Chapter 7:

Comparing three or more means

7.1 The Multiple Comparisons Issue 175
7.2 Analysis of Variance 177
7.3 Non-Parametric Means Tests I 190
7.4 Non-Parametric Means Tests II* 195
7.5 Post-Hoc Testing 196
7.6 Further Examples 201
7.7 Conclusion 208
7.8 End of Chapter Materials 209

This chapter continues examining procedures concerning population means. In Chapter 5, we introduced tests and confidence intervals covering means of a *single* population. In Chapter 6, we introduced tests and confidence intervals comparing means of *two* populations. In this chapter, we formulate tests comparing the means of more than two populations. Along with the usual assumptions of independence and of Normally-distributed measurements in each population, the parametric test requires the populations to have the same variance.

There are a total of six major conferences and five mid-major conferences in the Division I Football Bowl Subdivision (FBS). The six major conferences in 2009 were the Atlantic Coast Conference (ACC), Big East Conference, Big Ten Conference, Big 12 Conference, Pacific-10 Conference (Pac-10), and the Southeastern Conference (SEC). In terms of points scored in the games, does any major conference score significantly more points than any of the other major conferences?

Thus far, we have only examined tests that help us to compare one sample to a proposed population mean or to compare the means of two samples—in either case, running *one* test. Unfortunately, we often have several samples or groups among which we want to compare means. We may be tempted to continue using the methods from the previous chapter and just apply these pairwise tests to each possible pair of populations. There are two problems with this, however. The first problem is the sheer number of pairwise tests one would have to perform. For the introductory example, one would need to perform $\binom{6}{2} = 15$ pairwise tests.

multiple comparisons

The second issue is the inflation of the Type I Error rate if you do perform all 15 tests. Recall that the level of a test is the *actual* Type I Error rate, the true probability of rejecting a true null hypothesis. Each test we perform has that same nominal error rate, α. Performing multiple tests increases the Type I Error rate of the entire *experiment*. The amount of increase depends on several factors, including the level of independence between the tests. We can, however, easily calculate an upper bound on how much the Type I Error rate increases.

experiment-wise error rate

7.1: The Multiple Comparisons Issue

To see the multiple comparisons issue, let use suppose our null hypothesis is

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

That is, we wish to test if the three groups have the same mean. This null hypothesis actually contains three pairwise tests:

$$H_0^1 : \mu_1 = \mu_2$$
$$H_0^2 : \mu_1 = \mu_3$$
$$H_0^3 : \mu_2 = \mu_3$$

Recall that the Type I Error rate is the probability of rejecting a true null hypothesis. We would like this error rate to be held to (at most) α. However, if we reject each of the three sub-tests at the α level, then the Type I Error rate for the entire experiment, our "*experiment-wise*" Type I Error rate (also known as the familywise error rate, FWER), is *not* α. It could be upwards of *three times* the claimed α level.

Theorem 7.1. *Let the null hypothesis require k tests. Requiring that each sub-test has a Type I Error rate of α/k, also called the per-comparison error rate, controls the experiment-wise error rate to α.*

per-comparison error rate

Proof. For each of the k tests, let E_i be defined as the event of "rejecting the true null hypothesis." Let us reject each of those k tests at a rate of ϵ, the per-comparison error rate. That is, let

$$\mathbb{P}[E_i] = \epsilon, \text{ for all } i.$$

Then, if we require the experiment-wise Type I Error rate to be no greater than α, we have

$$\begin{aligned}
\alpha &= \mathbb{P}\left[\bigcup_{i=1}^{k} E_i\right] \\
&\leq \sum_{i=1}^{k} \mathbb{P}[E_i] \\
&= \sum_{i=1}^{k} \epsilon \\
&= k\epsilon
\end{aligned}$$

Solving for the per-comparison error rate gives $\epsilon = \alpha/k$. □

> ***Note***: This adjustment method was devised by Edward Paulson (1952) and Olive Dunn (1958) and named after Italian mathematician Carlo Emilio Bonferroni whose inequalities made the original proofs possible.

The Bonferroni Correction is a method used to address the issue of inflated Type I Error rates caused by multiple testing in statistics. Its strength is its ease of use. Its weakness is its high level of conservatism. To use the Bonferroni method, merely divide your stated α level by the number of tests, k. (*Equivalently*, you can multiply your p-values by k.) Thus, if you are performing the 15 pairwise tests of the opening example, you would reject any null hypotheses that produced a p-value of less than $0.05/15 = 0.0033$.

conservative

power loss

Again, the Bonferroni correction is conservative; that is, the true experiment-wise Type I Error rate is never more than α, and will most likely be less. How-

ever, unless we know the relationships between the tests and hypotheses, it is our best upper-bound.[1]

Since this method was created, many other statisticians have developed multiple testing methods based on assumptions of the relationships between the tests and hypotheses. We will cover some of these in Section 7.5.

The only sure way to avoid multiple testing issues is to perform only one test; that is, create a single test statistic with a single distribution. Thus, much of the early 20th Century was spent creating a procedure to compare multiple group means by way of a single test. Ronald A. Fisher (1918) finally created the analysis of variance procedure (ANOVA), which achieves this goal.

7.2: Analysis of Variance

The extension of the t-test was developed by Ronald A. Fisher (1918, 1921, 1925). At its simplest, the analysis of variance procedure is merely an extension of the t-test. Recall from Section 6.1, where we formulated the t-test comparing the means of two populations. The test statistic was similar to

t-test

$$t = \frac{\overline{x} - \overline{y}}{\text{standard error}}$$

which has a t_ν distribution, where ν is the number of degrees of freedom. If we want to compare three populations, it would seem very straight-forward to create a test statistic being the sum of the individual pairwise t-statistics. Let us see the difficulty with this test statistic.

Example 7.1: Four varieties of rice are each grown in each of four different fields and their yields are measured (Table 7.1). Are the four varieties essentially the same with respect to yield, or does one variety tend to do much better than the other three?

[1] Actually, the Holm procedure (1979) is always better than the Bonferroni procedure. To perform the Holm procedure, rank the p-values from smallest to largest. Multiply the smallest by k, the second smallest by $k-1$, the third-smallest by $k-2$, etc. Reject all hypotheses where the adjusted p-value is less than α.

Variety	Yields			
A	934	1041	1028	935
B	880	963	924	946
C	987	951	976	840
D	992	1143	1140	1191

Table 7.1: *Yields from four varieties of rice planted in each of four different fields. This* `rice` *data is used in Example 7.2.*

A Really Bad Solution: Let us calculate the t-test statistics for each possible pair of Rice Varieties. Recall that the test statistic for the two-sample t-test is

$$t = \frac{\overline{x} - \overline{y}}{s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$

Using this formula on the Rice data, we have the following test statistics:

Varieties	Test Statistic	Varieties	Test Statistic
A-B	1.6496	B-C	-0.2685
A-C	1.0350	B-D	-4.0308
A-D	-2.5407	C-D	-3.2533

Adding these values together gives –7.4088.

So far, so good. However, if we change the order of the testing—if we change the order of the *nominal* variable —we get the following test statistics:

nominal

Varieties	Test Statistic	Varieties	Test Statistic
A-C	1.0350	C-D	3.2533
A-D	-2.5407	C-B	0.2685
A-B	1.6496	D-B	4.0308

This has a sum of **7.6965**.[2]

[2]Note that the magnitudes of the individual test statistics is the same. They only differ in their signs.

This *is* a problem. The test statistic depends on the ordering of a nominal variable—something without an inherent ordering. This means the test is **very bad**. ◇

Fisher's first solution to this problem was to sum the *squares* of the individual t-statistics. This solves the above issue, since squaring makes the positive/negative issue vanish. Unfortunately, this sum was not standardized. Without standardization, the distribution of the test statistic is more difficult to write. And, without having such a distribution, we cannot know if the test statistic is "big enough" to reject the null hypothesis.

Fisher was able to perform enough adjustments and make enough assumptions. He created a test statistic that was "simple" to write and had a knowable probability distribution.

F distribution

Fisher's test statistic is the ratio of the variance *between* the groups (MSB) to the variances *within* the groups (MSW), a ratio of variances:

$$F = \frac{\text{MSB}}{\text{MSW}} \tag{7.1}$$

The distribution of this test statistic is (conveniently) the F distribution. For this test statistic, note the following two points:

- Larger ratios indicate that the variance *within* the groups is small compared to that between, that the grouping is appropriate, that the groups are not all the same with respect to the measurement.

  dependent

- Smaller ratios indicate that the variance within the groups is large compared to that between, therefore knowing the group tells us little about the expected values within the groups, which also implies that the groups are not significantly different.

  independent

To get a better understanding of the calculations, let us work through an example.

