

## One-way Analysis of Variance (Chapter 12)

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	$\dots$	population $k$
mean	$\mu_1$	$\mu_2$	$\dots$	$\mu_k$
std dev	$\sigma_1$	$\sigma_2$	$\dots$	$\sigma_k$

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

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

Let  $Y_{ij}$  denote the  $j^{\text{th}}$  observation in the  $i^{\text{th}}$  sample and define the total sample size  $n = n_1 + n_2 + \cdots + n_k$ . Let  $\bar{\bar{Y}}$  be the average response over all samples (combined):

$$\bar{\bar{Y}} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij} = \sum_{i=1}^k \left( \frac{n_i}{n} \right) \bar{Y}_i.$$

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

A test statistic  $F_{obs}$ , called an  $F$ -statistic, is used to test  $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$  against  $H_1 : \text{at least two of the } \mu_1, \dots, \mu_k \text{ are different.}$

The assumptions needed for the standard ANOVA  $F$ -test generalize those for the pooled two-sample  $t$ -test assumptions:

1. Independent random samples from each population.
2. The population density curves are normal.
3. The populations have equal standard deviations,

$$\sigma_1 = \sigma_2 = \cdots = \sigma_k.$$

The  $F$ -statistic is computed from two measures of spread which break 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:

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

The notation  $\sum_{ij}$  is the sum over all  $i$  and  $j$ , i.e.

$$\sum_{ij} (Y_{ij} - \bar{Y}_i)^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2.$$

Note that

$$\text{SS(Within)} = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2,$$

where  $s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$  is the unbiased variance estimate from group  $i$ .

The **Between SS**, often called the Model SS, measures the spread among the sample means

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

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

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

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:  $n - k$ . These two *df* add to give  $df(\text{Total}) = (k - 1) + (n - k) = n - 1$ .

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

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

R fits ANOVA models with the `aov()` function. R does not include the `SS(Total)` row in output.

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)

$$\left(\frac{n_1 - 1}{n - k}\right) s_1^2 + \left(\frac{n_2 - 1}{n - k}\right) s_2^2 + \cdots + \left(\frac{n_k - 1}{n - k}\right) s_k^2 = s_{pooled}^2$$

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. The MS(Within) is identical to the **pooled variance estimator** in a two-sample problem when  $k = 2$ .

The MS(Between)

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

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

The MS(Total)

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

is the variance in the combined data set.

The MS(Within) estimates  $\sigma^2$ :

$$E\{\text{MS(Within)}\} = \sigma^2.$$

The MS(Between) estimates something that's larger than  $\sigma^2$ :

$$E\{\text{MS(Between)}\} = \frac{1}{k-1} \sum_{i=1}^k n_i (\mu_i - \bar{\mu})^2 + \sigma^2,$$

where  $\bar{\mu} = \sum_{i=1}^k \left(\frac{n_i}{n}\right) \mu_i$ . When  $H_0 : \mu_1 = \dots = \mu_k$  is true, then  $\mu_i = \bar{\mu}$  for  $i = 1, \dots, k$  and  $E\{\text{MS(Between)}\} = \sigma^2$ .

When  $H_0 : \mu_1 = \dots = \mu_k$  is *not true*, then  $E\{\text{MS(Between)}\} > \sigma^2$ .

The ratio of MS(Between) to MS(Within) then provides a measure of how unlikely  $H_0$  is. The more spread out the  $\mu_1, \dots, \mu_k$  are, the larger MS(Between) will be relative to MS(Within).

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

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

Large values of  $F_{obs}$  indicate large variability among the sample means  $\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_k$  relative to the spread of the data within samples. That is, large values of  $F_{obs}$  suggest that  $H_0$  is false. When  $H_0$  is true and all groups have the same mean,  $F_{obs}$  has an  $F$  distribution with  $df$   $k - 1$  in the numerator and  $n - k$  in the denominator (i.e. the  $df$  for the numerators and denominators in the  $F$ -ratio).

Formally, for a size  $\alpha$  test, reject  $H_0$  if  $F_{obs} \geq F_{crit}$ , where  $F_{crit}$  is the upper- $\alpha$  percentile from an  $F$  distribution with numerator degrees of freedom  $k - 1$  and denominator degrees of freedom  $n - k$ . The  $p$ -value for the test is the area under the  $F$ -probability curve to the right of  $F_{obs}$ .

