (----*----)
fat3   -9.518    0.333  10.185            (----*----)
fat4  -22.352  -12.500  -2.648       (----*----)
                                 ------+---------+---------+---------+---
                                    -20         0        20        40
```

```
fat = fat2 subtracted from:

fat     Lower   Center    Upper  ------+---------+---------+---------+---
fat3  -20.018  -10.167   -0.315        (----*----)
fat4  -32.852  -23.000  -13.148  (----*---)
                                ------+---------+---------+---------+---
                                   -20         0        20        40

fat = fat3 subtracted from:

fat     Lower   Center   Upper  ------+---------+---------+---------+---
fat4  -22.685  -12.833  -2.982       (----*----)
                                ------+---------+---------+---------+---
                                    -20         0        20        40



Fisher 99.167% Individual Confidence Intervals  <<<<<<<<-- Bonferroni comparisons
All Pairwise Comparisons among Levels of fat

Simultaneous confidence level = 96.16%

fat = fat1 subtracted from:

fat     Lower   Center   Upper  --------+---------+---------+---------+-
fat2   -3.325   10.500  24.325                (------*------)
fat3  -13.492    0.333  14.159           (------*------)
fat4  -26.325  -12.500   1.325      (------*------)
                                --------+---------+---------+---------+-
                                      -20         0        20        40

fat = fat2 subtracted from:

fat     Lower   Center   Upper  --------+---------+---------+---------+-
fat3  -23.992  -10.167   3.659       (------*------)
fat4  -36.825  -23.000  -9.175  (------*-----)
                                --------+---------+---------+---------+-
                                      -20         0        20        40

fat = fat3 subtracted from:

fat     Lower   Center  Upper  --------+---------+---------+---------+-
fat4  -26.659  -12.833  0.992      (------*-----)
                               --------+---------+---------+---------+-
                                     -20         0        20        40
```

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's, Bonferroni's and Tukey's methods for comparing all pairs of means. These approaches are implemented in **Minitab**.

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).

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. If all the sample sizes are equal $n_1 = n_2 = \cdots = n_k$ then the LSD is the same for each comparison:

$$LSD = t_{crit} s_{pooled}\sqrt{\frac{2}{n_1}},$$

where n_1 is the common sample size.

I will illustrate Fisher's method on the doughnut data, using $\alpha = .05$. At the first step, you reject the hypothesis that the population mean absorptions are equal because $p - value = .001$. At the second step, compare all pairs of fats at the 5% level. Here, $s_{pooled} = 8.18$ and $t_{crit} = 2.086$ for a two-sided test based on 20 *df* (the *df* for Residual SS). Each sample has six observations, so the LSD for each comparison is

$$LSD = 2.086 * 8.18 * \sqrt{\frac{2}{6}} = 9.85.$$

Any two sample means that differ by at least 9.85 in magnitude are **significantly different** at the 5% level.

An easy way to compare all pairs of fats is to order the samples by their sample means. The samples can then be grouped easily, noting that two fats are in the same group if the absolute difference between their sample means is smaller than the LSD.

Fats	Sample Mean
2	85.00
3	74.83
1	74.50
4	62.00

There are six comparisons of two fats. From this table, you can visually assess which sample means differ by at least the LSD=9.85, and which ones do not. For completeness, the table below summarizes each comparison:

Comparison	Absolute difference in means	Exceeds LSD?
Fats 2 and 3	10.17	Yes
2 and 1	10.50	Yes
2 and 4	23.00	Yes
Fats 3 and 1	0.33	No
3 and 4	12.83	Yes
Fats 1 and 4	12.50	Yes

The end product of the multiple comparisons is usually presented as a collection of **groups**, where a group is defined to be a set of populations with sample means that not significantly different from each other. Overlap among groups is common, and occurs when one or more populations appears in two or more groups. Any overlap requires a more careful interpretation of the analysis.

There are three groups for the doughnut data, with no overlap. Fat 2 is in a group by itself, and so is Fat 4. Fats 3 and 1 are in a group together. This information can be summarized by ordering the samples from lowest to highest average, and then connecting the fats in the same group using an underscore:

```
FAT 4    FAT 1    FAT 3     FAT 2
-----    ----------------   -----
```

The results of a multiple comparisons must be interpreted carefully. At the 5% level, you have sufficient evidence to conclude that the population mean absorption for Fat 2 exceeds the other population means, whereas the mean absorption for Fat 4 is smallest. However, there is insufficient evidence to conclude that the population mean absorptions for Fats 1 and 3 differ.

Be Careful with Interpreting Groups in Multiple Comparisons!

To see why you must be careful when interpreting groupings, suppose you obtain two groups in a three sample problem. One group has samples 1 and 3. The other group has samples 3 and 2:

```
 1      3      2
---------
       ------------
```

This occurs, for example, when $|\overline{Y}_1 - \overline{Y}_2| \geq LSD$, but both $|\overline{Y}_1 - \overline{Y}_3|$ and $|\overline{Y}_3 - \overline{Y}_2|$ are less than the LSD. There is a tendency to conclude, and please try to avoid this line of attack, that populations 1 and 3 have the same mean, populations 2 and 3 have the same mean, but populations 1 and 2 have different means. This conclusion is illogical. The groupings imply that we have sufficient evidence to conclude that population means 1 and 2 are different, but insufficient evidence to conclude that population mean 3 differs from either of the other population means.

FSD Multiple Comparisons in Minitab

To get Fisher comparisons in **Minitab**, check on COMPARISONS in the one-way ANOVA dialog box. Then choose Fisher, with individual error rate = 5 to get the individual comparisons at the 5% level, as considered above. One slight difficulty relative to our presentation is that **Minitab** summarizes the multiple comparisons in terms of all possible 95% CIs for differences in population means. This output can be used to generate groupings by noting that the individual CIs will cover zero if and only if the corresponding 5% tests of equal means is not significant. Thus a CI for the difference in the population means that covers zero implies that the two populations are in the same group. A summary of the CIs is given below; see the earlier output. Let us see that we can recover the groups from this output.

95% CI for	Limits
$\mu_2 - \mu_1$	0.65 to 20.35
$\mu_3 - \mu_1$	-9.52 to 10.19
$\mu_3 - \mu_1$	-22.35 to -2.65
$\mu_3 - \mu_2$	-20.02 to -0.32
$\mu_4 - \mu_2$	-32.85 to -13.15
$\mu_4 - \mu_3$	-22.69 to -2.98

Discussion of the FSD Method

There are $c = .5k(k-1)$ 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** by SAS and the **individual error rate** by **Minitab**.

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 or 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

$$\alpha < FER < c\alpha.$$

For example, in the doughnut problem where $k = 4$, there are $c = .5 * 4 * 3 = 6$ possible comparisons of pairs of fats. If each comparison is carried out at the 5% level, then $.05 < FER < .30$. At the second step of the FSD method, you could have up to a 30% chance of claiming one or more pairs of population means are different if no differences existed between population means. **Minitab** gives the actual FER for this problem as .192. SAS and most other statistical packages do not evaluate the exact FER, so the upper bound is used.

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.

Bonferroni Comparisons

The Bonferroni method controls the FER by reducing the individual comparison error rate. The FER is guaranteed to be no larger than a prespecified amount, say α, by setting the individual error rate for each of the c comparisons of interest to α/c. Larger differences in the sample means are needed before declaring statistical significance using the Bonferroni adjustment than when using the FSD method at the α level.

Assuming all comparisons are of interest, you can implement the Bonferroni adjustment in **Minitab** by specifying the Fisher comparisons with the appropriate **individual error rate**. **Minitab** gives the actual FER, and $100(1-\alpha/c)\%$ CI for all pairs of means $\mu_i - \mu_j$. A by-product of the Bonferroni adjustment is that we have at least $100(1-\alpha)\%$ confidence that all CI statements hold simultaneously!

If you wish to guarantee a $FER \leq .05$ on all six comparisons in the doughnut problem, then set the individual error rate to $.05/6 = .0083$. **Minitab** gives $100(1-.0083)\% = 99.17\%$ CIs for all $\mu_i - \mu_j$, and computes the actual FER. Here FER=.0382. The Bonferroni output was given earlier. Looking at the output, can you create the groups? You should get the groups given below, which implies you have sufficient evidence to conclude that the population mean absorption for Fat 2 exceeds that for Fat 4.