Example 7.2: Four varieties of rice are each grown in each of four different fields and their yields are measured (Table 7.1). Are the four varieties essentially the same with respect to yield, or does one variety tend to do much better than the other three?

non-directional

Solution: Were we only comparing two different varieties of rice, we would use a t-test. Here, however, we are comparing four different varieties. Thus, we will use the analysis of variance method. The null hypothesis is

$$H_0 : \mu_A = \mu_B = \mu_C = \mu_D$$

The alternate hypothesis is that *at least one* of the four rice varieties has a different mean. We can write this as

$$H_A : \mu_A \neq \mu_B \text{ or, } \mu_A \neq \mu_C \text{ or, } \mu_A \neq \mu_D \text{ or,}$$
$$\mu_B \neq \mu_C \text{ or, } \mu_B \neq \mu_D \text{ or, } \mu_C \neq \mu_D$$

or as

$$H_A : \text{At least one mean differs from the others}$$

ANOVA

To test this hypothesis using the analysis of variance method, we first calculate the mean yield for each variety. Knowing the means allows us to calculate the sums of squares for each variety. The mean is calculated as usual, allowing for the added complexity in varying group sizes. The mean measurement in group i is

$$\overline{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{i,j}$$

group means

The four group sample means are $\overline{y}_A = 984.50$, $\overline{y}_B = 928.25$, $\overline{y}_C = 938.50$, and $\overline{y}_D = 1116.50$.

The formula for the sums of squares within each group is

$$\text{SS}_i = \sum_{j=1}^{n_i} \left(y_{i,j} - \overline{y}_i\right)^2$$

SSW

The four sums of squares are $\text{SS}_A = 10085$, $\text{SS}_B = 3868.75$, $\text{SS}_C = 13617$, and $\text{SS}_D = 22305$. Thus, the overall sum of squares within is SSW = 49875.75, the sum of these. As there are $L = 4$ groups and $n = 16$ data points, the number of degrees of freedom are $16 - 4 = 12$. Thus, the total variance within the groups is

$$\text{MSW} = \frac{\text{SSW}}{n-L} = \frac{\sum_j (y_{i,j} - \overline{y}_i)^2}{n-L} = \frac{\sum_i SS_i}{n-L} = \frac{SSW}{n-L} = \frac{49875.75}{12} = 4156.31$$

This is the denominator in our ratio. It measures how much of the variation in the data is *not* explained by the grouping. Smaller values tend to indicate the grouping is more appropriate.

The numerator is the variance *between* the groups. Its calculation is slightly different. First, we start with the sum of squares *between*:

$$\begin{aligned}\text{SSB} &= \frac{1}{\nu}\sum_{i=1}^{n}(y_{i,\cdot} - y_{\cdot,\cdot})^2 \\ &= \sum_{i=1}^{L}\frac{y_{i,\cdot}^2}{n_i} - \frac{y_{\cdot,\cdot}}{\sum_{i=1}^{L} n_i}\end{aligned}$$

Here $y_{i,\cdot}$ is the sum of the yields for variety i, and $y_{\cdot,\cdot}$ is the sum of all yield values. For this data, $y_{A,\cdot} = 3938$, $y_{B,\cdot} = 3713$, $y_{C,\cdot} = 3754$, and $y_{D,\cdot} = 4466$, and $y_{\cdot,\cdot} = 15,871$. Thus, the SSB = 89931.19. The number of degrees of freedom is one less than the number of groups, $L - 1 = 3$. Thus, the variance between the groups is

$$\text{MSB} = \frac{\text{SSB}}{L-1} = \frac{89931.19}{3} = 29977.06$$

This number is the variation *explained by* the grouping. A larger value here (as compared to the residual variation) indicates that the grouping is more appropriate.

explained variation

The test statistic is the ratio of these two variances:

test statistic

$$F = \frac{\text{MSB}}{\text{MSW}} = \frac{29977.06}{4156.31} = 7.21$$

This means that the explained variation is 7.21 times higher than the unexplained variation. A larger number means that the grouping is more appropriate. But, how big is enough to conclude that this reduction is due to the grouping and not just to randomness?

From work by Fisher and George W. Snedecor (1934), this test statistic is distributed according to the F distribution with two types of degrees of freedom: numerator and denominator. The "numerator degrees of freedom" are the degrees of freedom for the MSB: one less than the number of groups, $L - 1$. The "denominator degrees of freedom" are the degrees of freedom for the MSW, the number of data points less the number of groups, $n - L$.

This test will always be a one-tailed test. (Why?) A plot of the distribution, along with the value of the test statistic (F) is provided in Figure 7.1.

non-directional

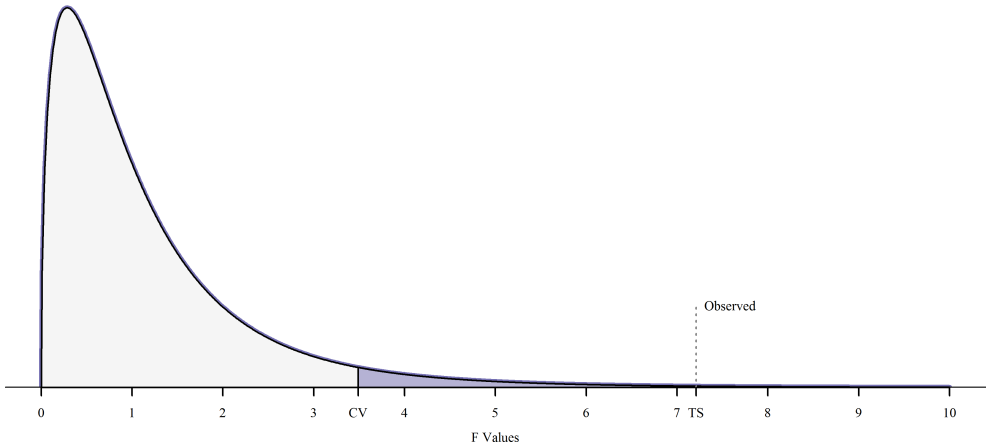

Figure 7.1: *A plot of the distribution of the test statistic in the Rice example. The value of the calculated test statistic is shown on the x-axis. As the test statistic is in the rejection region, we reject the null hypothesis and conclude that there is a difference among the rice varieties.*

With that said, the critical value is

$$F_{0.95,3,12} = \texttt{qf(0.95, df1=3, df2=12)} = 3.490295$$

As our test statistic is greater than the critical value, we reject the null hypothesis at the $\alpha = 0.05$ level and conclude that at least one of the rice varieties is different from the others ($F = 7.21, cv = 3.49$).

Alternatively, as we have access to a computer, we could have calculated the actual p-value for our test statistic. Using the command

p-value

```
pf(7.21, df1=3, df2=12, lower.tail=FALSE)
```

gives us a p-value of $p = 0.00504$. From this, we also conclude that there is a statistically significant difference amongst the four rice varieties ($F = 7.21, \nu_n = 3, \nu_d = 12, p = 0.00504$). ◇

Example 7.3: Looking back to the opening example, an associate of mine claimed that all football conferences scored essentially the same number of points per game on average. Using the data from 2009, let us test her claim.

research hypothesis

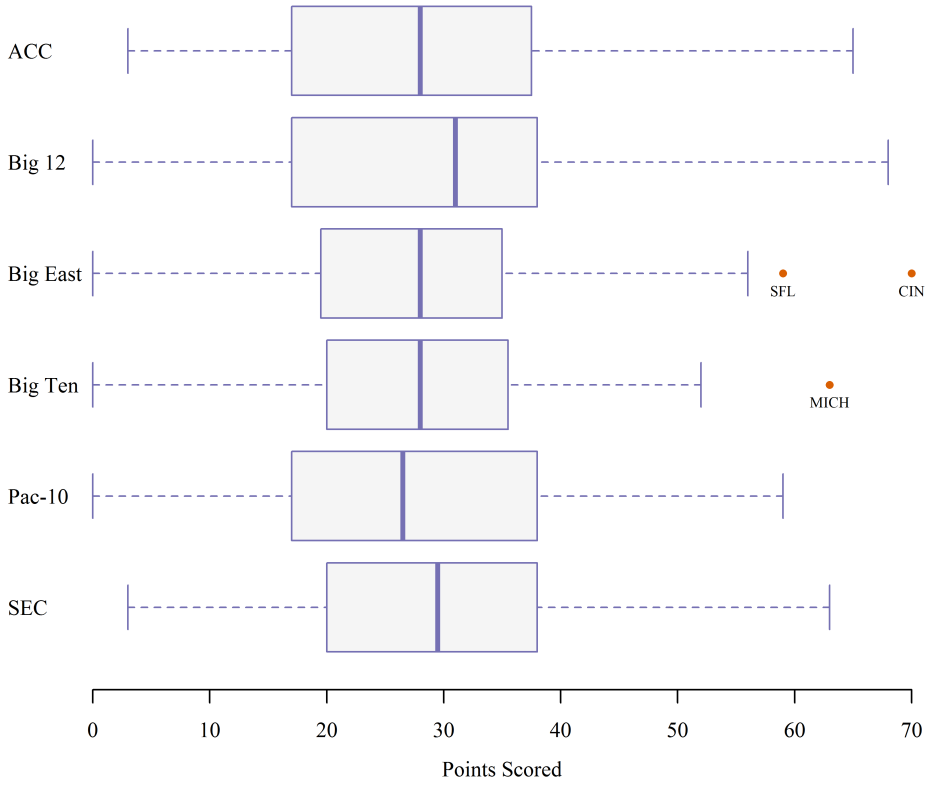

Figure 7.2: *A boxplot of the number of points scored in a football game, compared across the six major conferences. Note that the medians are all approximately equal across the six conferences. As such, we would expect to see no statistically significant difference among the six conferences.*

