Chapter 5

One-Way Analysis of Variance

Learning objectives

After completing this topic, you should be able to:

select graphical displays that meaningfully compare independent populations.

assess the assumptions of the ANOVA visually and by formal tests.

decide whether the means between populations are different, and how.

Achieving these goals contributes to mastery in these course learning outcomes:

1. organize knowledge.

5. define parameters of interest and hypotheses in words and notation.

6. summarize data visually, numerically, and descriptively.

8. use statistical software.

12. make evidence-based decisions.

5.1 ANOVA

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	$\bar{Y}_1$	$\bar{Y}_2$	$\cdots$	$\bar{Y}_k$
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 $\bar{\bar{Y}}$ be the average response over all samples (combined), that is

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

Note that $\bar{\bar{Y}}$ 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 (that is, at least two means are different). The assumptions needed for the standard ANOVA F-test are analogous to the independent pooled 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:
SS(Within) $= (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2 = \sum_{ij}(Y_{ij} - \bar{Y}_i)^2$.
The **Between SS**, often called the Model SS, measures the spread between the sample means

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

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

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

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	F
Between Groups (Model)	$dfM = k-1$	$SSM = \sum_i n_i(\bar{Y}_i - \bar{\bar{Y}})^2$	$MSM = SSM/dfM$	MSM/MSE
Within Groups (Error)	$dfE = n^* - k$	$SSE = \sum_i (n_i - 1)s_i^2$	$MSE = SSE/dfE$	
Total	$dfT = n^* - 1$	$SST = \sum_{ij}(Y_{ij} - \bar{\bar{Y}})^2$	$MST = SST/dfT$	

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)

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

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)

$$\text{MS(Between)} = \frac{\sum_i n_i(\bar{Y}_i - \bar{\bar{Y}})^2}{k-1}$$

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

The MS(Total)

$$\text{MS(Total)} = \frac{\sum_{ij}(Y_{ij} - \bar{\bar{Y}})^2}{n^* - 1}$$

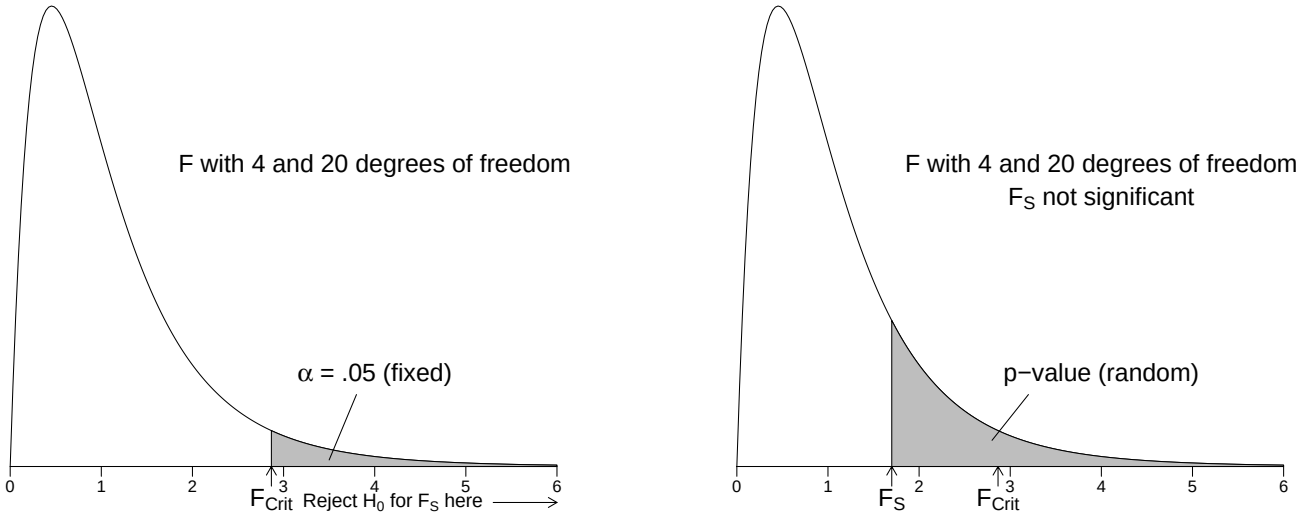
is the variance in the combined data set.

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):

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

Large values of F_s indicate large variability among the sample means $\bar{Y}_1,\ \bar{Y}_2, \ldots, \bar{Y}_k$ 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.

The specification of a one-way analysis of variance is analogous to a regression analysis (discussed later, though we've seen the specification in some plotting functions). The only difference is that the descriptive (x) variable

needs to be a factor and not a numeric variable. We calculate a model object using `lm()` and extract the analysis of variance table with `anova()`.

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.

Let

μ_i = pop mean grams of fat i 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.

Row	fat1	fat2	fat3	fat4
1	164	178	175	155
2	172	191	186	166
3	168	197	178	149
4	177	182	171	164
5	190	185	163	170
6	176	177	176	168

Let's take a short detour to read the wide table and convert it into long format.

R skills: wide to long table format Many functions in R expect data to be in a long format rather than a wide format. Let's use the fat data as an example of how to read a table as text, convert the wide format to long, and then back to wide format.

```
#### Example: Comparison of Fats
fat <- read.table(text="
Row  fat1  fat2  fat3  fat4
  1   164   178   175   155
  2   172   191   186   166
  3   168   197   178   149
  4   177   182   171   164
  5   190   185   163   170
  6   176   177   176   168
", header=TRUE)
fat
```

```
##   Row fat1 fat2 fat3 fat4
## 1   1  164  178  175  155
## 2   2  172  191  186  166
## 3   3  168  197  178  149
## 4   4  177  182  171  164
## 5   5  190  185  163  170
## 6   6  176  177  176  168
```

From wide to long: Use `melt()` from the `reshape2` package.

```
#### From wide to long format
library(reshape2)
fat.long <- melt(fat,
              # id.vars: ID variables
              #   all variables to keep but not split apart on
              id.vars=c("Row"),
              # measure.vars: The source columns
              #   (if unspecified then all other variables are measure.vars)
              measure.vars = c("fat1", "fat2", "fat3", "fat4"),
              # variable.name: Name of the destination column identifying each
              #   original column that the measurement came from
              variable.name = "type",
              # value.name: column name for values in table
              value.name = "amount"
            )
## naming variables manually, the variable.name and value.name not working 11/2012
#names(fat.long) <- c("Row", "type", "amount")
fat.long
```

```
##    Row type amount
## 1    1 fat1    164
## 2    2 fat1    172
## 3    3 fat1    168
## 4    4 fat1    177
## 5    5 fat1    190
## 6    6 fat1    176
## 7    1 fat2    178
## 8    2 fat2    191
## 9    3 fat2    197
## 10   4 fat2    182
## 11   5 fat2    185
## 12   6 fat2    177
## 13   1 fat3    175
## 14   2 fat3    186
## 15   3 fat3    178
## 16   4 fat3    171
## 17   5 fat3    163
## 18   6 fat3    176
## 19   1 fat4    155
## 20   2 fat4    166
## 21   3 fat4    149
```

```
## 22   4 fat4    164
## 23   5 fat4    170
## 24   6 fat4    168
```

```
# or as simple as:
# melt(fat, "Row")
```

If you don't specify variable.name, it will name that column "variable", and if you leave out value.name, it will name that column "value".