```
FAT 4    FAT 1    FAT 3    FAT 2
----------------------
         ------------------------
```

The Bonferroni method tends to produce "coarser" groups than the FSD method, because the individual comparisons are conducted at a lower level. Equivalently, the minimum significant difference is inflated for the Bonferroni method. For example, in the doughnut problem with $FER \leq .05$, the critical value for the individual comparisons at the .0083 level is $t_{crit} = 2.929$. You can read this off the **Minitab** output or estimate it from a $t-$table with $df = 20$. The minimum significant difference for the Bonferroni comparisons is

$$LSD = 2.929 * 8.18 * \sqrt{\frac{2}{6}} = 13.824$$

versus an LSD=9.85 for the FSD method. Referring back to our table of sample means on page 71, we see that the sole comparison where the absolute difference between sample means exceeds 13.824 involves Fats 2 and 4.

Example from Koopmans: Facial Tissue Thickness

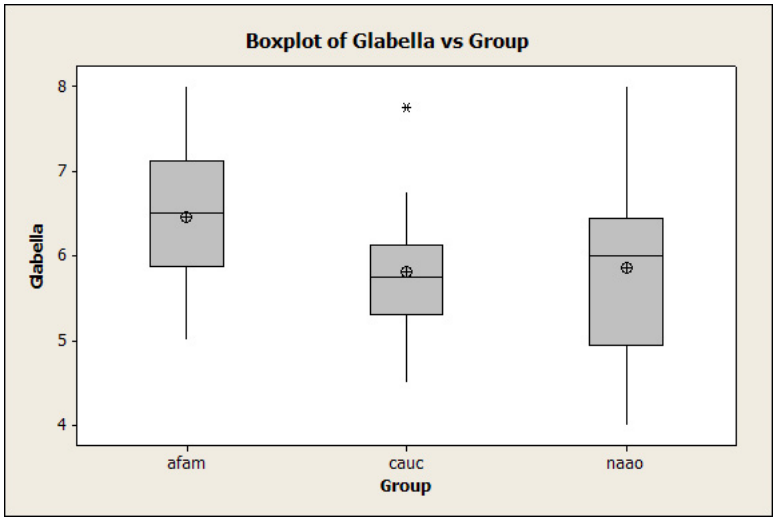
In an anthropological study of facial tissue thickness for different racial groups, data were taken during autopsy at several points on the faces of deceased individuals. The Glabella measurements taken at the bony ridge for samples of individuals from three racial groups (cauc = Caucasian, afam = African American and naao = Native American and Oriental) follow. The data values are in mm.

There are 3 groups, so there are 3 possible pairwise comparisons. If you want a Bonferroni analysis with FER of no greater than .05, you should do the individual comparisons at the $.05/3 = .0167$ level. **Minitab** output is given below. Except for the mild outlier in the Caucasian sample, the observed distributions are fairly normal, with similar spreads. I would expect the standard ANOVA to perform well here.

Let μ_c = population mean Glabella measurement for Caucasians, μ_a = population mean Glabella measurement for African Americans, and μ_n = population mean Glabella measurement for Native Americans and Orientals. At the 5% level, you would not reject the hypothesis that the population mean Glabella measurements are identical. That is, you do not have sufficient evidence to conclude that these racial groups differ with respect to their average Glabella measurement.

The Bonferroni intervals reinforce this conclusion, since each interval for a difference in population means contains zero. You can think of the Bonferroni intervals as simultaneous CI. We're (at least) 95% confident that all of the following statements hold simultaneously: $-1.62 \leq \mu_c - \mu_a \leq .32$, $-.91 \leq \mu_n - \mu_c \leq 1.00$, and $-1.54 \leq \mu_n - \mu_a \leq .33$. The individual CI have level $100(1 - .0167)\% = 98.33\%$. Any further comments?

```
 CONTENTS OF WORKSHEET: Data in Columns c1-c3, labeled

Row  cauc  afam  naao
  1  5.75  6.00  8.00
  2  5.50  6.25  7.00
  3  6.75  6.75  6.00
  4  5.75  7.00  6.25
  5  5.00  7.25  5.50
  6  5.75  6.75  4.00
  7  5.75  8.00  5.00
  8  7.75  6.50  6.00
  9  5.75  7.50  7.25
 10  5.25  6.25  6.00
 11  4.50  5.00  6.00
 12  6.25  5.75  4.25
 13        5.00  4.75
 14              6.00

 Descriptive Statistics: cauc, afam, naao

 Variable   N  N*   Mean  SE Mean  StDev  Minimum     Q1  Median     Q3  Maximum
 cauc      12   0  5.813    0.241  0.833    4.500  5.313   5.750  6.125    7.750
 afam      13   0  6.462    0.248  0.895    5.000  5.875   6.500  7.125    8.000
 naao      14   0  5.857    0.298  1.117    4.000  4.938   6.000  6.438    8.000
```