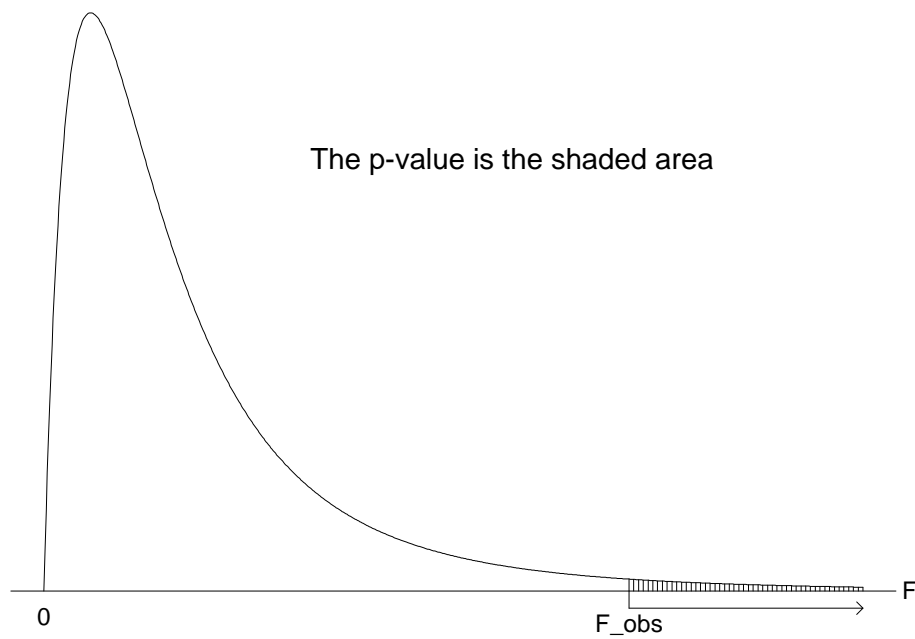


Figure 1: Reject at level  $\alpha$  when  $p\text{-value} \leq \alpha$ .

## 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  $k = 4$  fats,  $n_i = 6$  batches of 24 doughnuts were prepared. The data are grams of fat absorbed per batch (minus 100).

Let  $\mu_i =$  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  $H_1 : \text{not } H_0$ . There is no strong evidence against normality here.

Furthermore the spreads across the 4 groups are close. The standard ANOVA appears to be appropriate here.

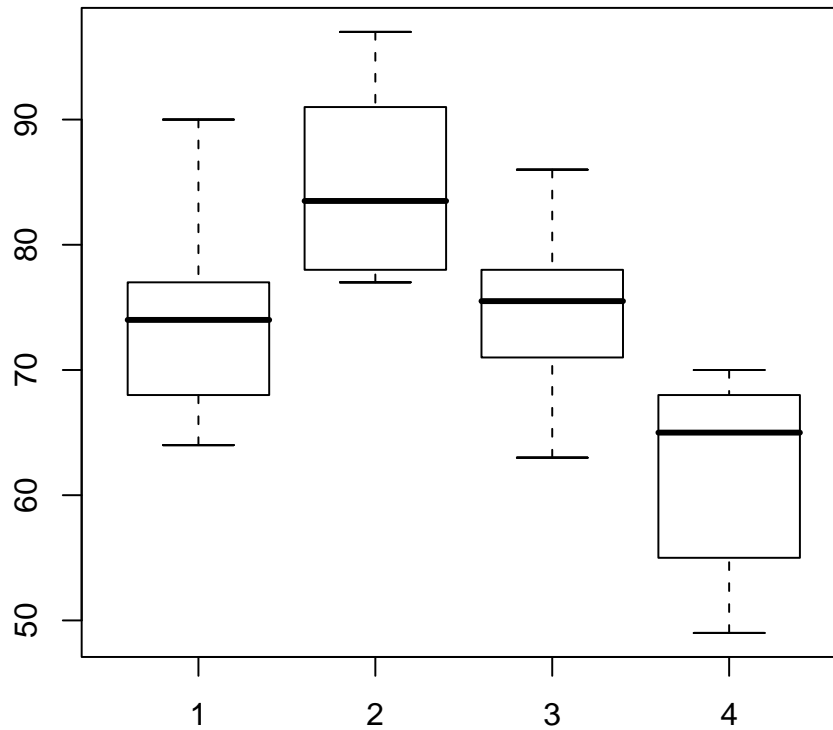


Figure 2: Side-by-side boxplots for fat data.