From long to wide: Use `dcast()` from the `reshape2` package.

```
#### From long to wide format
fat.wide <- dcast(fat.long, Row ~ type, value.var = "amount")
fat.wide
```

```
##   Row fat1 fat2 fat3 fat4
## 1   1  164  178  175  155
## 2   2  172  191  186  166
## 3   3  168  197  178  149
## 4   4  177  182  171  164
## 5   5  190  185  163  170
## 6   6  176  177  176  168
```

Now that we've got our data in the long format, let's return to the ANOVA.

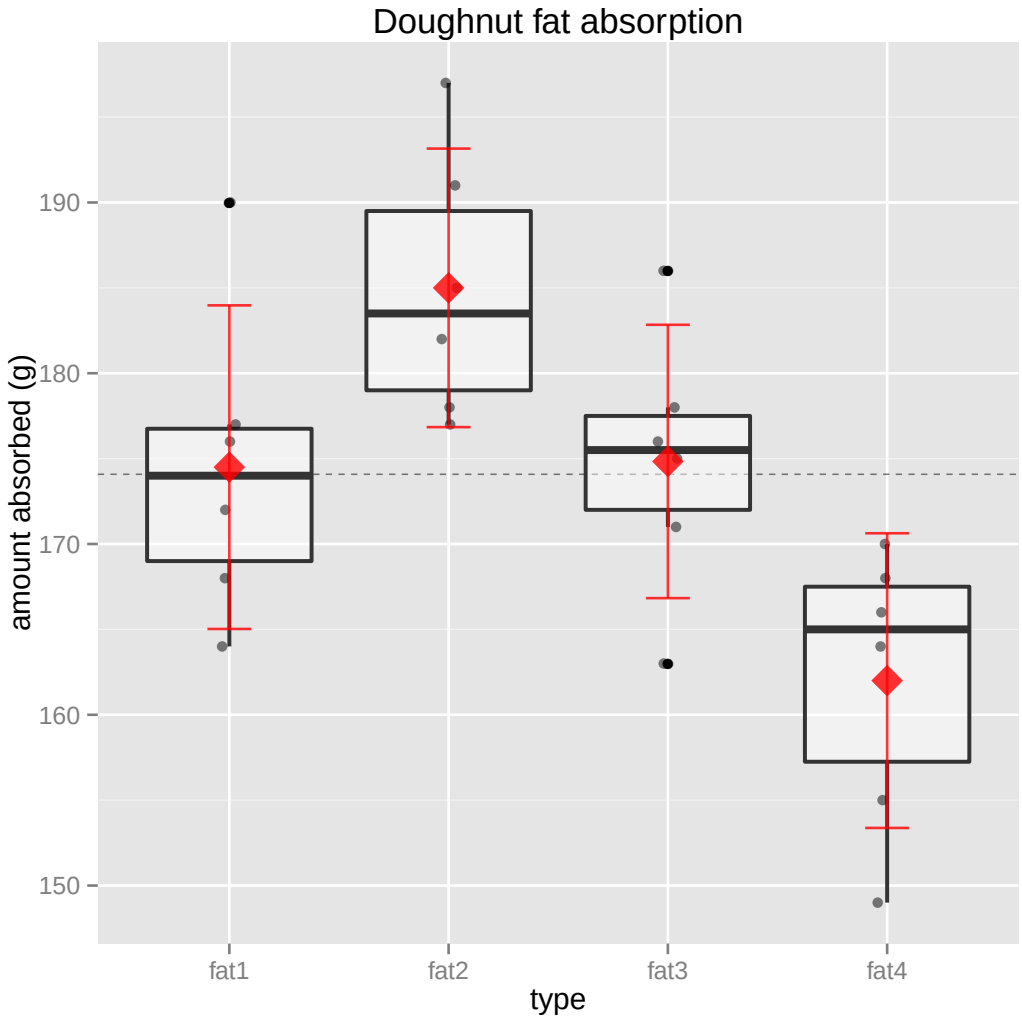
Back to ANOVA: Let's look at the numerical summaries. We've seen other ways of computing these so I'll show you another way.

```
#### Back to ANOVA
# Calculate the mean, sd, n, and se for the four fats

# The plyr package is an advanced way to apply a function to subsets of data
#   "Tools for splitting, applying and combining data"
library(plyr)
# ddply "dd" means the input and output are both data.frames
fat.summary <- ddply(fat.long,
                     "type",
                     function(X) {
                       data.frame( m = mean(X$amount),
                                   s = sd(X$amount),
                                   n = length(X$amount)
                                 )
                     }
                    )
# standard errors
fat.summary$se   <- fat.summary$s/sqrt(fat.summary$n)
# individual confidence limits
fat.summary$ci.l <- fat.summary$m - qt(1-.05/2, df=fat.summary$n-1) * fat.summary$se
fat.summary$ci.u <- fat.summary$m + qt(1-.05/2, df=fat.summary$n-1) * fat.summary$se
fat.summary
```

```
##   type        m        s n       se     ci.l     ci.u
## 1 fat1 174.5000 9.027735 6 3.685557 165.0260 183.9740
## 2 fat2 185.0000 7.771744 6 3.172801 176.8441 193.1559
## 3 fat3 174.8333 7.626707 6 3.113590 166.8296 182.8371
## 4 fat4 162.0000 8.221922 6 3.356586 153.3716 170.6284
```

Let's plot the data with boxplots, individual points, mean, and CI by fat type.

```
# Plot the data using ggplot
library(ggplot2)
p <- ggplot(fat.long, aes(x = type, y = amount))
# plot a reference line for the global mean (assuming no groups)
p <- p + geom_hline(yintercept = mean(fat.long$amount),
                    colour = "black", linetype = "dashed", size = 0.3, alpha = 0.5)
# boxplot, size=.75 to stand out behind CI
p <- p + geom_boxplot(size = 0.75, alpha = 0.5)
# points for observed data
p <- p + geom_point(position = position_jitter(w = 0.05, h = 0), alpha = 0.5)
# diamond at mean for each group
p <- p + stat_summary(fun.y = mean, geom = "point", shape = 18, size = 6,
                      colour = "red", alpha = 0.8)
# confidence limits based on normal distribution
p <- p + stat_summary(fun.data = "mean_cl_normal", geom = "errorbar",
                      width = .2, colour = "red", alpha = 0.8)
p <- p + labs(title = "Doughnut fat absorption") + ylab("amount absorbed (g)")
print(p)
```



The p-value for the F-test is 0.001. The scientist would reject H_0 at any of the usual test levels (such as, 0.05 or 0.01). The data suggest that the population mean absorption rates differ across fats *in some way*. The F-test does not say *how* they differ. The pooled standard deviation $s_{\text{pooled}} = 8.18$ is the "Residual standard error". We'll ignore the rest of this output for now.

```
fit.f <- aov(amount ~ type, data = fat.long)
summary(fit.f)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## type         3   1596   531.8   7.948 0.0011 **
## Residuals   20   1338    66.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit.f
```

```
## Call:
##    aov(formula = amount ~ type, data = fat.long)
##
## Terms:
##                     type Residuals
```

```
## Sum of Squares  1595.500  1338.333
## Deg. of Freedom        3        20
##
## Residual standard error: 8.180261
## Estimated effects may be unbalanced
```

■ CLICKER Qs — ANOVA, Fat 1/2 ■

■ CLICKER Qs — ANOVA, Fat 1/2 ■

5.2 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.