Solution: Notice that her claim is actually the null hypothesis (it contains the "no difference" condition). As there are six conferences, the null hypothesis is

equal

$$H_0 : \mu_{\text{ACC}} = \mu_{\text{Big East}} = \mu_{\text{Big Ten}} = \mu_{\text{Big 12}} = \mu_{\text{PAC-10}} = \mu_{\text{SEC}}$$

The alternative hypothesis is that at least one of the conferences scored significantly more points than the others.

alternative hypothesis

Let us first examine a boxplot of the data to determine if the null hypothesis seems reasonable. The boxplot in Figure 7.2 strongly suggests that we will find no statistically significant difference in scores between any of the football conferences. Let us perform analysis of variance to determine if this conclusion is actually supported by the data.

ANOVA table

I will leave it as an exercise for you to verify the results summarized in Table 7.2, which is referred to as an ANOVA table . Note that MST (Mean Squared Total) is just the variance of the original data (the scores).

Using the results of the analysis of variance procedure, we can conclude at the $\alpha = 0.05$ level that there is no statistically significant difference among the six NCAA conferences in terms of points scored ($F = 0.6343, \nu_n = 5, \nu_d = 774, p = 0.6736$). ◇

Note: This is equivalent to concluding that knowledge of the conference give no additional information about our best guess for the number of points scored by a given team. In other words, the grouping variable and the dependent variable are independent of each other.

independent

weakness

Note that the null hypothesis was that the (population) means in each group are *equal*. The analysis of variance procedure cannot easily test a hypothesis such as "The Big 12 Conference scored more points, on average, than any other single conference."

$$\begin{aligned} H_R: \quad & \mu_{\text{Big 12}} > \mu_{\text{ACC}} && \text{and} \\ & \mu_{\text{Big 12}} > \mu_{\text{Big East}} && \text{and} \\ & \mu_{\text{Big 12}} > \mu_{\text{Big Ten}} && \text{and} \\ & \mu_{\text{Big 12}} > \mu_{\text{Pac-10}} && \text{and} \\ & \mu_{\text{Big 12}} > \mu_{\text{SEC}} \end{aligned}$$

multiple comparisons

To test *this* hypothesis, we would have to perform multiple comparisons with the Big 12 conference singly compared to each of the other five conferences; that is, we would have to perform five t-tests (or Mann-Whitney tests or permutation tests).

Bonferroni

Additionally, we would also need to perform a Bonferroni adjustment (Section 7.1) since we are performing multiple tests on the same research hypothesis. The number of tests is $k = 5$, thus we would reject the null hypothesis only when the calculated p-value was less than $\frac{\alpha}{k} = \frac{0.05}{5} = 0.01$.

7.2.1 Assumptions

As the analysis of variance procedure grew out of the t-test, the assumptions are the same as for the original t-test: The measurements are Normally distributed in each sub-population; the variances are the

Source	Sum of Squares		ν	Mean Squares		F-statistic	p-value
Between	SSB =	584	5	MSB =	117	0.6343	0.6736
Within	SSW =	142478	774	MSW =	184		
Total	SST =	143061	779	MST =	183		

Table 7.2: *The analysis of variance table for the NCAA 2009 Football data associated with Example 7.3. The large p-value indicates that there is no statistically significant difference among the six conferences in terms of points scored per game.*

same in each sub-population. In other words, the populations are homogeneous *except* for the additive group effect.

This can be symbolized as

additive effect

$$X_{i,j} \sim \mathcal{N}(\mu_j, \sigma^2)$$

Here, $X_{i,j}$ is the ith measure in Group j, and μ_j is the population mean of Group j.

Normality: The first assumption is that of Normality: The measurements within each sub-population is distributed Normally. This assumption can be tested using either graphical or numeric means. The graphical test performed is often the Quantile-Quantile plot.

Q-Q plot

The Quantile-Quantile plot, also known as the Q-Q plot, graphs the observed quantiles of the data against the hypothesized quantiles (of the Normal distribution). If the data are distributed Normally, then the Q-Q plot will consist of points perfectly lined up along the diagonal. No real data will line up perfectly along the diagonal; Normally distributed data should be close, however.

sample v. population
observed v. expected

How 'close' is close? That is a very good question, for which there is no absolute answer. This lack of an answer leads many to shun graphical means and solely use numerical methods for determining Normality. However, I generally avoid numerical tests as they are only powerful for detecting lack-of-Normality when the sample size is large, which is when the Normality assumption is least needed.

eschew

When looking at the Q-Q plots, there are two things to keep in mind: First, concern yourself more about systematic patterns than about random fluctuations. Second, worry more about deviations in the center half of the plot than near the tails. Tails are highly variable by their very nature, so

variability

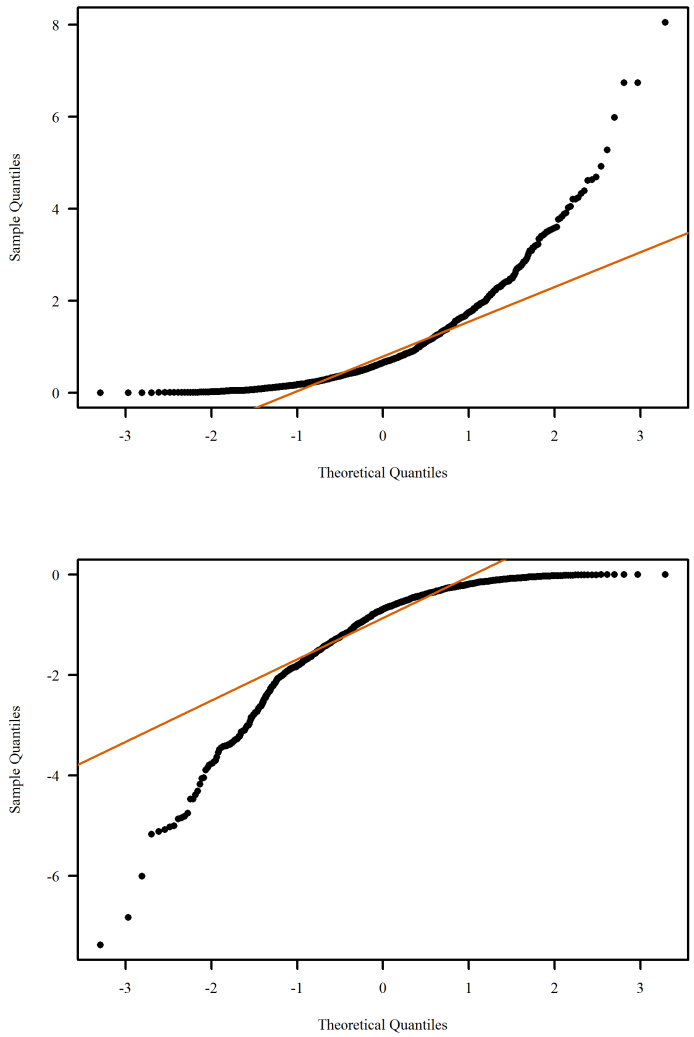

Figure 7.3: *Quantile-Quantile plots of two non-Normally distributed data against the Normal distribution. The top Q-Q plot indicates severe right-skew; the bottom, severe left-skew.*

it is easy to have large deviations in the tails even when the distribution is Normal.

Figure 7.3 contains two examples of Q-Q plots that indicate severe deviation from Normality. The top plot indicates severe right-skew; the bottom plot, severe left skew. In both instances, the severity of the violation indicates that the analysis of variance model you fit is inappropriate for the data.