Fat 1	Fat 2	Fat 3	Fat 4
64	78	75	55
72	91	86	66
68	97	78	49
77	82	71	64
90	85	63	70
76	77	76	68

```

> f<-c(64,72,68,77,90,76,78,91,97,82,85,77,75,86,78,71,63,76,55,66,49,64,70,68)
> g<-factor(c(1,1,1,1,1,1,2,2,2,2,2,2,3,3,3,3,3,3,4,4,4,4,4,4))
> boxplot(f~g)
> fit <- aov(f~g)
> summary(fit)

```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
g	3	1595.50	531.83	7.9477	0.001104 **
Residuals	20	1338.33	66.92		

The  $p$ -value for the  $F$ -test is 0.0011. The scientist would reject  $H_0$  at any of the usual test levels (i.e.  $\alpha = 0.05$  or  $\alpha = 0.01$ ). The data suggest that the population mean absorption rates differ across fats.

## 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. We will discuss Fisher's, Bonferroni's and Tukey's methods for comparing all pairs of means.

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

1. Carry out the ANOVA  $F$ -test of  $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$  at the  $\alpha$  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  $\alpha$  level. Use  $s_{pooled}$  from the ANOVA table and  $df(\text{Residual})$ .

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

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

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

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

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  $n_1 = n_2 = \dots = n_k$  then the LSD is the same for each comparison, where  $n_1$  is the common sample size:

$$|\bar{Y}_i - \bar{Y}_j| \geq LSD = t_{crit} s_{pooled} \sqrt{\frac{2}{n_1}}.$$

Let's perform Fisher's method on the doughnut data by hand, using  $\alpha = 0.05$ . At the first step, you reject the hypothesis that the population mean absorptions are equal because  $p\text{-value} = 0.0011$ . At the second step, compare all pairs of fats at the 5% level.

```
> fit
```

```
Terms:
```

	g	Residuals
Sum of Squares	1595.500	1338.333
Deg. of Freedom	3	20

```
Residual standard error: 8.18026
```

```
> qt(c(0.975),20)
```

```
[1] 2.085963
```

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

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

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.

```
> fm <- matrix(f,6)
> apply(fm,2,mean)
[1] 74.50000 85.00000 74.83333 62.00000
```

Fat	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.

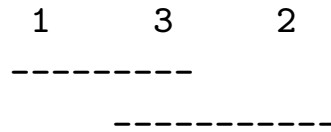
There are three groups for the doughnut data, with no overlap. Fat 2 is in a group by itself, and so is Fat 4. Fats 3 and 1 are in a group together. This information can be summarized by ordering the samples from lowest to highest average, and then connecting the fats in the same group using an underscore:

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.

## A note of caution

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:



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 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 doesn't make sense!

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 R

To get Fisher comparisons in R, use the `pairwise.t.test()` function and specify the method for adjusting the  $p$ -value as “none.”

```
> pairwise.t.test(f,g,p.adjust.method="none")
      Pairwise comparisons using t tests with pooled SD
data:  f and g
      1      2      3
2 0.038 -        -
3 0.944 0.044   -
4 0.015 9.3e-05 0.013

P value adjustment method: none
```

This gives  $p$ -values for the  $\binom{6}{2} = 6$  pairwise comparisons using the LSD method above. We see that we reject  $\mu_1 = \mu_2$ ,  $\mu_1 = \mu_4$ ,  $\mu_2 = \mu_3$ ,  $\mu_2 = \mu_4$  and  $\mu_3 = \mu_4$ . We accept  $\mu_1 = \mu_3$ ; the same conclusion as by hand.

## Discussion of the FSD Method

There are  $c = \binom{k}{2} = 0.5k(k - 1)$  pairs of means to compare in the second step of the FSD method. Each comparison is done at the  $\alpha$  level, where for a generic comparison of the  $i^{th}$  and  $j^{th}$  populations

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

This probability is called 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) 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  $\alpha$  level, then

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

For example, in the doughnut problem where  $k = 4$ , there are  $c = 6$  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.

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 deflates 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  $\alpha$ , by setting the individual error rate for each of the  $c$  comparisons of interest to  $\alpha/c$ . Larger differences in the sample means are needed before declaring statistical significance using the Bonferroni adjustment than when using the FSD method at the  $\alpha$  level.