```
One-way ANOVA: Glabella versus Group

Source  DF      SS     MS     F      P
Group    2   3.398  1.699  1.83  0.175
Error   36  33.461  0.929
Total   38  36.859

S = 0.9641   R-Sq = 9.22%   R-Sq(adj) = 4.18%

                          Individual 95% CIs For Mean Based on
                          Pooled StDev
Level   N    Mean   StDev  -----+---------+---------+---------+----
afam   13  6.4615  0.8947              (----------*----------)
cauc   12  5.8125  0.8334  (----------*----------)
naao   14  5.8571  1.1168    (---------*---------)
                          -----+---------+---------+---------+----
                             5.50      6.00      6.50      7.00

Pooled StDev = 0.9641


Fisher 98.33% Individual Confidence Intervals
All Pairwise Comparisons among Levels of Group

Simultaneous confidence level = 95.69%


Group = afam subtracted from:

Group    Lower   Center   Upper  ---+---------+---------+---------+------
cauc   -1.6178  -0.6490  0.3198  (-------------*-------------)
naao   -1.5365  -0.6044  0.3277   (------------*-------------)
                                 ---+---------+---------+---------+------
                                -1.40     -0.70      0.00      0.70


Group = cauc subtracted from:

Group    Lower  Center   Upper  ---+---------+---------+---------+------
naao   -0.9074  0.0446  0.9967           (-------------*------------)
                                 ---+---------+---------+---------+------
                                -1.40     -0.70      0.00      0.70
```

Further Discussion of Multiple Comparisons

The FSD and Bonferroni methods comprise the ends of the spectrum of multiple comparisons methods. Among multiple comparisons procedures, the FSD method is most likely to find differences, whether real or due to sampling variation, whereas Bonferroni is often the most conservative method. You can be reasonably sure that differences suggested by the Bonferroni method will be suggested by almost all other methods, whereas differences not significant under FSD will not be picked up using other approaches.

The Bonferroni method is conservative, but tends to work well when the number of comparisons is small, say 4 or less. A smart way to use the Bonferroni adjustment is to focus attention only on the comparisons of interest (generated independently of looking at the data!), and ignore the rest. I will return to this point later.

Two commonly used alternatives to FSD and Bonferroni are **Tukey's** honest significant difference method (HSD) and Newman-Keuls studentized range method. Tukey's method can be implemented in **Minitab** by specifying Tukey multiple comparisons (typically with FER=5%) in the one-way ANOVA dialog box. SW discuss the Newman-Keuls approach, which is not implemented in **Minitab**.

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. For the doughnut fats, the groupings based on Tukey and Bonferroni comparisons are identical; see the **Minitab** output below.

```
Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons among Levels of fat

Individual confidence level = 98.89%

fat = fat1 subtracted from:

fat     Lower   Center   Upper  --------+---------+---------+---------+-
fat2   -2.725   10.500  23.725                  (-----*------)
fat3  -12.891    0.333  13.558            (-----*------)
fat4  -25.725  -12.500   0.725       (------*-----)
                                --------+---------+---------+---------+-
                                      -20         0        20        40

fat = fat2 subtracted from:

fat     Lower   Center   Upper  --------+---------+---------+---------+-
fat3  -23.391  -10.167   3.058        (------*------)
fat4  -36.225  -23.000  -9.775  (------*-----)
                                --------+---------+---------+---------+-
                                      -20         0        20        40

fat = fat3 subtracted from:

fat     Lower   Center  Upper  --------+---------+---------+---------+-
fat4  -26.058  -12.833  0.391       (------*-----)
                              --------+---------+---------+---------+-
                                    -20         0        20        40
```

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 normal scores tests, which we discussed earlier. An alternative approach that is useful with three or more samples is to make a single normal scores plot for the entire data set. The samples must be centered at the same location for this to be meaningful. (WHY?) This is done by subtracting the sample mean from each observation in the sample, giving the so-called **residuals**. A normal scores plot or histogram of the residuals should resemble a sample from a normal population. These two plots can be generated with the ANOVA procedure in **Minitab**, but the normal probability plot does not include a p-value for testing normality. However, the residuals can be stored in the worksheet, and then a formal test of normality is straightforward.

In a previous lecture, I illustrated the use of Bartlett's test and Levene's test for equal population variances, and showed how to evaluate these tests in **Minitab**. I will now define **Bartlett's test**, which assumes normally distributed data. As above, let $n^* = n_1 + n_2 + \cdots + n_k$, where the n_is are the sample sizes from the k groups, and define

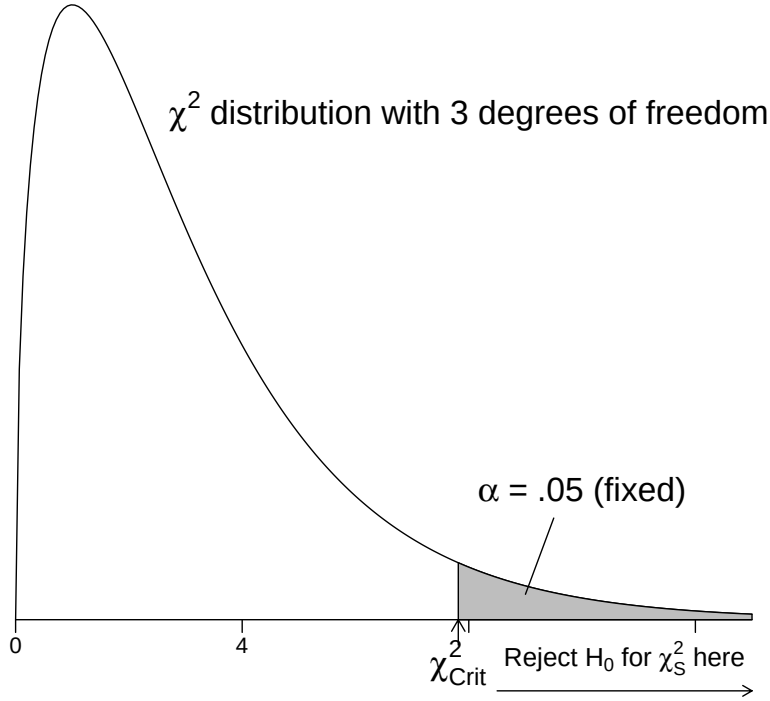
$$v = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{n^* - k}\right).$$

Bartlett's statistic for testing $H_0 : \sigma_1^2 = \cdots = \sigma_k^2$ is given by