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 $df = dfE$ (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

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

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

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

The minimum absolute difference between $\bar{Y}_i$ and $\bar{Y}_j$ 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:

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

where n_1 is the common sample size.

I will illustrate Fisher's method on the doughnut data, using $\alpha = 0.05$. At the first step, you reject the hypothesis that the population mean absorptions are equal because p-value= 0.001. At the second step, compare all pairs of fats at the 5% level. Here, $s_{\text{pooled}} = 8.18$ and $t_{\text{crit}} = 2.086$ for a two-sided test based on 20 *df* (the dfE for Residual SS). Each sample has six observations, so the LSD for each comparison is

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

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	185.00
3	174.83
1	174.50
4	162.00

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 are 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 and Fat 4 are each different than the other population means. 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 $|\bar{Y}_1 - \bar{Y}_2| \geq LSD$, but both $|\bar{Y}_1 - \bar{Y}_3|$ and $|\bar{Y}_3 - \bar{Y}_2|$ 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.

5.2.1 FSD Multiple Comparisons in R

One way to get Fisher comparisons in R uses `pairwise.t.test()` with `p.adjust.method` The resulting summary of the multiple comparisons is in terms of p-values for all pairwise two-sample t-tests using the pooled standard deviation from the ANOVA using `pool.sd = TRUE`. This output can be used to generate groupings. A summary of the p-values is given below. Let us see that we can recover the groups from this output.

```
#### Multiple Comparisons
# all pairwise comparisons among levels of fat
# Fisher's LSD (FSD) uses "none"
pairwise.t.test(fat.long$amount, fat.long$type,
                pool.sd = TRUE, p.adjust.method = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  fat.long$amount and fat.long$type
##
##      fat1  fat2    fat3
## fat2 0.038 -       -
## fat3 0.944 0.044   -
## fat4 0.015 9.3e-05 0.013
##
## P value adjustment method: none
```

Discussion of the FSD Method: family error rate

There are $c = k(k-1)/2$ 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

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

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

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

For example, in the doughnut problem where $k = 4$, there are $c = 4(3)/2 = 6$ possible comparisons of pairs of fats. If each comparison is carried out at the 5% level, then $0.05 < FER < 0.30$. 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. Most statistical packages do not evaluate the 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 (about 45 populations) each at the 5% level.

5.2.2 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. Therefore, $\alpha/c < FER < c\alpha/c = \alpha$, thus the upper bound for FER is α. 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.

■ CLICKER Qs — ANOVA, Bonferroni ■

Assuming all comparisons are of interest, you can implement the Bonferroni adjustment in R by specifying `p.adjust.method = "bonf"` A by-product of the Bonferroni adjustment is that we have at least $100(1-\alpha)\%$ confidence that all pairwise t-test statements hold simultaneously!

```
# Bonferroni 95% Individual p-values
# All Pairwise Comparisons among Levels of fat
pairwise.t.test(fat.long$amount, fat.long$type,
                pool.sd = TRUE, p.adjust.method = "bonf")

##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  fat.long$amount and fat.long$type
##
##      fat1    fat2    fat3
## fat2 0.22733 -       -
## fat3 1.00000 0.26241 -
## fat4 0.09286 0.00056 0.07960
##
## P value adjustment method: bonferroni
```

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 is different than that for Fat 4.

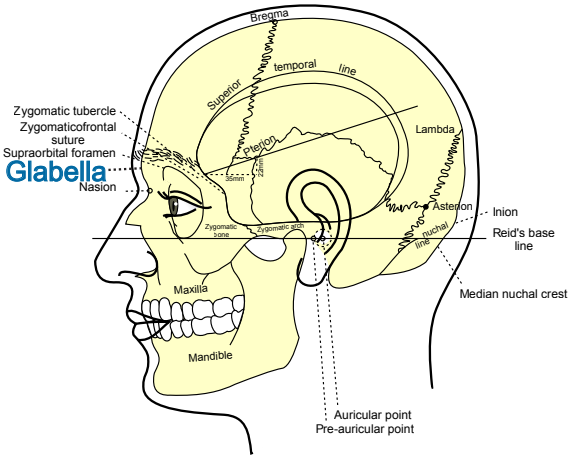
FAT 4 FAT 1 FAT 3 FAT 2

The Bonferroni method tends to produce "coarser" groups than the FSD method, because the individual comparisons are conducted at a lower (alpha/error) level. Equivalently, the minimum significant difference is inflated for the Bonferroni method. For example, in the doughnut problem with $FER \leq 0.05$, the critical value for the individual comparisons at the 0.0083 level is $t_{\text{crit}} = 2.929$ with $df = 20$. The minimum significant difference for the Bonferroni comparisons is

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

versus an LSD=9.85 for the FSD method. Referring back to our table of sample means on page 171, we see that the sole comparison where the absolute difference between sample means exceeds 13.824 involves Fats 2 and 4.