The function `pairwise.t.test()` performs Bonferroni comparisons by simply multiplying each individual  $p$ -value by  $c = 0.5k(k - 1)$ . We simply reject those  $H_0 : \mu_i = \mu_j$  with *adjusted*  $p$ -values smaller than the FER.

```

> pairwise.t.test(f,g,p.adjust.method="bonferroni")
      Pairwise comparisons using t tests with pooled SD
data:  f and g
      1      2      3
2 0.22733 -      -
3 1.00000 0.26241 -
4 0.09286 0.00056 0.07960

```

P value adjustment method: bonferroni

Each of these  $p$ -values are the same as the individual pairwise  $p$ -values multiplied by  $c = 6$ . We now reject  $\mu_2 = \mu_4$  *only* with an overall FER of 0.05:

```

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

```

The Bonferroni method produces “coarser” groups than the FSD method, because the individual comparisons are conducted at a lower significance level to cap off the FER. Equivalently, the LSD is inflated for the Bonferroni method. For example, in the doughnut problem with  $FER \leq 0.05$ , LSD is (without belaboring the computations involved)

$$LSD = 2.929 \times 8.18 \times \sqrt{\frac{2}{6}} = 13.82$$

versus an  $LSD=9.85$  for the FSD method. Referring back to our table of sample means, we see that the sole comparison where the absolute difference between sample means exceeds 13.82 involves Fats 2 and 4.

## **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 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 (defined up front), and ignore the rest.

A commonly used alternative to FSD and Bonferroni is **Tukey's** honest significant difference method (HSD). Tukey's method can be implemented in R through `TukeyHSD(fit)` where `fit` is from the `aov()` function, e.g. `fit <- aov(f~g)`.

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

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

where  $q_{crit}$  is the  $\alpha$  level critical value of the studentized range distribution. For the doughnut fats, the groupings based on Tukey and Bonferroni comparisons are identical.

```

> TukeyHSD(fit)
  Tukey multiple comparisons of means
    95% family-wise confidence level

$g
      diff      lwr      upr    p adj
2-1 10.5000000 -2.719028 23.7190277 0.1510591
3-1  0.3333333 -12.885694 13.5523611 0.9998693
4-1 -12.5000000 -25.719028  0.7190277 0.0679493
3-2 -10.1666667 -23.385694  3.0523611 0.1709831
4-2 -23.0000000 -36.219028 -9.7809723 0.0004978
4-3 -12.8333333 -26.052361  0.3856944 0.0590077

```

A bonus from using `TukeyHSD(fit)` is that 95% CI's are obtained for all possible pairings of means. These are *simultaneous* 95% CI's. The probability that each difference  $\mu_i - \mu_j$  will be in its respective interval *simultaneously* is at least 95% before data are collected.

The  $p$ -values are adjusted to keep  $\text{FER} \leq 0.05$ . Other FER levels can be specified. Here, as in Bonferroni, we reject only  $H_0 : \mu_4 = \mu_2$  with  $\text{FER} \leq 0.05$ .

## 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 normality tests, which we discussed earlier. An alternative approach that is useful with three or more samples is to consider the normality of the residuals  $r_{ij} = y_{ij} - \bar{y}_i$  from all groups instead, obtained as `r <- fit$res` in R. A histogram of the residuals should resemble a sample from a normal population.