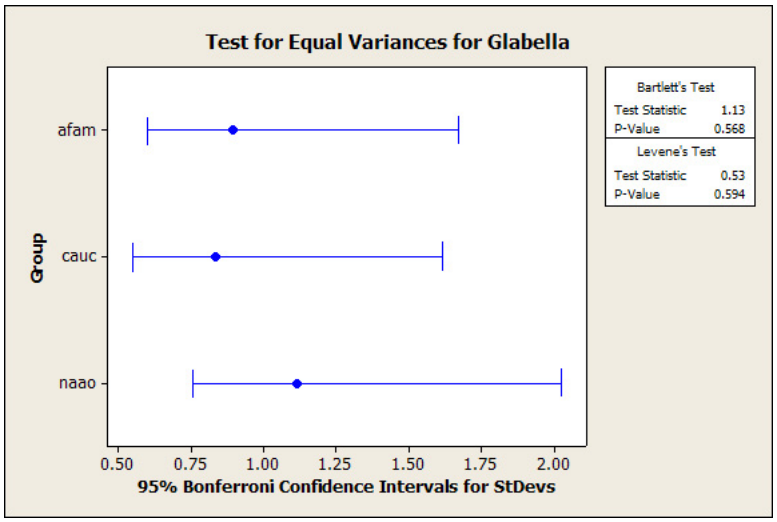
$$B_{obs} = \frac{2.303}{v}\left\{(n-k)\log s^2_{pooled} - \sum_{i=1}^{k}(n_i - 1)\log s_i^2\right\},$$

where s^2_{pooled} is the pooled estimator of variance and s_i^2 is the estimated variance based on the i^{th} sample.

Large values of B_{obs} suggest that the population variances are unequal. For a size α test, we reject H_0 if $B_{obs} \geq \chi^2_{k-1,crit}$, where $\chi^2_{k-1,crit}$ is the upper-α percentile for the χ^2_{k-1} (chi-squared) probability distribution with $k-1$ degrees of freedom. A generic plot of the χ^2 distribution is given below. SW give a chi-squared table on p. 653. A p-value for the test is given by the area under the chi-squared curve to the right of B_{obs}.



Minitab does the calculation for us, as illustrated below. Follow the menu path `Stat > ANOVA > Test for equal variances`. This result is not surprising given how close the sample variances are to each other.



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.

Several plots were generated as part of the analysis: dotplots and boxplots, normal probability plots for each sample, and a normal probability plot and histogram of the residuals from the ANOVA. These are included at the end of the notes.

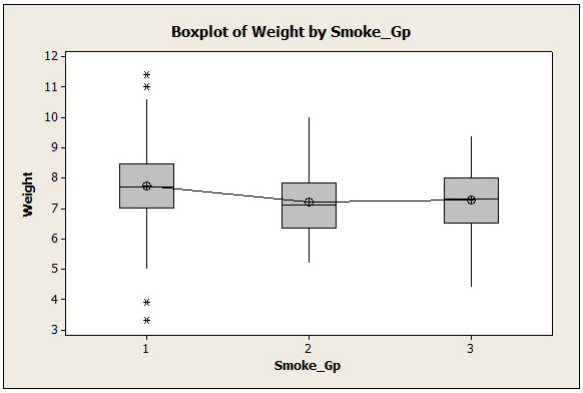
Looking at the boxplots, there is some evidence of non-normality here. Although there are outliers in the no smoking group, we need to recognize that the sample size for this group is fairly large - 381. Given that boxplots are calibrated in such a way that 7 outliers per 1000 observations are expected