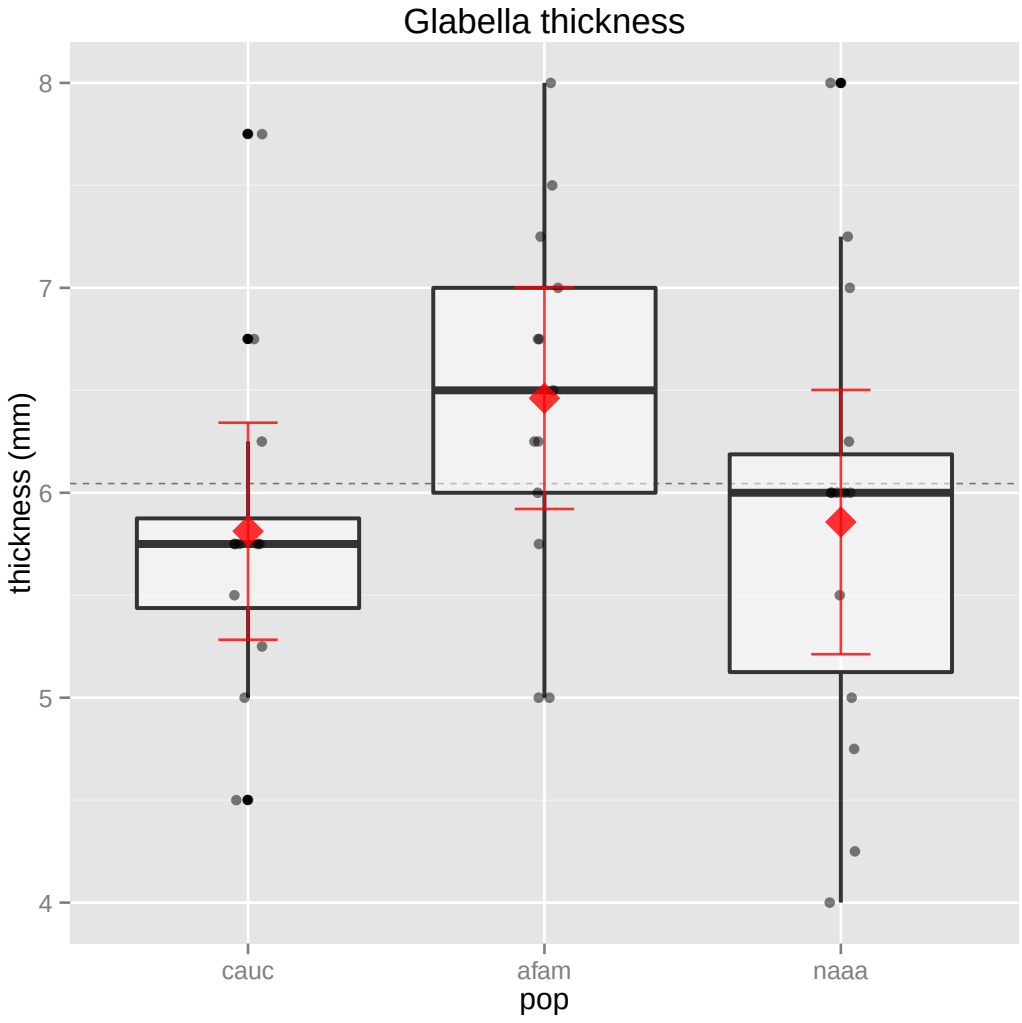
Example from Koopmans: glabella facial tissue thickness In an anthropological study of facial tissue thickness for different racial groups, data were taken during autopsy at several points on the faces of deceased individuals. The Glabella measurements taken at the bony ridge for samples of individuals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




```
#### Example from Koopmans: glabella facial tissue thickness
glabella <- read.table(text="
Row  cauc  afam  naaa
  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    NA  5.00  4.75
 14    NA    NA  6.00
", header=TRUE)

glabella.long <- melt(glabella,
              id.vars=c("Row"),
              variable.name = "pop",
              value.name = "thickness",
              # remove NAs
              na.rm = TRUE
             )
# naming variables manually, the variable.name and value.name not working 11/2012
names(glabella.long) <- c("Row", "pop", "thickness")
# another way to remove NAs:
#glabella.long <- subset(glabella.long, !is.na(thickness))
```

```
# Plot the data using ggplot
library(ggplot2)
p <- ggplot(glabella.long, aes(x = pop, y = thickness))
# plot a reference line for the global mean (assuming no groups)
p <- p + geom_hline(yintercept = mean(glabella.long$thickness),
                    colour = "black", linetype = "dashed", size = 0.3, alpha = 0.5)
# boxplot, size=.75 to stand out behind CI
p <- p + geom_boxplot(size = 0.75, alpha = 0.5)
# points for observed data
p <- p + geom_point(position = position_jitter(w = 0.05, h = 0), alpha = 0.5)
# diamond at mean for each group
p <- p + stat_summary(fun.y = mean, geom = "point", shape = 18, size = 6,
                      colour = "red", alpha = 0.8)
# confidence limits based on normal distribution
p <- p + stat_summary(fun.data = "mean_cl_normal", geom = "errorbar",
                      width = .2, colour = "red", alpha = 0.8)
p <- p + labs(title = "Glabella thickness") + ylab("thickness (mm)")
print(p)
```



There are 3 groups, so there are 3 possible pairwise comparisons. If you want a Bonferroni analysis with FER of no greater than 0.05, you should do the individual comparisons at the $0.05/3 = 0.0167$ level. Except for the mild outlier in the Caucasian sample, the observed distributions are fairly symmetric, 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 Asians.

```
glabella.summary <- ddply(glabella.long, "pop",
     function(X) { data.frame( m = mean(X$thickness),
                               s = sd(X$thickness),
                               n = length(X$thickness) ) } )
glabella.summary
```

```
##    pop        m         s  n
## 1 cauc 5.812500 0.8334280 12
```

```
## 2 afam 6.461538 0.8946959 13
## 3 naaa 5.857143 1.1168047 14
```

```
fit.g <- aov(thickness ~ pop, data = glabella.long)
summary(fit.g)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## pop          2   3.40  1.6991   1.828  0.175
## Residuals   36  33.46  0.9295
```

```
fit.g
```

```
## Call:
##    aov(formula = thickness ~ pop, data = glabella.long)
##
## Terms:
##                       pop Residuals
## Sum of Squares   3.39829  33.46068
## Deg. of Freedom        2        36
##
## Residual standard error: 0.9640868
## Estimated effects may be unbalanced
```

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. **This is the end of the analysis!**

The Bonferroni intervals reinforce this conclusion, all the p-values are greater than 0.05. If you were to calculate CIs for the difference in population means, each would contain 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: $-1.62 \leq \mu_c - \mu_a \leq 0.32$, $-0.91 \leq \mu_n - \mu_c \leq 1.00$, and $-1.54 \leq \mu_n - \mu_a \leq 0.33$. The individual CIs have level $100(1 - 0.0167)\% = 98.33\%$.

```
# Bonferroni 95% Individual p-values
# All Pairwise Comparisons among Levels of glabella
pairwise.t.test(glabella.long$thickness, glabella.long$pop,
                pool.sd = TRUE, p.adjust.method = "bonf")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  glabella.long$thickness and glabella.long$pop
##
##      cauc afam
```

```
## afam 0.30 -
## naaa 1.00 0.34
##
## P value adjustment method: bonferroni
```

5.3 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.