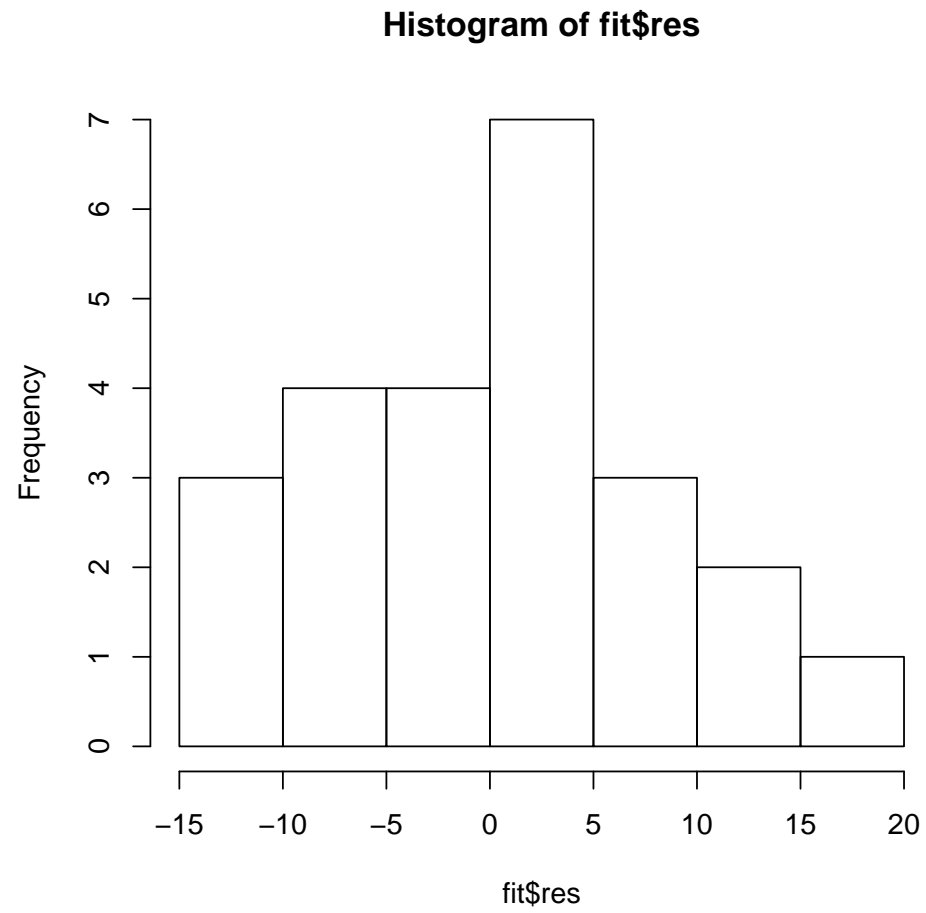


Figure 3: From `hist(fit$res)` – the residuals look fairly normal.

**Bartlett's test** tests the null  $H_0 : \sigma_1 = \sigma_2 = \dots = \sigma_k$ . Bartlett's test assumes normally distributed data. As above, let  $n = n_1 + n_2 + \dots + n_k$ , where  $n_i$ s is the sample sizes from the  $i^{th}$  group, 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  $H_0 : \sigma_1^2 = \dots = \sigma_k^2$  is given by

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

where  $s_{pooled}^2$  is the pooled estimator of variance and  $s_i^2$  is the unbiased estimated variance based on the  $i^{th}$  sample

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2.$$

Large values of  $B_{obs}$  suggest that the population variances are unequal. If  $H_0$  is true then  $B_{obs}$  has approximately a  $\chi_{k-1}^2$  distribution, a chi-squared distribution with  $k - 1$  *df*. The *p*-value for the test is given by the area under the chi-squared curve to the right of  $B_{obs}$ .

```
> bartlett.test(f,g)
      Bartlett test of homogeneity of variances
data:  f and g
Bartlett's K-squared = 0.1628, df = 3, p-value = 0.9834
```

We accept  $H_0 : \sigma_1 = \sigma_2 = \sigma_3 = \sigma_4$  at the  $\alpha = 0.10$  level.

## Example from the Child Health and Development Study

We consider data from the birth records of  $n = 680$  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  $k = 3$  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  $\mu_i =$  population mean birth weight (in lbs) for children in group  $i$  where  $i = 1, 2, 3$ . We wish to test  $H_0 : \mu_1 = \mu_2 = \mu_3$  against  $H_1 :$  at least two of  $\mu_1, \mu_2, \mu_3$  are different.