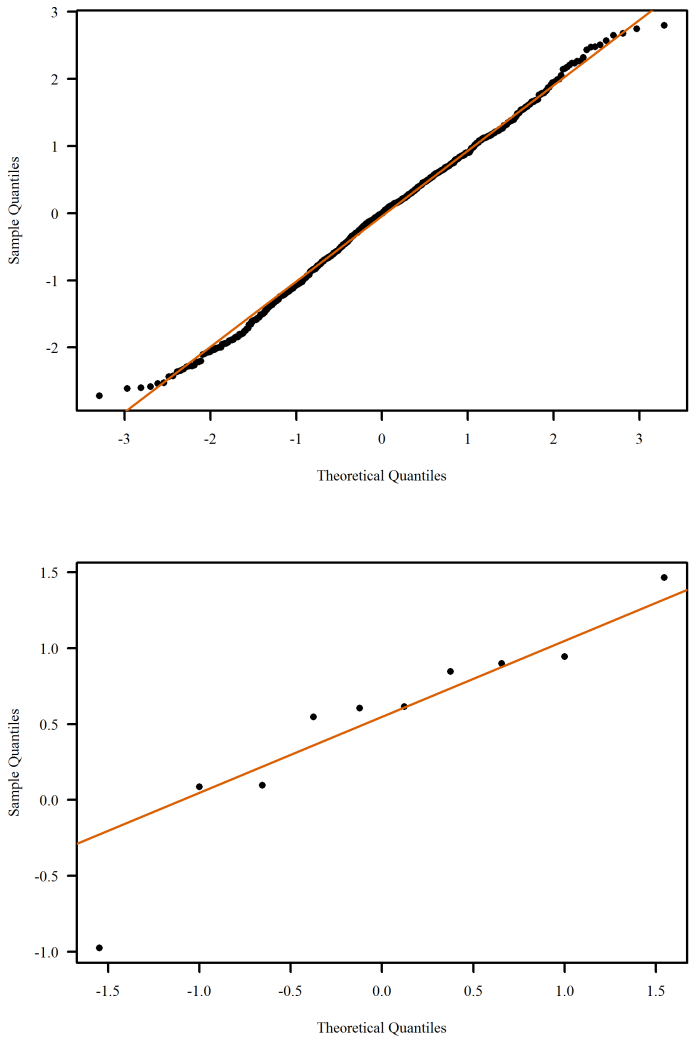

Figure 7.4: *Quantile-Quantile plots of two Normally distributed sets of data: top, $n = 10$; bottom, $n = 1000$. Take note of the variability.*

In contrast, Figure 7.4 shows two Q-Q plots of two sets of data that *are* distributed Normally. The top has a sample size of $n = 1000$; the bottom, $n = 10$. Note two things: First, the variability is much greater in the tails than in the center. Second, smaller sample sizes will produce Q-Q plots which are not necessarily close to the diagonal line even when they come from a Normal distribution.

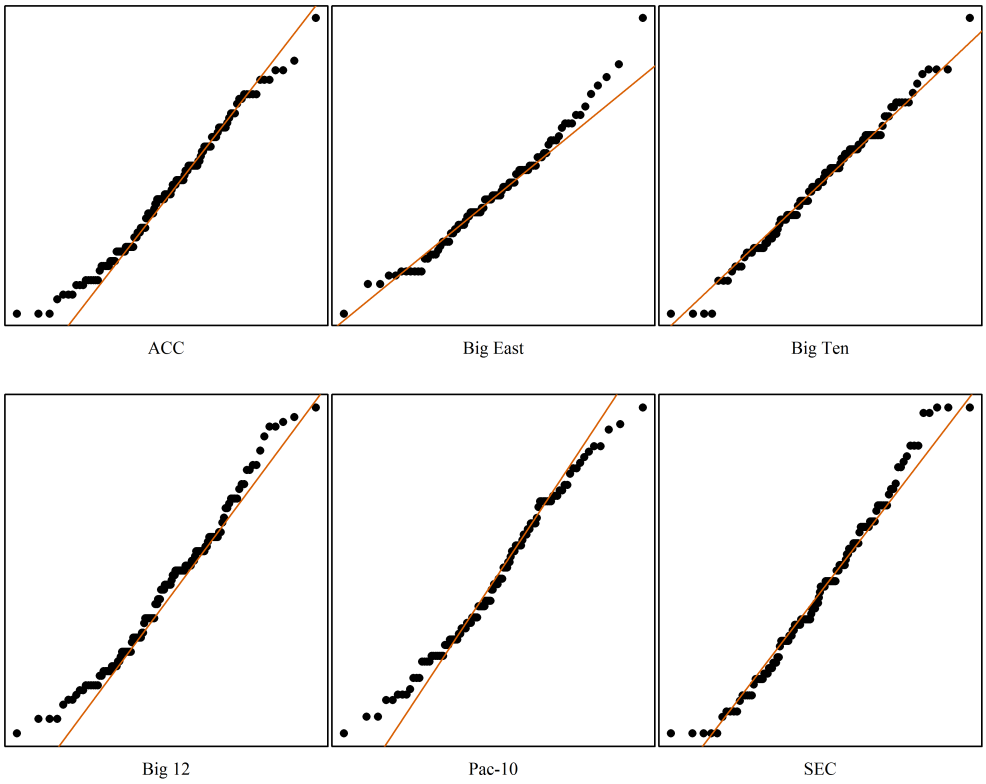

Figure 7.5: *Quantile-Quantile plots of the team scores in the six conferences. There does not appear to be any systematic deviation from the diagonal lines in any of the plots.*

Remember, the assumption is that each sub-population is distributed Normally. As such, you need to perform the test on *each* group. In R, let us parse the scores of the Pac-10 using

```
scorePAC10 = score[conference=="Pac-10"]
```

With that, the command to produce a Q-Q plot is

```
qqnorm(scorePAC10)
```

The Q-Q plots for each of the six conferences are provided in Figure 7.5. Note that none of the six are perfectly Normal, but there does not appear to be any systematic deviation from the diagonal line.

Alternatively, we can use numerical methods to test the Normality of the residuals. Two popular tests are the Kolmogorov-Smirnov test (Massey 1951) and the Shapiro-Wilk test (1965). The K-S test is a general test that compares two specified distributions; the Shapiro-Wilk test is a custom-made test of Normality. As such, if you *must* use a numerical measure of Normality, this is the one I recommend.

Shapiro-Wilk test

Again, remember that the assumption is that the measurements in each sub-population are distributed Normal. As such, you need to perform the test on *each* group. Thus, the R command to use (for the Pac-10) would be

```
shapiro.test(scorePAC10)
```

Performing the Shapiro-Wilk test gives us p-values of ACC, 0.09; Big East, 0.21; Big Ten, 0.63; Big 12, 0.02; Pac-10, 0.09; and SEC, 0.04. Note that two of the conference tests fail at the $\alpha = 0.05$ level. *However*, remember the discussion about multiple testing at the beginning of this chapter (*v.s.*, Section 7.5). Using the Bonferroni correction indicates that none of the groups are sufficiently non-Normal to cause concern (compare the p-values with $\alpha/6 = 0.00833$).

Bonferroni

Equal-Variance: The second assumption we need to test is that of equal variances across the groups. As with the Normality tests, there are graphical tests and numerical tests. The graphical test of choice is the box-and-whiskers plot (Figure 7.2). Looking at the box-and-whiskers plot, we are not struck by any conference having much more (or less) spread than any other. In fact, the six conferences look quite similar in terms of distributions (including spread).

The numerical tests include the F-test (useful only for two groups) and the Bartlett test (1937). Performing the Bartlett test indicates that the variances across the six groups are not statistically different ($K2 = 8.304, \nu = 5, p = 0.1403$). Thus, we fail to reject the null hypothesis of different variances and conclude that the model and data do not violate the equal-variance assumption.

Bartlett test

equal-variances

In R, the command is

```
bartlett.test(score ~ conference, data=fb)
```

The tilde ('~') is the character that separates the dependent variable (`score`) from the independent variable/grouping variable (`conference`). This one line performs the Bartlett test comparing all six groups. As it performs a

formula

single test, there is no need to adjust the p-values using the Bonferroni adjustment.

Fligner-Killeen test

A second (and frequently better) test to run is the Fligner-Killeen test. This test is superior to the Bartlett test as it does not require the underlying distribution to be Normal. As such, if the Shapiro-Wilk test (*v.s.*, Section 7.2.1) is in the grey region, it will not affect this test of equal variances. According to the Fligner-Killeen test, there is also no compelling reason to reject the null hypothesis of equal variances ($X2 = 9.4431, \nu = 5, p = 0.093$).