A commonly-used alternative is **Tukey's** honest significant difference method (HSD). John Tukey's honest significant difference method is to reject the equality of a pair of means based, not on the t-distribution, but the studentized range distribution. To implement Tukey's method with a FER of α, reject $H_0 : \mu_i = \mu_j$ when

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

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.

```
#### Tukey's honest significant difference method (HSD)
## Fat
# Tukey 95% Individual p-values
# All Pairwise Comparisons among Levels of fat
TukeyHSD(fit.f)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = amount ~ type, data = fat.long)
##
## $type
##                   diff        lwr        upr     p adj
## fat2-fat1  10.5000000  -2.719028 23.7190277 0.1510591
## fat3-fat1   0.3333333 -12.885694 13.5523611 0.9998693
## fat4-fat1 -12.5000000 -25.719028  0.7190277 0.0679493
## fat3-fat2 -10.1666667 -23.385694  3.0523611 0.1709831
## fat4-fat2 -23.0000000 -36.219028 -9.7809723 0.0004978
## fat4-fat3 -12.8333333 -26.052361  0.3856944 0.0590077
```

```
## Glabella
# Tukey 95% Individual p-values
# All Pairwise Comparisons among Levels of pop
TukeyHSD(fit.g)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = thickness ~ pop, data = glabella.long)
##
## $pop
##                   diff        lwr       upr     p adj
## afam-cauc  0.64903846 -0.2943223 1.5923993 0.2259806
## naaa-cauc  0.04464286 -0.8824050 0.9716907 0.9923923
## naaa-afam -0.60439560 -1.5120412 0.3032500 0.2472838
```

Another popular method controls the **false discovery rate** (FDR) instead of the FER. The FDR is the expected proportion of false discoveries amongst the rejected hypotheses. The false discovery rate is a less stringent condition than the family-wise error rate, so these methods are more powerful than the others, though with a higher FER. I encourage you to learn more about the methods by Benjamini, Hochberg, and Yekutieli.

```
#### false discovery rate (FDR)
## Fat
# FDR
pairwise.t.test(fat.long$amount, fat.long$type,
                pool.sd = TRUE, p.adjust.method = "BH")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  fat.long$amount and fat.long$type
##
##      fat1    fat2    fat3
## fat2 0.05248 -       -
## fat3 0.94443 0.05248 -
## fat4 0.03095 0.00056 0.03095
##
## P value adjustment method: BH
```

```
## Glabella
# FDR
pairwise.t.test(glabella.long$thickness, glabella.long$pop,
                pool.sd = TRUE, p.adjust.method = "BH")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  glabella.long$thickness and glabella.long$pop
##
##      cauc afam
## afam 0.17 -
## naaa 0.91 0.17
##
## P value adjustment method: BH
```

■ ClickerQs — ANOVA, multiple comparisons ■

5.4 Checking Assumptions in ANOVA Problems

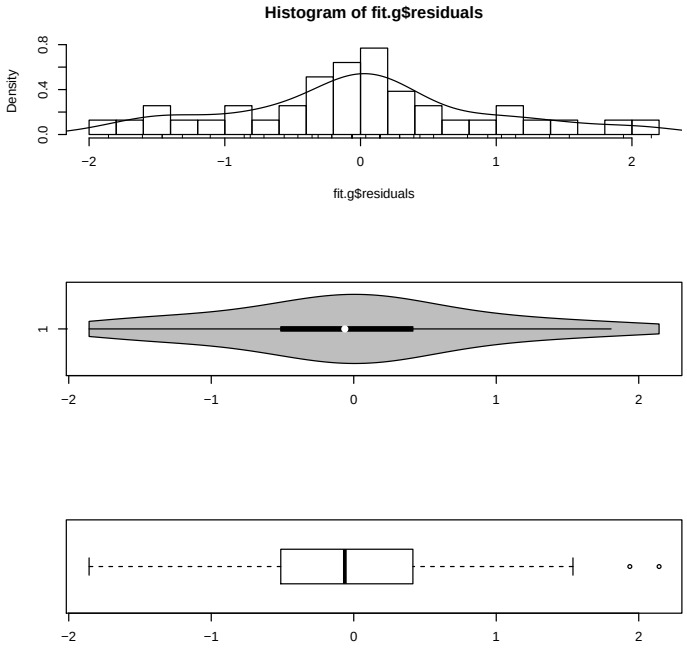
The classical ANOVA assumes that the populations have normal frequency curves and the populations have equal variances (or spreads). You can test the normality assumption using multiple normal QQ-plots and normal scores tests, which we discussed in Chapter 4. An alternative approach that is useful with 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

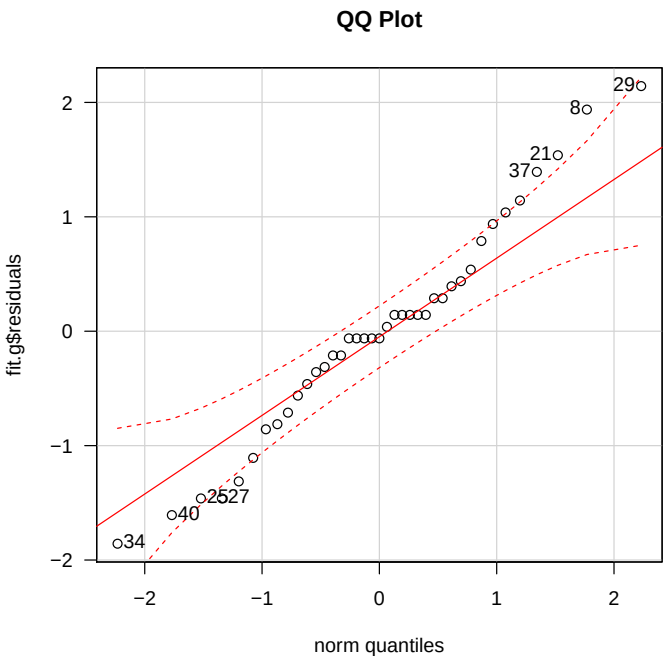
meaningful. (WHY?) This is done by subtracting the sample mean from each observation in the sample, giving the so-called **residuals**. A normal scores plot or histogram of the residuals should resemble a sample from a normal population. These two plots can be generated with output in `$residuals` from the `anova()` procedure.

For the glabella residuals, there are a few observations outside the confidence bands, but the formal normality tests each have p-values > 0.2, so there's weak but unconvincing evidence of nonnormality.

```
#### Checking Assumptions in ANOVA Problems
# plot of data
par(mfrow=c(3,1))
# Histogram overlaid with kernel density curve
hist(fit.g$residuals, freq = FALSE, breaks = 20)
points(density(fit.g$residuals), type = "l")
rug(fit.g$residuals)

# violin plot
library(vioplot)
vioplot(fit.g$residuals, horizontal=TRUE, col="gray")

# boxplot
boxplot(fit.g$residuals, horizontal=TRUE)


# QQ plot
par(mfrow=c(1,1))
library(car)
qqPlot(fit.g$residuals, las = 1, id.n = 8, id.cex = 1, lwd = 1, main="QQ Plot")
```

```
## 29  8 34 40 21 25 27 37
## 39 38  1  2 37  3  4 36
```






```
shapiro.test(fit.g$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit.g$residuals
## W = 0.97693, p-value = 0.5927
```