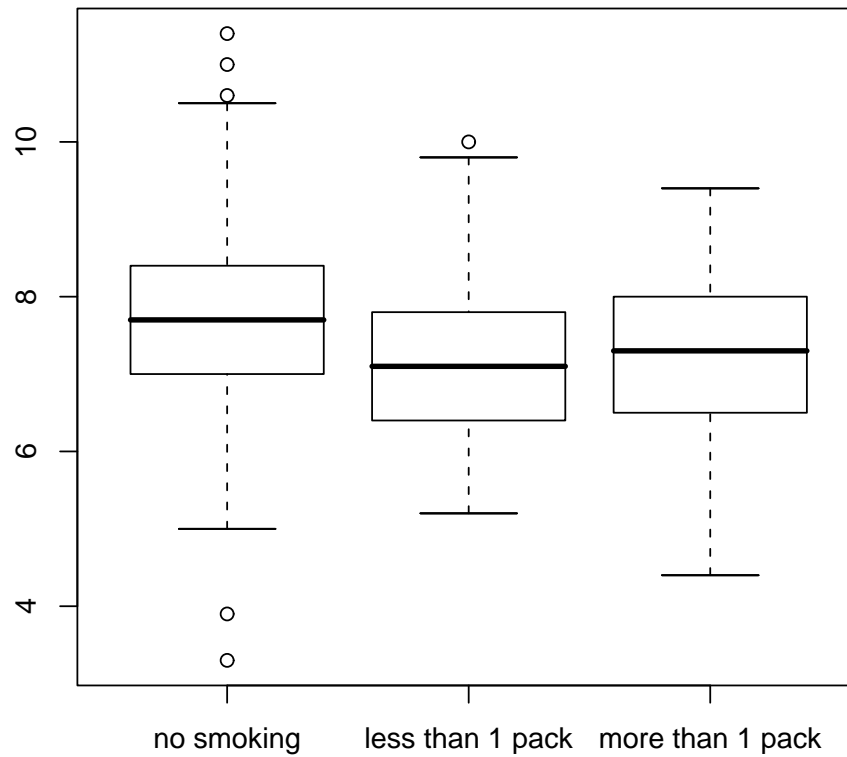


Figure 4: Boxplots of birthweight (lbs) by smoking group.

There are several variables in this data set including

- 1) ID
- 2) child's head circumference (inches)
- 3) child's length (inches)
- 4) child's birth weight (pounds)
- 5) gestation (weeks)
- 6) maternal age (years)
- 7) maternal smoking (cigarettes/day)
- 8) maternal height (inches)
- 9) maternal pre-pregnancy weight (pounds)
- 10) paternal age (years)
- 11) paternal years of education
- 12) paternal smoking (cigarettes/day)
- 13) paternal height (inches)

The full data set will be read into R and the needed columns plucked out and manipulated to perform the ANOVA...

```

> c <- read.table("c:/tim/PubH7400/chds.txt")
> bw <- c[,4] # birthweight in lbs
> smoke <- c[,7] # number of cigarettes smoke per day
> g <- rep(1,680) # smoking indicator initially set=1
> for(i in 1:680){
+   if(smoke[i]>0) g[i]=2
+   if(smoke[i]>19) g[i]=3
+ }
> g <- factor(g,labels=c("no smoking","less than 1 pack","more than 1 pack"))
> boxplot(bw~g)
> one.pack <- subset(bw,g=="less than 1 pack")
> no.smoke <- subset(bw,g=="no smoking")
> more.one.pack <- subset(bw,g=="more than 1 pack")
> shapiro.test(no.smoke)
      Shapiro-Wilk normality test
data:  no.smoke W = 0.9872, p-value = 0.001991
> shapiro.test(one.pack)
      Shapiro-Wilk normality test
data:  one.pack W = 0.9785, p-value = 0.009926
> shapiro.test(more.one.pack)
      Shapiro-Wilk normality test
data:  more.one.pack W = 0.9813, p-value = 0.06962
> fit <- aov(bw~g)
> summary(fit)

```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
g	2	40.70	20.35	17.904	2.648e-08 ***
Residuals	677	769.49	1.14		

The  $p$ -value for testing  $H_0 : \mu_1 = \mu_2 = \mu_3$  is 0.000000026. We reject that there is no difference in mean birthweight across smoking groups. However, the presence of 5 outliers out of  $n_1 = 381$  indicates the non-smoker's infant birthweights are probably not normal. The other two groups fare similarly.

```
> bartlett.test(bw,g)
      Bartlett test of homogeneity of variances
data:  bw and g Bartlett's K-squared = 0.3055, df = 2, p-value =
0.8583
> hist(fit$res)
> shapiro.test(fit$res)
      Shapiro-Wilk normality test
data:  fit$res W = 0.9955, p-value = 0.04758
```

We *do not* reject that the variances are the same across groups, i.e. we accept  $H_0 : \sigma_1 = \sigma_2 = \sigma_3$ . However, we do reject that the residuals come from a normal population with a  $p$ -value of 0.048.