In R, the command is

```
fligner.test(score ~ conference, data=fb)
```

Thus, we can conclude that the assumptions of analysis of variance are not violated here. As such, we can be confident in the results of the analysis of variance test: We conclude that there is no statistically significant difference in terms of points scored among the six major NCAA conferences in 2009 ($F = 0.6343, \nu_n = 5, \nu_d = 774, p = 0.6736$).

independent

Note: Again, this is equivalent to concluding that the grouping variable and the dependent variable are independent.

7.3: Non-Parametric Means Tests I

Let us perform the analysis of variance procedure for a different hypothesis and a different set of data.

Example 7.4: Last week, a professor I know made the statement that Africa is more poor (has a lower average GDP per capita) than each of the other regions of the world. This seems to be the common wisdom. However, it is true? Does Africa indeed have a significantly lower GDP per capita than the rest of the world?

Solution: To answer this question, one could perform multiple t-tests (or Mann-Whitney tests), suitably adjusting for multiple tests using the Bonferroni correction. Alternatively, one can use analysis of variance.

Bonferroni

Using ANOVA, the first step is to load the data set into memory so that we can perform analysis on it: `read.csv`. Second, let us attach the data so

Source	SSx	ν	MSx	F	$p-value$
Between	1.34×10^{10}	5	2.68×10^{9}	13.154	$\ll 0.0001$
Within	3.53×10^{10}	173	2.04×10^{8}		
Total	4.87×10^{10}	178	2.74×10^{8}		

Table 7.3: *The analysis of variance table produced from the* `gdpcap` *data being fit by the model explaining the State's GDP per capita by the world region. Note that the p-value is much less than our usual $\alpha = 0.05$. As such, we are tempted to reject the null hypothesis based on this test.*

that we can avoid the '`$`' notation: `attach`. Next, let us perform the analysis of variance procedure on the data: `aov(gdpcap~region)`. Finally, we summarize the results: `summary`. From this, we have the ANOVA table in Table 7.3.

Note that the p-value is much less than our usual α-level of $\alpha = 0.05$. From this, we would like to conclude that there is a statistically significant different among the six world regions — at least one of the six regions is *difference* with respect to GDP per capita than the others. **p-value**

Let us check the assumptions of the analysis of variance procedure. First, a side-by-side box-and-whiskers plot of the data: Figure 7.6 compares the box-and-whiskers plots of the six regions of the world. Recall that we need to test for Normality and for equal variances across the six regions. The box-and-whiskers plot suggests both a lack of Normality and a lack of equal variances across the six world regions. The vast number of outliers (signified by solid dots) is not consistent with the assumption of Normality. **assumption**

The asymmetry in several of the regions (Africa, Eastern, and Islamic) is also inconsistent with the assumption of Normality. Finally, the spread of the Eastern region is much larger than that of the Africa region. From this, we must conclude that the two assumptions of the analysis of variance procedure are violated in this data and model. **skew**

If we feel more comfortable with numerical tests, then we can use the Shapiro-Wilk test of Normality and the Fligner-Killeen test of equal variances. The Shapiro-Wilk test indicates severe departure from Normality: Africa has a p-value $\ll 0.0001$; Eastern, $p = 0.0007$; Islamic, $p \ll 0.0001$; Latin America, $p = 0.49$; Other, $p = 0.002$; and Western, $p = 0.014$. The Fligner-Killeen test indicates that the variances are not equal across the six regions ($X2 = 98.84, \nu = 5, p \ll 0.0001$). From these tests, we can conclude $p < \alpha \rightarrow$ **fail**

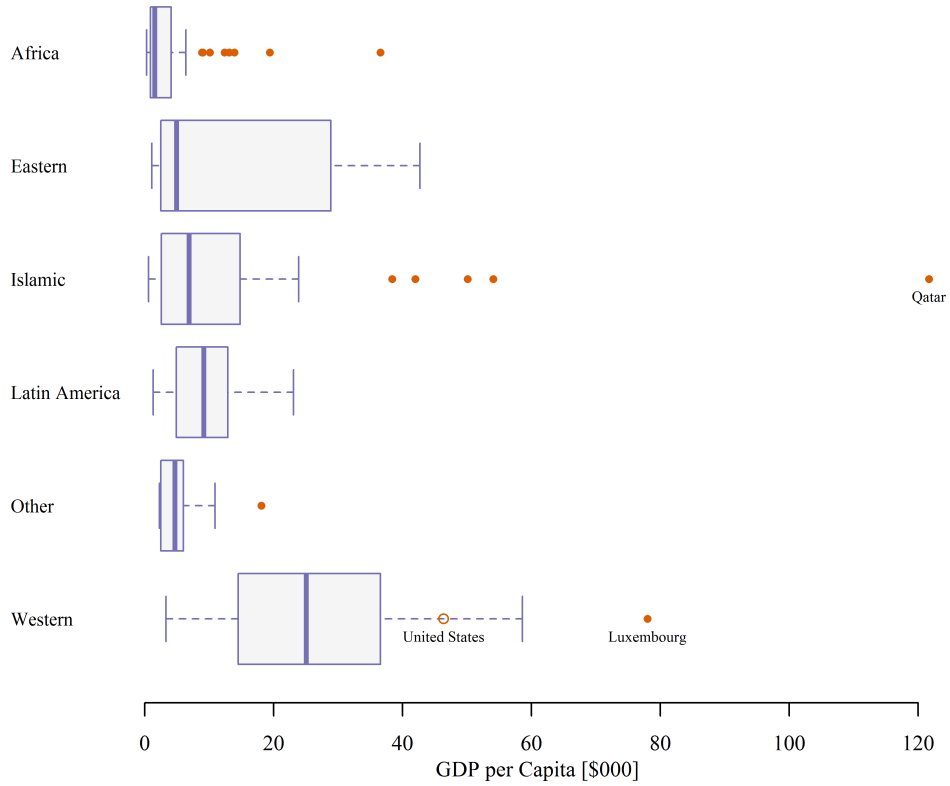

Figure 7.6: *A box-and-whiskers plot of the GDPs per capita for each of the six regions. Note the presence of many outliers and the lack of symmetry in the plots — both indicators of lack of Normality.*

that the analysis of variance procedure is *not* appropriate for this model and data. ◇

What do we do to answer the original question? We have two options. First, we can transform the data so that the assumptions are not violated (*v.i.*, Chapter 14). Second, we can perform a non-parametric method, which I do here.

7.3.1 Another Rank-Sum Test

Recall that in Chapters 5 and 6 we discussed two non-parametric methods that allow us to statistically compare the mean of one population to a hypothesized mean (Wilcoxon test) and to sta-

tistically compare the means of two populations (Mann-Whitney test). The idea behind both tests was to rank the data, count the ranks above zero or the ranks corresponding to one group, then compare that test statistic to a distribution to get the p-value. The Kruskal-Wallis rank-sum test (1952) uses the same idea. However, as there are $L > 2$ groups instead of 1 or 2, the calculation is more difficult.

The assumption of the Kruskal-Wallis test is that each group has the same distribution *except for* the mean. This assumption is less restrictive than that of the analysis of variance procedure which required the additional assumption of Normality. As such, the Kruskal-Wallis test is more general, it is also less powerful — it fails to reject too often. As it is less powerful, we will want to use analysis of variance when possible.

Kruskal-Wallis test

power

The Kruskal-Wallis test is also robust to violations of its assumptions. This means that even when the groups do not have the same distribution, we can use the Kruskal-Wallis test as long as the differences are not "too big" and as long as we are not "slaves to α."

Performing the Kruskal-Wallis test in R is quite easy. If our analysis of variance command was

ANOVA

```
aov(gdpcap ~ region)
```

then, our Kruskal-Wallis test command is

```
kruskal.test(gdpcap ~ region)
```

Using this command, we can conclude that the six world regions are not the same with respect to average GDP per capita ($X2 = 79.96, \nu = 5, p \ll 0.0001$). While this is the same conclusion as that of the analysis of variance test, we are more confident in these conclusions as they are not based on the faulty assumption of Normality.

p-value

> ***Note***: Both tests assume the distributions of the sub-populations *are the same* except for the center. This is not supported by our tests. However, the Kruskal-Wallis test is still better for this data than the analysis of variance test because it does not make the additional faulty assumption of Normality.

> ***Note***: The Kruskal-Wallis test is a non-parametric test. As such, it has less power than the analysis of variance test. This means that it will fail to

non-parametric test

reject a false null hypothesis more often than will the analysis of variance test; that is, the Type II Error rate is higher. Here, the null hypothesis was rejected. This increases our confidence that the null hypothesis *should* be rejected.

7.4: Non-Parametric Means Tests II*

In Section 5.6 (page 122), we introduced the permutation test. In Section 6.5 (page 154), we expanded it to two groups. The logic behind the permutation test is that the measurements in each group come from the same distribution under the null hypothesis. Thus, permuting the observed values amongst the groups changes nothing. This idea extends to more than two groups. However, because the number of permutations becomes large quickly, the randomization test version is usually used.

permutation test

randomization test

The function to use, `permKS()`, is from the same package as before (`perm`).