when sampling from a normal population, 5 outliers (you only see 4!) out of 381 seems a bit excessive. A formal test rejects the hypothesis of normality in the no and low smoker groups. The normal probability plot and the histogram of the residuals also suggests that the population distributions are heavy tailed. I also saved the residuals from the ANOVA and did a formal test of normality on the combined sample, which was significant (p-value=.029). However, I am not overly concerned about this for the following reasons - in large samples, small deviations from normality are often statistically significant and in my experience, the small deviations we are seeing here are not likely to impact our conclusions, in the sense that non-parametric methods that do not require normality will lead to the same conclusions.

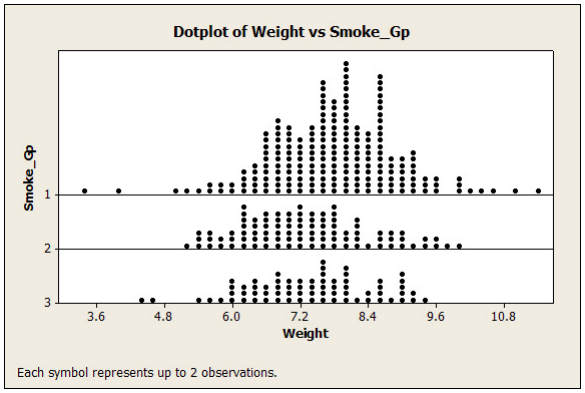
Looking at the summaries, we see that the sample standard deviations are close. Formal tests of equal population variances are far from significant. The p-values for Bartlett's test and Levene's test are greater than .4. Thus, the standard ANOVA appears to be appropriate here.

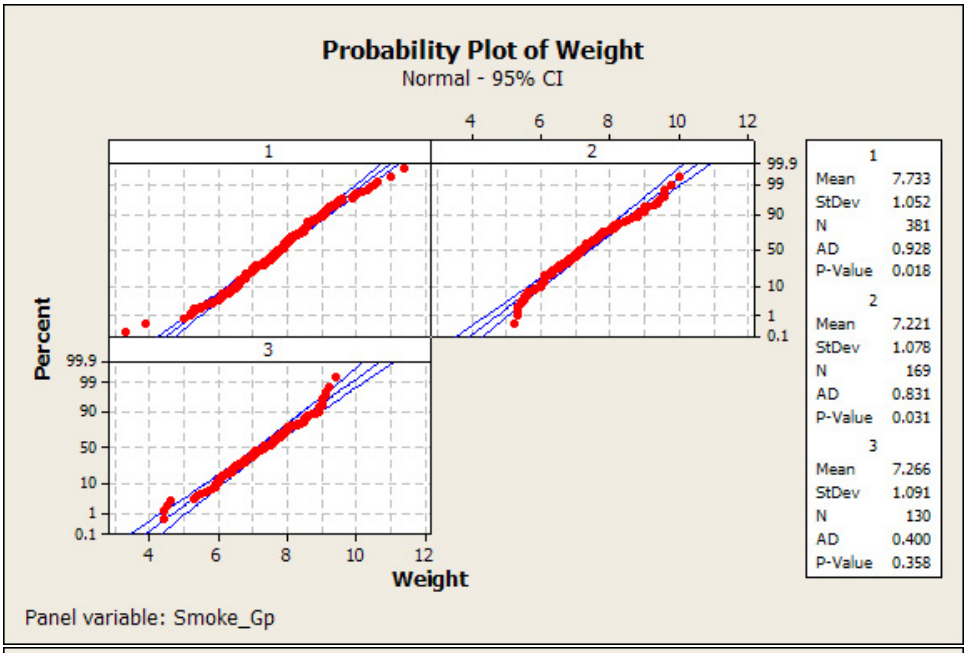
The p-value for the $F-$test is less than .0001. We would reject H_0 at any of the usual test levels (i.e. .05 or .01). The data suggest that the population mean birth weights differ across smoking status groups. The Tukey multiple comparisons suggest that the mean birth weights are higher for children born to mothers that did not smoke during pregnancy.

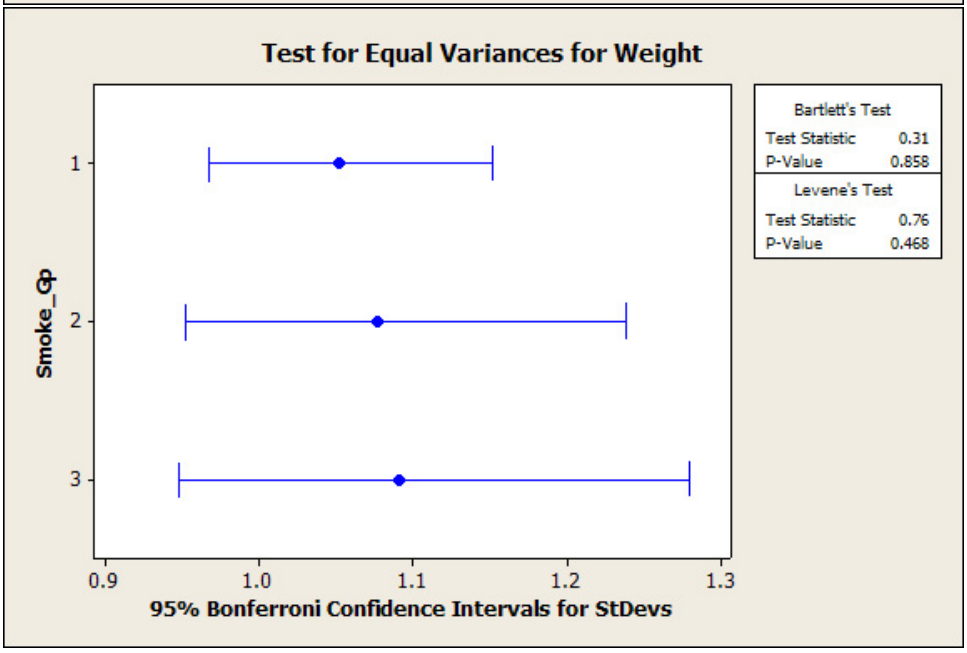
```
Descriptive Statistics: Weight