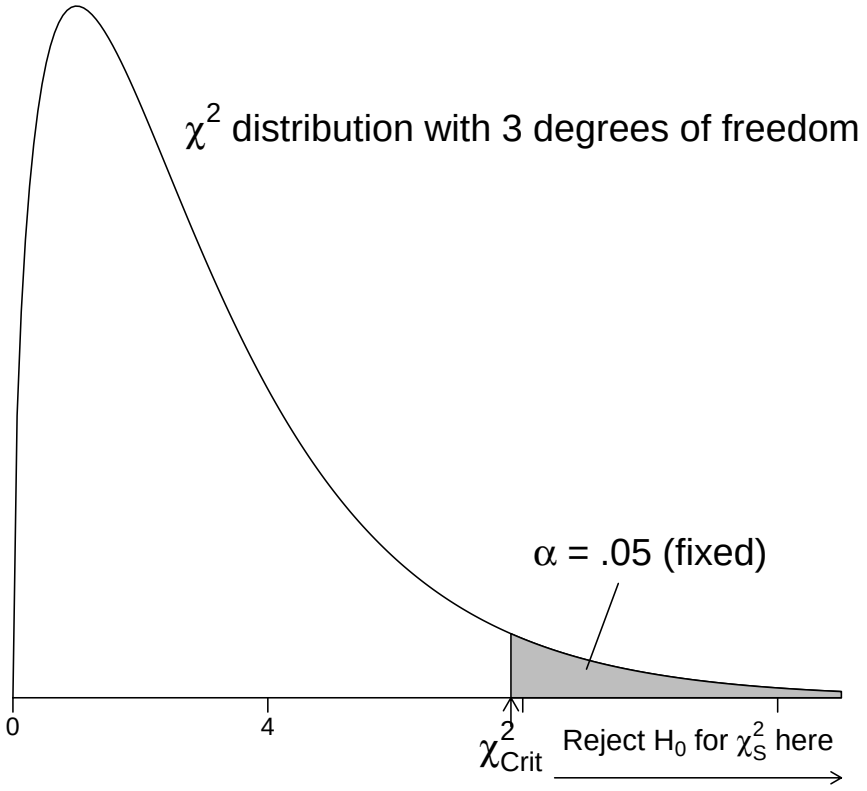
```
library(nortest)
ad.test(fit.g$residuals)
```

```
##
##  Anderson-Darling normality test
##
## data:  fit.g$residuals
## A = 0.37731, p-value = 0.3926
```

```
# lillie.test(fit.g$residuals)
cvm.test(fit.g$residuals)
```

```
##
##  Cramer-von Mises normality test
##
## data:  fit.g$residuals
## W = 0.070918, p-value = 0.2648
```

In Chapter 4, I illustrated the use of Bartlett's test and Levene's test for equal population variances, and showed how to evaluate these tests in R.



R does the calculation for us, as illustrated below. Because the p-value > 0.5, we fail to reject the null hypothesis that the population variances are equal. This result is not surprising given how close the sample variances are to each other.

```
## Test equal variance
# Barlett assumes populations are normal
bartlett.test(thickness ~ pop, data = glabella.long)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  thickness by pop
## Bartlett's K-squared = 1.1314, df = 2, p-value = 0.568
```

Levene's and Flinger's tests are consistent with Bartlett's.

```
# Levene does not assume normality, requires car package
library(car)
leveneTest(thickness ~ pop, data = glabella.long)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
```

```
## group  2  0.5286 0.5939
##       36
```

```
# Fligner is a nonparametric test
fligner.test(thickness ~ pop, data = glabella.long)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  thickness by pop
## Fligner-Killeen:med chi-squared = 1.0311, df = 2, p-value =
## 0.5972
```

5.5 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, and (3) mother smoked at least one pack of cigarettes a day during pregnancy.

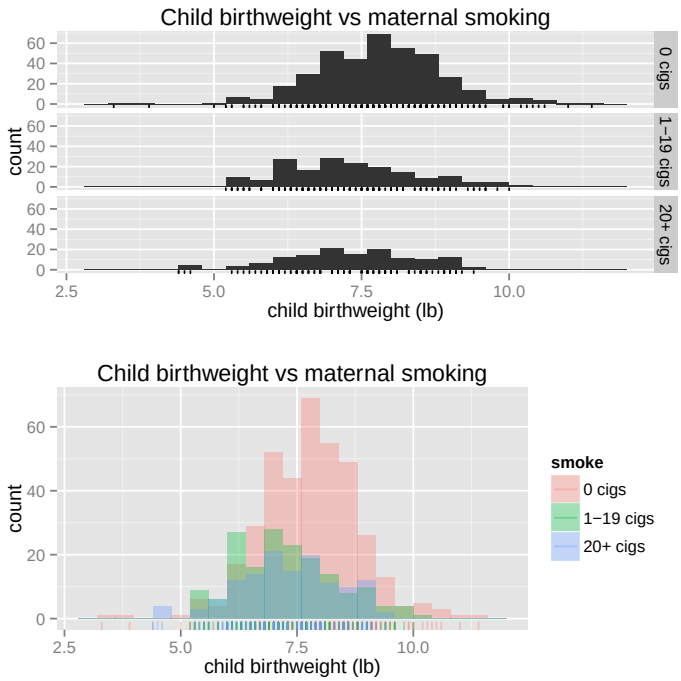
Let μ_i = pop mean birth weight (lb) for children in group i, $(i = 1, 2, 3)$. We wish to test $H_0 : \mu_1 = \mu_2 = \mu_3$ against H_A : not H_0.

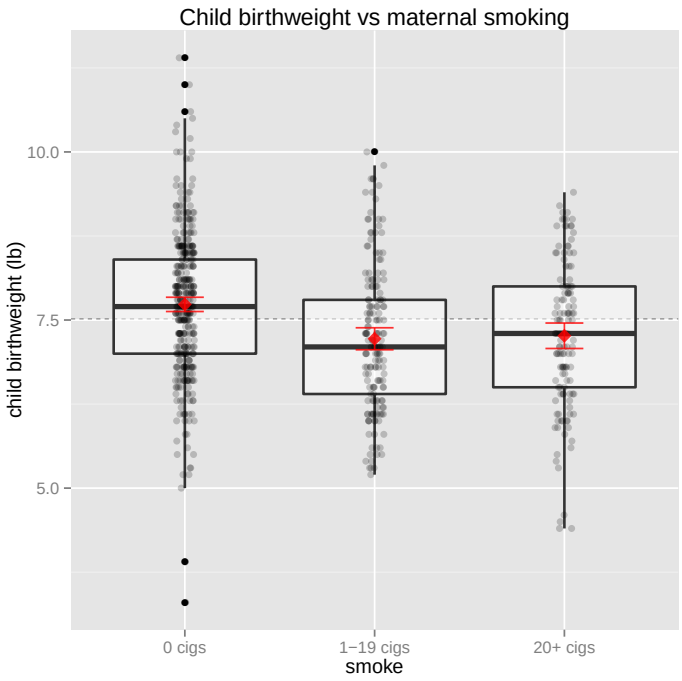
We read in the data, create a `smoke` factor variable, and plot the data by smoking group.

```
#### Example from the Child Health and Development Study (CHDS)
# description at http://statacumen.com/teach/ADA1/ADA1_notes_05-CHDS_desc.txt
# read data from website
chds <- read.csv("http://statacumen.com/teach/ADA1/ADA1_notes_05-CHDS.csv")

chds$smoke <- rep(NA, nrow(chds));
# no cigs
chds[(chds$m_smok == 0), "smoke"] <- "0 cigs";
# less than 1 pack (20 cigs = 1 pack)
chds[(chds$m_smok > 0) & (chds$m_smok < 20),"smoke"] <- "1-19 cigs";
# at least 1 pack (20 cigs = 1 pack)
chds[(chds$m_smok >= 20),"smoke"] <- "20+ cigs";
```

```
chds$smoke <- factor(chds$smoke)
```