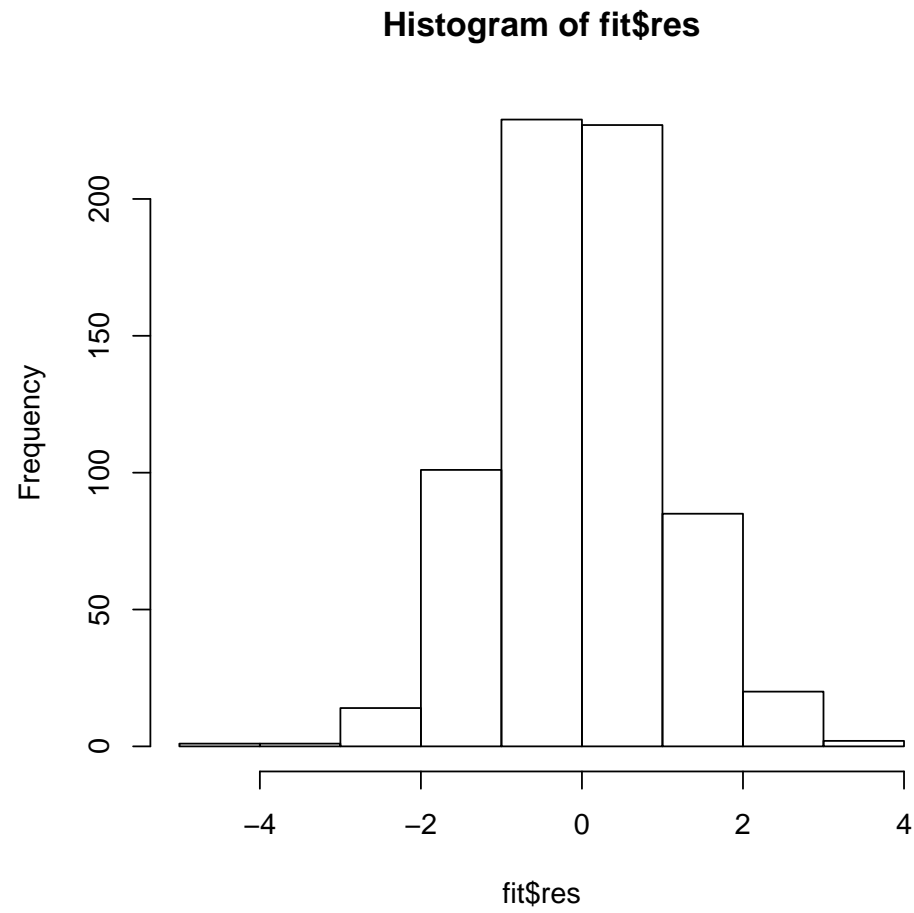


Figure 5: Histogram of residuals from ANOVA fit.

Normal probability plots (not discussed but in your book) and the histogram of the residuals above suggest that the population distributions are heavy tailed. A formal test of the normality of the residuals rejects at the 0.05 level.

However, I am not overly concerned about this for the following reasons: in large samples, small deviations from normality are *often* statistically significant – i.e. *no data* are truly normal.

The small deviations we are seeing here are not likely to impact our conclusions; non-parametric methods that do not require normality (e.g. the Kruskal-Wallis test) will lead to the same conclusions. In fact `kruskal.test(bw~g)` yields a  $p$ -value of 0.000000011.

We would reject  $H_0$  at any of the usual test levels (i.e. 0.05 or 0.01). 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.

```
> TukeyHSD(fit)
```

```
  Tukey multiple comparisons of means
```

```
 95% family-wise confidence level
```

```
$g
```

	diff	lwr	upr	p adj
less than 1 pack-no smoking	-0.51150	-0.7429	-0.2800	0.0000
more than 1 pack-no smoking	-0.46665	-0.7210	-0.2122	0.0000
more than 1 pack-less than 1 pack	0.04485	-0.2472	0.3369	0.9308

## Alternative ANOVA methods

1. Welch's ANOVA method is appropriate for normal populations with unequal variances. The test is a generalization of Satterthwaite's two-sample test discussed earlier. The `oneway.test()` function performs this test.
2. The Wilcoxon or Kruskal-Wallis non-parametric ANOVA is appropriate with non-normal populations with similar spreads. The null hypothesis can be interpreted as  $H_0 : \eta_1 = \dots = \eta_k$  where  $\eta_i$  is the median response in the  $i^{th}$  group. The `kruskal.test()` function carries out the test.

Multiple pairwise comparisons can proceed in either case using Bonferroni's correction on pairwise tests for means  $H_0 : \mu_i = \mu_j$  for (1) using `t.test()`, or medians  $H_0 : \eta_i = \eta_j$  for (2) using `wilcox.test()`.