Variable  Smoke_Gp    N  N*    Mean  SE Mean   StDev  Minimum      Q1  Median
Weight    1         381   0  7.7328   0.0539  1.0523   3.3000  7.0000  7.7000
          2         169   0  7.2213   0.0829  1.0778   5.2000  6.3500  7.1000
          3         130   0  7.2662   0.0957  1.0909   4.4000  6.5000  7.3000

Variable  Smoke_Gp      Q3  Maximum
Weight    1         8.4500  11.4000
          2         7.8500  10.0000
          3         8.0000   9.4000

One-way ANOVA: Weight versus Smoke_Gp

Source     DF      SS     MS      F      P
Smoke_Gp    2   40.70  20.35  17.90  0.000
```

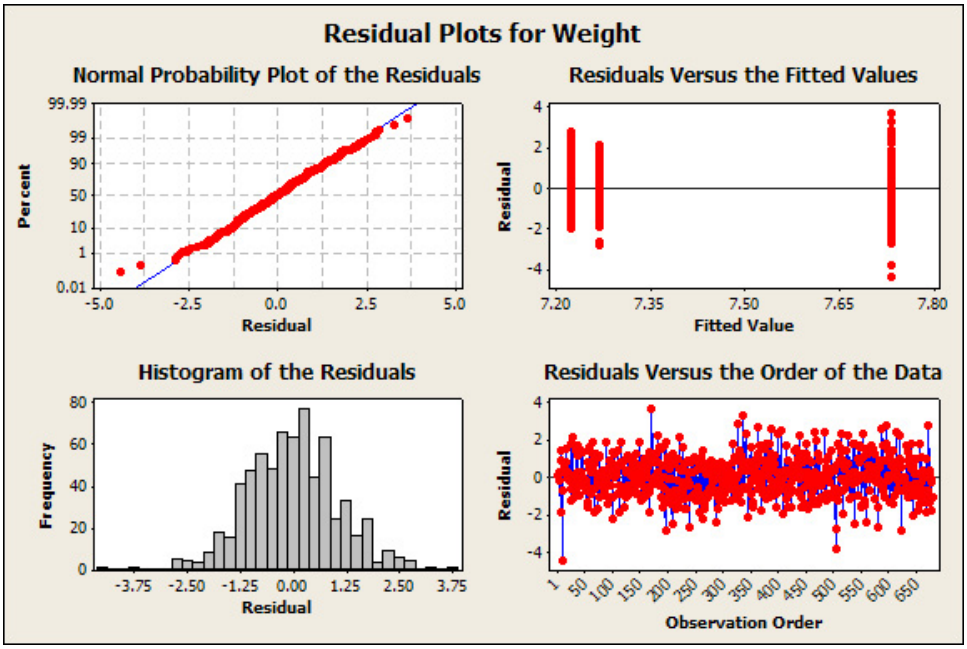
```
Error     677  769.49   1.14
Total     679  810.20

S = 1.066   R-Sq = 5.02%   R-Sq(adj) = 4.74%

                                Individual 95% CIs For Mean Based on
                                Pooled StDev
Level    N   Mean  StDev  -------+---------+---------+---------+--
1      381  7.733  1.052                              (-----*----)
2      169  7.221  1.078  (-------*-------)
3      130  7.266  1.091   (--------*--------)
                           -------+---------+---------+---------+--
                                7.20      7.40      7.60      7.80

Pooled StDev = 1.066

Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons among Levels of Smoke_Gp

Individual confidence level = 98.05%

Smoke_Gp = 1 subtracted from:

Smoke_Gp   Lower  Center   Upper  -----+---------+---------+---------+----
2         -0.742  -0.512  -0.281  (-------*-------)
3         -0.720  -0.467  -0.213   (--------*--------)
                                 -----+---------+---------+---------+----
                                   -0.60     -0.30      0.00      0.30

Smoke_Gp = 2 subtracted from:

Smoke_Gp   Lower  Center  Upper  -----+---------+---------+---------+----
3         -0.246   0.045  0.336                  (--------*---------)
                                 -----+---------+---------+---------+----
                                   -0.60     -0.30      0.00      0.30
```






Probability Plot of Weight
Normal - 95% CI
Percent
Weight
Panel variable: Smoke_Gp
1
Mean 7.733
StDev 1.052
N 381
AD 0.928
P-Value 0.018
2
Mean 7.221
StDev 1.078
N 169
AD 0.831
P-Value 0.031
3
Mean 7.266
StDev 1.091
N 130
AD 0.400
P-Value 0.358


Test for Equal Variances for Weight
Smoke_Gp
95% Bonferroni Confidence Intervals for StDevs
Bartlett's Test
Test Statistic 0.31
P-Value 0.858
Levene's Test
Test Statistic 0.76
P-Value 0.468


Residual Plots for Weight
Normal Probability Plot of the Residuals
Percent
99.99
99
90
50
10
1
0.01
-5.0
-2.5
0.0
2.5
5.0
Residual
Residuals Versus the Fitted Values
Residual
4
2
0
-2
-4
7.20
7.35
7.50
7.65
7.80
Fitted Value
Histogram of the Residuals
Frequency
80
60
40
20
0
-3.75
-2.50
-1.25
0.00
1.25
2.50
3.75
Residual
Residuals Versus the Order of the Data
Residual
4
2
0
-2
-4
1
50
100
150
200
250
300
350
400
450
500
550
600
650
Observation Order