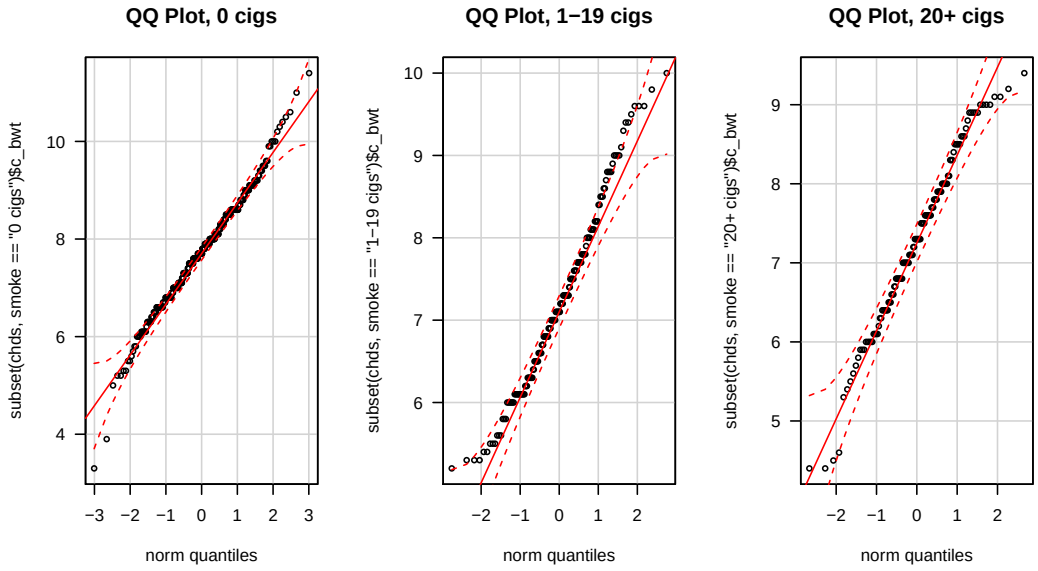
```
# histogram using ggplot
p1 <- ggplot(chds, aes(x = c_bwt))
p1 <- p1 + geom_histogram(binwidth = .4)
p1 <- p1 + geom_rug()
p1 <- p1 + facet_grid(smoke ~ .)
p1 <- p1 + labs(title = "Child birthweight vs maternal smoking") +
          xlab("child birthweight (lb)")
#print(p1)

p2 <- ggplot(chds, aes(x = c_bwt, fill=smoke))
p2 <- p2 + geom_histogram(binwidth = .4, alpha = 1/3, position="identity")
p2 <- p2 + geom_rug(aes(colour = smoke), alpha = 1/3)
p2 <- p2 + labs(title = "Child birthweight vs maternal smoking") +
          xlab("child birthweight (lb)")
#print(p2)

library(gridExtra)
grid.arrange(p1, p2, ncol=1)


# Plot the data using ggplot
library(ggplot2)
p <- ggplot(chds, aes(x = smoke, y = c_bwt))
# plot a reference line for the global mean (assuming no groups)
p <- p + geom_hline(yintercept = mean(chds$c_bwt),
                    colour = "black", linetype = "dashed", size = 0.3, alpha = 0.5)
# boxplot, size=.75 to stand out behind CI
p <- p + geom_boxplot(size = 0.75, alpha = 0.5)
# points for observed data
p <- p + geom_point(position = position_jitter(w = 0.05, h = 0), alpha = 0.2)
# diamond at mean for each group
p <- p + stat_summary(fun.y = mean, geom = "point", shape = 18, size = 4,
                      colour = "red", alpha = 0.8)
# confidence limits based on normal distribution
p <- p + stat_summary(fun.data = "mean_cl_normal", geom = "errorbar",
                      width = .2, colour = "red", alpha = 0.8)
p <- p + labs(title = "Child birthweight vs maternal smoking") +
          ylab("child birthweight (lb)")
print(p)
```





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 (only 4 are visible!) 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.

```
library(car)
par(mfrow=c(1,3))
qqPlot(subset(chds, smoke == "0 cigs"   )$c_bwt, las = 1, id.n = 0,
  id.cex = 1, lwd = 1, main="QQ Plot, 0 cigs")
qqPlot(subset(chds, smoke == "1-19 cigs")$c_bwt, las = 1, id.n = 0,
  id.cex = 1, lwd = 1, main="QQ Plot, 1-19 cigs")
qqPlot(subset(chds, smoke == "20+ cigs" )$c_bwt, las = 1, id.n = 0,
  id.cex = 1, lwd = 1, main="QQ Plot, 20+ cigs")
```




```
library(nortest)
# 0 cigs    --------------------
shapiro.test(subset(chds, smoke == "0 cigs"   )$c_bwt)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subset(chds, smoke == "0 cigs")$c_bwt
## W = 0.98724, p-value = 0.00199
```

```
ad.test(     subset(chds, smoke == "0 cigs"   )$c_bwt)
```

```
##
##  Anderson-Darling normality test
##
## data:  subset(chds, smoke == "0 cigs")$c_bwt
## A = 0.92825, p-value = 0.01831
```

```
cvm.test(    subset(chds, smoke == "0 cigs"   )$c_bwt)
```

```
##
##  Cramer-von Mises normality test
##
## data:  subset(chds, smoke == "0 cigs")$c_bwt
## W = 0.13844, p-value = 0.03374
```

```
# 1-19 cigs ----------------------
shapiro.test(subset(chds, smoke == "1-19 cigs")$c_bwt)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subset(chds, smoke == "1-19 cigs")$c_bwt
## W = 0.97847, p-value = 0.009926
```

```
ad.test(     subset(chds, smoke == "1-19 cigs")$c_bwt)
```

```
##
```

```
##  Anderson-Darling normality test
##
## data:  subset(chds, smoke == "1-19 cigs")$c_bwt
## A = 0.83085, p-value = 0.03149
```

```
cvm.test(    subset(chds, smoke == "1-19 cigs")$c_bwt)
```

```
##
##  Cramer-von Mises normality test
##
## data:  subset(chds, smoke == "1-19 cigs")$c_bwt
## W = 0.11332, p-value = 0.07317
```

```
# 20+ cigs  --------------------
shapiro.test(subset(chds, smoke == "20+ cigs" )$c_bwt)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subset(chds, smoke == "20+ cigs")$c_bwt
## W = 0.98127, p-value = 0.06962
```

```
ad.test(     subset(chds, smoke == "20+ cigs" )$c_bwt)
```

```
##
##  Anderson-Darling normality test
##
## data:  subset(chds, smoke == "20+ cigs")$c_bwt
## A = 0.40008, p-value = 0.3578
```

```
cvm.test(    subset(chds, smoke == "20+ cigs" )$c_bwt)
```

```
##
##  Cramer-von Mises normality test
##
## data:  subset(chds, smoke == "20+ cigs")$c_bwt
## W = 0.040522, p-value = 0.6694
```