> ***Note***: As an aside, the "KS" stands for "k-sample," just as the "TS" from the previous chapter stands for "two-sample."

This function requires two slots: the measurement variable and the grouping variable. Thus, a command to perform a randomization test on the `rice` dataset is just

```
permKS(yield,variety)
```

This gives the following output:

```
    K-Sample Exact Permutation Test Estimated by Monte Carlo

data:  yield and variety
p-value = 0.006

p-value estimated from 999 Monte Carlo replications
99 percent confidence interval on p-value:
 0.001080589 0.014099183
```

As usual, the null hypothesis is that the four groups have the same mean. The p-value above is estimated to be $p = 0.006$. While it may be tempting to interpret the confidence interval as a confidence interval for the difference in mean yields, it is not. It is a 99% confidence interval for the *p-value*. The output tells us that we are 99% confident that the true p-value is between 0.001 and 0.014. As this interval does not contain values above the usual $\alpha = 0.05$, we can conclude that the true p-value is less than $\alpha = 0.05$.

null hypothesis

The analysis above used the default estimation method, the exact Monte Carlo method. We can specify that the function use that method, or we can

specify the method be based on the central limit theorem for permutations. These two lines do these two methods, respectively.

```
permKS(yield,variety, method="exact.mc")
permKS(yield,variety, method="pclt")
```

The method based on the central limit theorem for permutations provides just the estimated p-value, not a confidence interval for it.

Using a permutation test on the GDPs per capita across the six world regions gives output

```
    K-Sample Exact Permutation Test Estimated by Monte Carlo

data:  gdpcap and region
p-value = 0.001

p-value estimated from 999 Monte Carlo replications
99 percent confidence interval on p-value:
 0.000000000 0.005289582
```

Again, as the 99% confidence interval for the p-value is entirely less than our usual $\alpha = 0.05$, we firmly reject the null hypothesis that the regions all have the same average GDP per capita in favor of at least one differs from the others.

7.5: Post-Hoc Testing

Thus far, we have only been able to test whether several populations have different means based on the samples measured. The test we should use depends on whether and which assumptions are met. If the populations are Normally distributed and have the same variance, we use the analysis of variance test. If the populations have the same distribution except for the mean, we use the Kruskal-Wallis test. If none of these assumptions are met, we use the permutation test.

know more

However, we often want to know more than just that there *is* a difference. We want to know *which* group is different from the others. For instance, *Which* rice variety has the significantly different yield (see Example 7.2) or *Does* Africa have a lower GDP per capita than the other regions (see Example 7.4)

Our temptation is to perform pairwise t-tests (or Mann-Whitney tests or permutation tests) on all possible pairs and use the Bonferroni adjustment of Section 7.1 to correct the p-values for the experiment-wise error rate. However, the Bonferroni adjustment is very conservative; that is, it rejects *less often* than it should. As such, it is not the best answer.

Ronald Fisher, in the early 20th Century, hypothesized that if the analysis of variance test rejected the null hypothesis of no difference, then the follow-up tests would *not* have to be adjusted — they were "protected" tests.

pairwise tests

Figure 7.7: *Sir Ronald A. Fisher, FRS*

Further analysis showed that Fisher was not entirely correct (one of the few times). However, his non-adjustment produced experiment-wise error rates closer to α than did the Bonferroni correction. Regardless, it was still not good enough for Fisher (or the many who followed). Today, solving this problem of an inflated Type I Error rate is a rich area of statistical research.

7.5.1 Fisher's LSD Test

One of the earliest adjustments to the Bonferroni method was created by Ronald A. Fisher himself, one of the statistical luminaries of the 20th Century. The Least Significant Difference test (Fisher 1948) attempts to protect the experiment error rate by simultaneously performing all t-tests at once, but using a common variance measure (the MSW). Doing this allows us to calculate a least difference (LSD) corresponding to significantly different groups; that is, if two group means differ by at least the LSD then the group means are significantly different.

LSD test

significant difference

That a *single* number is produced and that this number corresponds to a minimum distance between non-different groups are the strengths of this test. Unfortunately, of the multiple comparison tests we discuss here, it is the least protective of the experiment-wise error rate.

protection

Fisher determined that for two groups to be significantly different, their difference must be at least

$$\text{LSD} := t_{\alpha/2}\sqrt{\frac{2\cdot \text{MSW}}{n}}$$

Here, $t_{\alpha/2}$ has degrees of freedom equal to that of the MSW, $L(n-1)$, where L is the number of groups and n is the number of observations in each group.

Thus, for the Rice Yield example (7.2), assuming $\alpha = 0.05$, with MSW = 4156.31, $L = 4$, and $n = 4$, we have

$$\text{LSD} = 2.178813\sqrt{\frac{2 \cdot 4156.31}{4}} = 99.33$$

Thus, when two means differ by 99.33 we are 95% confident that the two populations have statistically different means. Referring to the means we calculated in Example 7.2, we see that Variety D is significantly different from the other three varieties, but that none of the other three are significantly different from each other.

agricolae

Note: We can get the same results using the `R` function `LSD.test()` from the `agricolae` package. Its first parameter is the analysis of variance model; the second, the name of the grouping variable (`"variety"`). Thus, the above analysis is done using the command

```
print( LSD.test(mod, "variety") )
```

experiment-wise error rate

Note: While it is true that this procedure does not protect the experiment-wise error rate as well as the following tests, it does a good enough job if you have already rejected the null hypothesis that all of the means are equal (Carmer and Swanson 1973).

7.5.2 Tukey's Range Test

John W. Tukey (Tukey 1949, Kramer 1956) improved upon Fisher's LSD test above by using a different measure of variance. Whereas Fisher used the variance, Tukey used an adjusted variance corresponding to a different distribution.

Figure 7.8: *John W. Tukey, ForMemRS*

Tukey's Range Test test statistic, a.k.a. the *Honestly* Significant Difference (HSD) test statistic, is

$$W = q_\alpha(t, \nu)\sqrt{\frac{\text{MSW}}{n}}$$

Studentized Range

Here, $q_\alpha(t, \nu)$ is the critical value corresponding to the distribution named the "Studentized Range distribution."

An advantage of the HSD procedure is that it also provides a single number with which we can compare differences in sample means as did Fisher's LSD test. A second advantage is that Tukey's HSD test controls the experiment-wise error rate better than does the LSD test.

A disadvantage is that Tukey's HSD test uses a different distribution, one that is only 'well-known' in this context. For those using tables to determine the appropriate critical values, this matters; for the rest of us using computers, it does not. Except for this expense of calculation, the HSD test is superior to the LSD in all ways.

Studentized Range

According to Tukey's HSD procedure, the "honestly" significant difference for our rice example is

$$W = 4.19866\sqrt{\frac{4156.312}{4}} = 135.3427$$

This means that there needs to be a difference of 135.3427 between two means before they are considered significantly different. In the rice example, we see that the yield of Variety D is significantly better than that of Varieties B and C, but it is not significantly better than Variety A.

Note: We can get these results using `R`'s `TukeyHSD()` function. For the above analysis, the command to run is `TukeyHSD(mod)`. This command gives a table of differences in the means. The p-value column (`p adj`) gives the p-value associated with the difference in means against the null hypothesis that the means are not different. Thus, a p-value less than α indicates the means are significantly different.

This function also includes a function that plots the confidence intervals for the differences in means between each pair of levels. Just surround the above function call with a plot function. That is, run this line `plot(TukeyHSD(mod))`

Note: In addition to the `TukeyHSD()` function, we can get the same results using the `HSD.test()` function from the `agricolae` package. It requires two parameters, however. Its first parameter is the analysis of variance model; the second, the name of the grouping variable (here, `"variety"`).

ANOVA

Thus, the line of code to run to perform Tukey's HSD test is just `print( HSD.test(mod, "variety") )`. Note that this function gives a

lot more information than the `TukeyHSD()` function. It, alas, does not have a plotting option.

Note: This procedure is also based on calculating a single number that serves as the difference between statistically different and not. Other procedures, like Duncan's Multiple Range Test, create test statistics based on how far apart the means are in terms of ranks. Thus, Variety B and Variety C (adjacent ranks) would have a different "HSD" than would Variety B and Variety D (three ranks apart) when using Duncan's test.

Note: Other parametric multiple testing procedures used in `R` are the Student-Newman-Keuls (SNK) test (`SNK.test`), Duncan's new multiple range test (`duncan.test`), and the Waller-Duncan test (`waller.test`), among *many* others. In short, if you read about a multiple testing procedure, it is probably already a function in `R`.

7.5.3 Kruskal's Multiple Range Test

ANOVA

Tukey's HSD procedure, as well as Fisher's LSD procedure and most other multiple testing procedures (including Duncan's test and Scheffé's test) are based on the same assumptions as the analysis of variance procedure. Thus, if you cannot use analysis of variance, these tests will not work. You will have to use a non-parametric multiple-testing method such as the Kruskal method (Conover 1999).

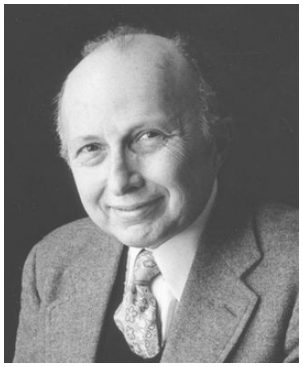

Figure 7.9: *Bill Kruskal*

Performing the Kruskal multiple comparison's method in `R` is simple. The command for the rice data would be

```
kruskal(gdpcap,region)
```

Note that the first parameter is the measurement and the second is the grouping variable (a.k.a. the treatment). Performing this test on the `gdpcap` data gives us that the Western Region has a significantly higher GDP per capita than the other regions, Africa has a lower GDP per capita than the other regions, and the rest are not significantly different in terms of GDP per capita.

Note: The `kruskal()` function also requires the `agricolae` package.

7.6: Further Examples

To further illustrate some of these processes, this section provides several additional examples.

Example 7.5: This example comes from Fisher's original introduction to the analysis of variance procedure (Fisher 1925). On Page 195, he presents data and asks the following question:

> In an experiment on the accuracy of counting soil bacteria, a soil sample was divided into four parallel samples, and from each of these after dilution seven plates were inoculated. The number of colonies on each plate is shown below. Do the results from the four samples agree within the limits of random sampling? In other words, is the whole set of 28 values homogeneous, or is there any perceptible intraclass correlation?

In other words, does the mean number of colonies differ across the four parallel samples? Let us answer his question.

Solution: The datafile `fisher38` contains the data. The two variables are the number of colonies observed and the sample number. We would like to test if the mean number of colonies observed is the same for the four populations. To do this, we would like to use the analysis of variance procedure as it is more powerful than our other options. This test makes two assumptions: the colony count in each population is Normally distributed and the variance of the colony count in each population is the same.

Let us use the Fligner-Killeen test to determine if the second assumption is reasonable. According to that test, the assumption is reasonable ($p = 0.8469$). The Shapiro-Wilk test also indicates the Normality assumption is reasonable ($p_{\min} = 0.062$). Thus, as the assumptions are met, we can use the analysis of variance test.

According to the analysis of variance test, there is no significant evidence that the average number of colonies varies across the four samples ($p = 0.669$). In effect, the number of colonies is independent of the sample number. The box-and-whiskers plot supports this conclusion (Figure 7.10).

This conclusion makes perfect sense knowing the source of the data: Fisher pulled the four samples from the same soil sample.

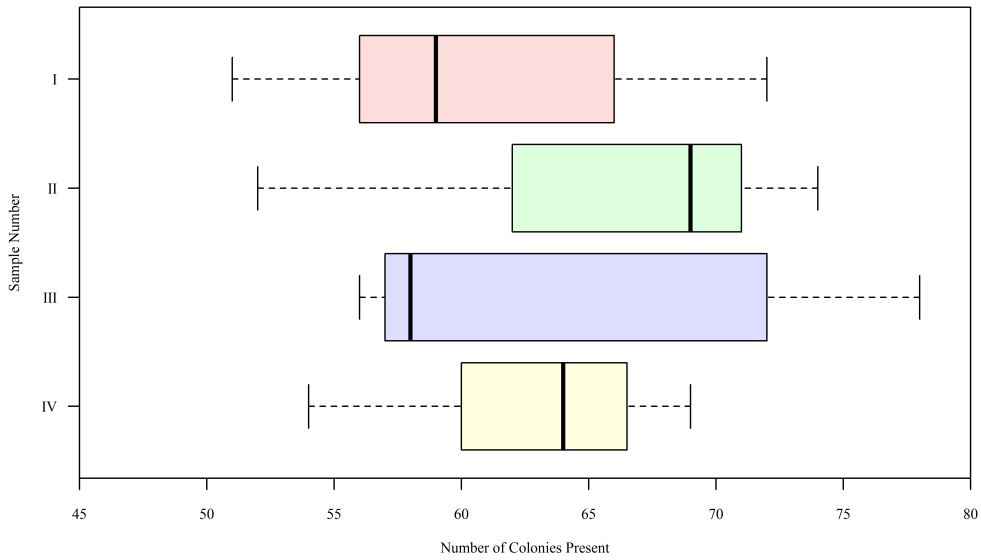

Figure 7.10: *A box-and-whiskers plot of the colony count across the four samples. Note that the medial lines vary wildly, but so too do the actual measurements.*

> ***Note***: One can see that the median (mean) lines vary wildly, there is also great variation in the measurements in each group. It is this latter fact that means we cannot conclude the means are different. This is an important point! We do not conclude the means are the same. We just *cannot detect a difference in the means*.

◇

Example 7.6: At its most general, a biome is a region with similar climactic conditions. There are many ways of categorizing the Earth's land into biomes. One common method was proposed by the World Wide Fund for Nature (WWF). The WWF biome scheme consists of 14 different biomes. These range from the Mangrove biome (subtropical and tropical land inundated by salt water) and Taiga (subarctic, humid land), to Tundra (artic land) and Xeric shrubland (temperate to tropical arid land).

Fires are expensive, both their fighting and their prevention. To reduce resources spent, many jurisdictions are thinking about taking the biome into consideration when planning future fires. It makes sense that different

biomes would have different fire probabilities, but do they? Furthermore, which biomes are associated with faster fire return intervals?

To answer these questions, a researcher randomly sampled 30 areas in the United States. These 30 samples consisted of five biomes with mean fire return intervals ranging from three years to 1000 years.

Let us answer the questions.

Solution: The datafile `biome2` contains the data. The two variables of interest are the biome and the mean fire return interval (`mfri`). We would like to test if the average mfri differs across the five biomes and (if so) which has the lowest mean time between fires (`mfri`). To do this, we would like to use the analysis of variance procedure as it is more powerful than our other options. This test makes two assumptions: the mean fire return interval in each biome (population) is Normally distributed and the variance of the mean fire return intervals in each biome (population) is the same.

Let us use the Fligner-Killeen test to determine if the second assumption is reasonable. According to that test, the assumption is not reasonable ($p = 0.0019 < 0.05 = \alpha$). Because the data fails this test, there is no need to test the Normality assumption; both the analysis of variance test and the Kruskal-Wallis test requires that the distributions be identical except for the medians. Such is not the case here. As such, we will use the permutation test.

According to the exact Monte Carlo permutation test, there is strong evidence that the average mean fire return rate varies across the five biomes ($p = 0.001$).

At least one of the five is not like the others, but which? There is no simple one-line test to answer this question when the equal-variance assumption is not met. Thus, we will need to perform pairwise permutation tests and adjust the p-values using the Bonferroni adjustment. This is not perfect, but it is the best we have. From those $\binom{5}{2} = 10$ tests, we have that the subtropical rainforest (STR) biome significantly differs from all others in terms of the average mean fire return interval. We cannot, however, detect a difference among the other four biomes. The box-and-whiskers plot only partially supports this conclusion (Figure 7.11).

Note: In effect, we can only conclude that the subtropical rainforest biome differs in the mean fire return interval. We cannot conclude that the other four biomes have the same average mean fire return interval. Looking at the box-and-whiskers plot (Figure 7.11) illustrates this. While it *looks* as

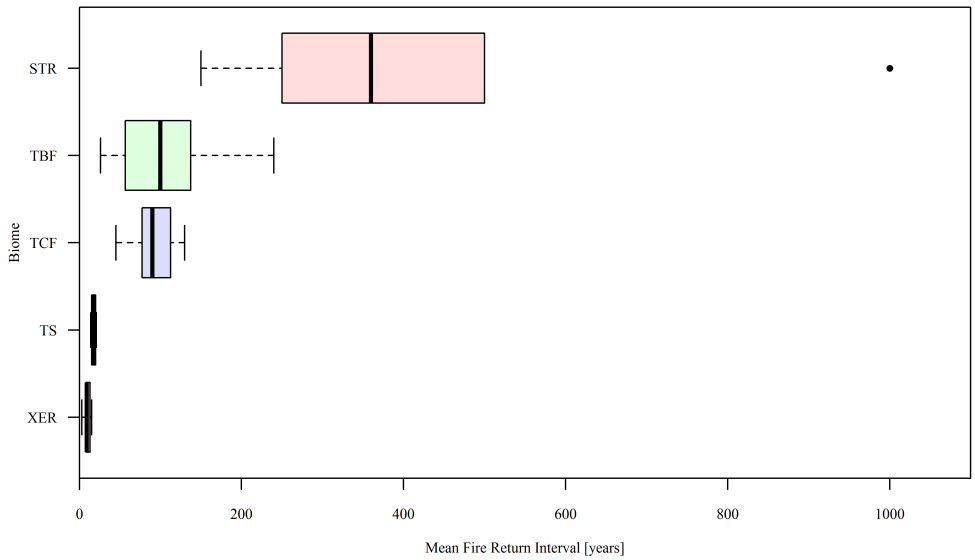

Figure 7.11: *A box-and-whiskers plot of the mean fire return interval across the five biomes. Note that it appears as though the TS and XER biomes differ from the TBF and TCF biomes. However, the permutation test was not powerful enough to detect the difference.*

though the xeric shrubland (XER) and the tree savanna (TS) have similar average mean fire return intervals that differ greatly from those of the temperate broadleaf and coniferous forests (TBF and TCF), we cannot detect a difference using this method.