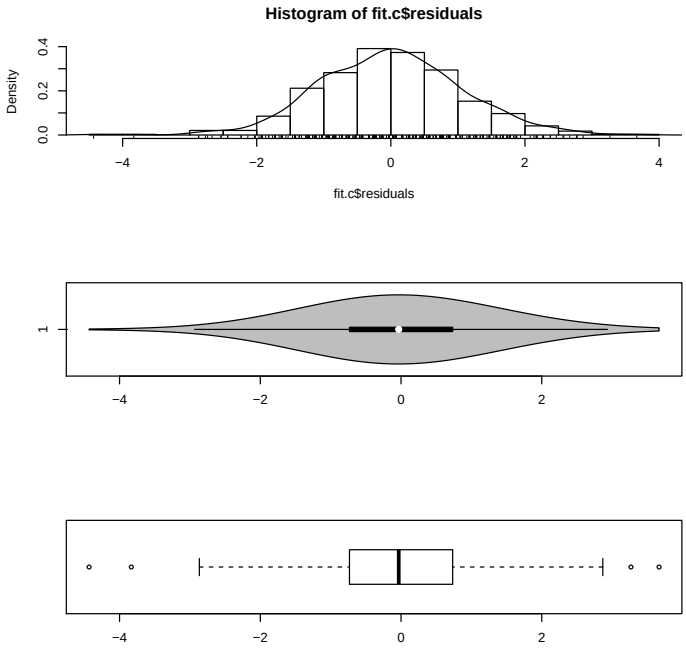
Fit the ANOVA (we'll look at the table below).

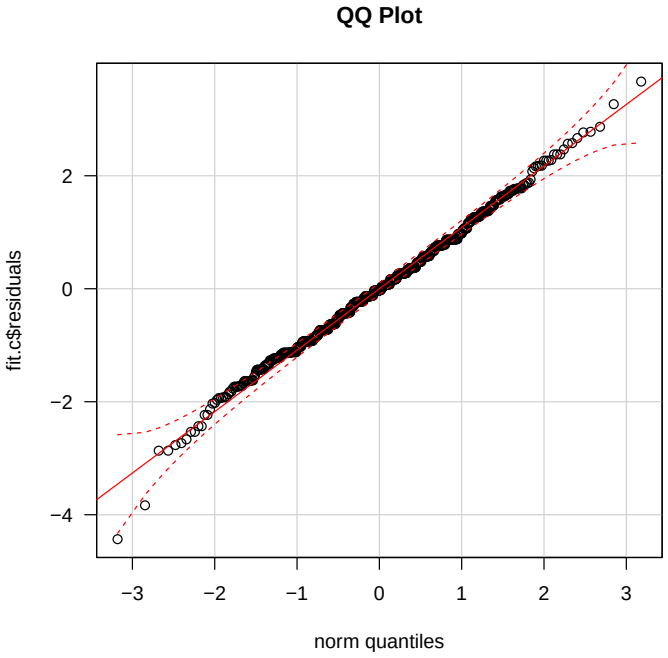
```
fit.c <- aov(c_bwt ~ smoke, data = chds)
```

A formal test of normality on the residuals of the combined sample is marginally significant (SW p-value= 0.047, others > 0.10). 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.

```
# plot of data
par(mfrow=c(3,1))
```

```r
# Histogram overlaid with kernel density curve
hist(fit.c$residuals, freq = FALSE, breaks = 20)
points(density(fit.c$residuals), type = "l")
rug(fit.c$residuals)

# violin plot
library(vioplot)
vioplot(fit.c$residuals, horizontal=TRUE, col="gray")

# boxplot
boxplot(fit.c$residuals, horizontal=TRUE)


# QQ plot
par(mfrow=c(1,1))
library(car)
qqPlot(fit.c$residuals, las = 1, id.n = 0, id.cex = 1, lwd = 1, main="QQ Plot")
```






```r
shapiro.test(fit.c$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit.c$residuals
## W = 0.99553, p-value = 0.04758
```

```r
library(nortest)
ad.test(fit.c$residuals)
```

```
##
##  Anderson-Darling normality test
##
```

```
## data:  fit.c$residuals
## A = 0.62184, p-value = 0.1051
```

```
cvm.test(fit.c$residuals)
```

```
##
##  Cramer-von Mises normality test
##
## data:  fit.c$residuals
## W = 0.091963, p-value = 0.1449
```

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 0.4. Thus, the standard ANOVA appears to be appropriate here.

```
# calculate summaries
chds.summary <- ddply(chds, "smoke",
     function(X) { data.frame( m = mean(X$c_bwt),
                               s = sd(X$c_bwt),
                               n = length(X$c_bwt) ) } )
chds.summary
```

```
##       smoke        m        s   n
## 1    0 cigs 7.732808 1.052341 381
## 2 1-19 cigs 7.221302 1.077760 169
## 3  20+ cigs 7.266154 1.090946 130
```

```
## Test equal variance
# assumes populations are normal
bartlett.test(c_bwt ~ smoke, data = chds)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  c_bwt by smoke
## Bartlett's K-squared = 0.3055, df = 2, p-value = 0.8583
```

```
# does not assume normality, requires car package
library(car)
leveneTest(c_bwt ~ smoke, data = chds)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   2  0.7591 0.4685
##       677
```

```
# nonparametric test
fligner.test(c_bwt ~ smoke, data = chds)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  c_bwt by smoke
## Fligner-Killeen:med chi-squared = 2.0927, df = 2, p-value =
## 0.3512
```

The p-value for the F-test is less than 0.0001. We would reject H_0 at any of the usual test levels (such as 0.05 or 0.01). The data suggest that the population mean birth weights differ across smoking status groups.

```
summary(fit.c)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## smoke         2   40.7  20.351    17.9 2.65e-08 ***
## Residuals   677  769.5   1.137
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit.c
```

```
## Call:
##    aov(formula = c_bwt ~ smoke, data = chds)
##
## Terms:
##                     smoke Residuals
## Sum of Squares   40.7012  769.4943
## Deg. of Freedom        2       677
##
## Residual standard error: 1.066126
## Estimated effects may be unbalanced
```

The Tukey multiple comparisons suggest that the mean birth weights are different (higher) for children born to mothers that did not smoke during pregnancy.

```
## CHDS
# Tukey 95% Individual p-values
TukeyHSD(fit.c)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = c_bwt ~ smoke, data = chds)
##
## $smoke
##                           diff        lwr        upr     p adj
## 1-19 cigs-0 cigs   -0.51150662 -0.7429495 -0.2800637 0.0000008
## 20+ cigs-0 cigs    -0.46665455 -0.7210121 -0.2122970 0.0000558
## 20+ cigs-1-19 cigs  0.04485207 -0.2472865  0.3369907 0.9308357
```