An Alternative to the Permutation Test* This shows that permutation tests are of low power. It would have been better to transform the data so that we could have used the analysis of variance test. Perhaps a logarithm transformation on the mean fire return intervals would make the population variances sufficiently similar to allow us to avoid using the permutation tests.

I leave it as an exercise for you to transform all mean fire return intervals with the logarithm function, then determine if the variances are sufficiently the same to pass the Fligner-Killeen test (it is). The logarithm transform also produces log-mean fire return intervals that are sufficiently Normal. Thus, we *can* use the analysis of variance test on the logged mean fire return interval values.

Here is the code to accomplish this analysis:

```
lmfri = log(mfri)

fligner.test(lmfri~biome)
mod2 = aov(lmfri~biome)
shapiro.test( residuals(mod2) )

summary(mod2)
TukeyHSD(mod2)
```

The first line performs the logarithm transformation of the mean fire return interval variable. The second line performs the Fligner-Killeen test. Since the data passed this test, we fit it with the model and test the model's residuals for Normality. This is a quicker way to test the Normality assumption. We could have performed five Shapiro-Wilk tests, one for each biome, but we would have had to perform a Bonferroni adjustment. This way is faster and better in terms of being less conservative.

This analysis shows that the temperate broadleaf and coniferous forests are not significantly different in terms of their average mean fire return interval and that the xeric shrubland and tree savanna biomes are not significantly different in their average mean fire return interval. All other pairs of biomes *do* significantly differ in their average mean fire return intervals. The last two lines perform the hypothesis tests. The first tests if there is a difference among the five biomes. The second determines which are different.

This conclusion feels better, as it seems to better agree with the box-and-whiskers plot (Figure 7.11).

◇

Example 7.7: The Cold War ran from the late 1940s until the early 1990s. During that period, Europe was divided into three groups: Those states who were members of NATO, those who were members of the Warsaw Pact, and those who were members of neither. NATO members were allied with the United States; Warsaw Pact members; the Soviet Union. The neutral states were officially non-allied. After the fall of the Soviet Union in 1992, the Warsaw Pact ceased to exist and several new states came into being.

In 2010, the European Union polled citizens in its member states and asked them if they thought the United States was, overall, a positive force in the world or a negative force in the world. A researcher thought that there

may be lingering dislike of the United States based on the Cold War divisions, so she decided to test if there was a significant difference in the perception of the United States across the three alliance types named above.

Solution: To test this, the data must be properly divided into the three groups, and the three groups must be delineated, especially as there are many new countries that did not exist during the Cold War. The rule is to place the current country based on what it was and where it was during the Cold War. It is also to ignore Germany. With those rules, there were 11 European NATO members: Belgium, Denmark, France, Greece, Italy, Luxembourg, Portugal, The Netherlands, Turkey, the United Kingdom, and Spain. There were nine European Warsaw Pact members: Bulgaria, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Romania, and Slovakia. There were eight neutral countries: Austria, Croatia, Cyprus, Finland, Ireland, Malta, Slovenia, and Sweden.

We will also use the proportion of the population viewing the United States as having an overall negative force in the world as our dependent variable. The following three lines puts this data into R:

```
NATO = c(0.70, 0.48, 0.67, 0.88, 0.38, 0.66, 0.61, 0.50,
    0.76, 0.53, 0.54)

WSWP = c(0.39, 0.30, 0.43, 0.43, 0.40, 0.28, 0.38, 0.22,
    0.46)

NEUT = c(0.55, 0.32, 0.84, 0.60, 0.59, 0.65, 0.67, 0.62)
```

We would like to use the analysis of variance test as it is the most powerful test we have. It assumes the measurements are Normally distributed in the populations and that the measurements have equal variances in the populations. Here, we test the first:

```
shapiro.test(NATO)
shapiro.test(WSWP)
shapiro.test(NEUT)
```

All three groups pass the Normality test. We now test the equal-variance assumption. Here is one way:

```
fligner.test( list(NATO,WSWP,NEUT) )
```

The data also pass the Fligner-Killeen test. Thus, we can perform the analysis of variance test.

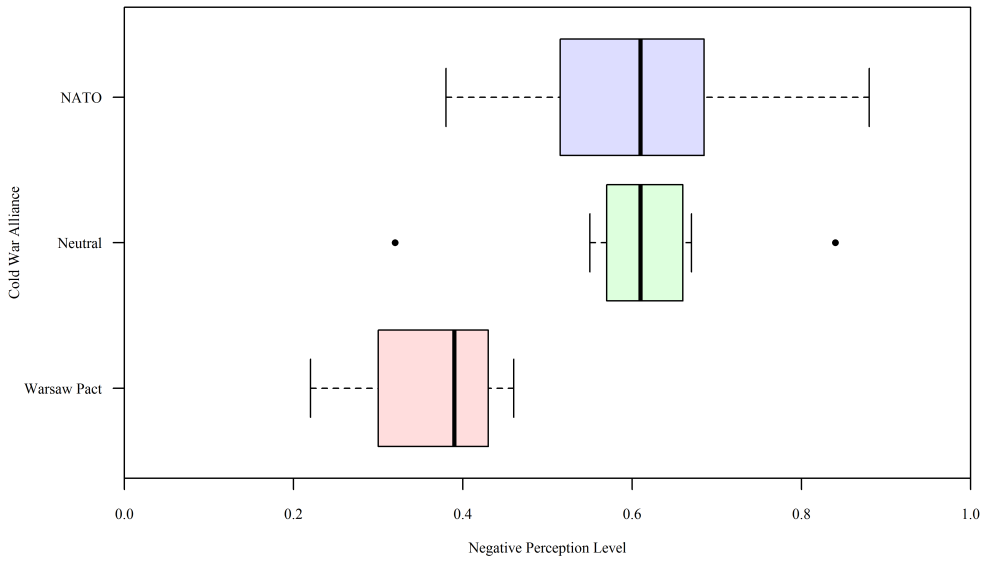

Figure 7.12: *A box-and-whiskers plot of the perception that the United States has a net negative effect on the world across the three Cold War alliances. Note that there appears to be no difference between the NATO group and the neutral group, and that the Warsaw Pact group has a much less negative perception of the United States.*

This gets the data in the right format and performs the analysis of variance test:

```
Negative = c(NATO,WSWP,NEUT)
Alliance = c(rep("NATO",11), rep("WSWP",9), rep("NEUT",8))
Alliance = as.factor(Alliance)

mod = aov( Negative~Alliance )
summary(mod)
```

According to the analysis of variance test, the three means are not the same ($p = 0.00034$). But, which is different? Are there lingering Cold War effects? Do the former members of the Warsaw Pact see the United States much more negatively than the members of the former NATO?

```
TukeyHSD(mod)
```

According to Tukey's HSD test, the answers are the Warsaw Pact, perhaps, and no. While there is no detectable difference in how NATO members and neutral members see the United States, there is a significant difference in how the former Warsaw Pact countries see the US. The former Cold War en-

emies have a significantly *less negative* view of the US than do the other two alliance categories. This conclusion is supported by the bow-and-whiskers plot (Figure 7.12). ◇

7.7: Conclusion

In this chapter, we discovered many methods of comparing the centers of more than two groups. The methods were either parametric (analysis of variance) or non-parametric (Kruskal-Wallis, and permutation). Before we discussed the tests, we explored the multiple testing issue. The first (and easiest) correction method is the Bonferroni adjustment. Its drawback is that it is extremely conservative; it fails to reject more often than it should.

We then discussed several adjustments and improvements to the Bonferroni method: Fisher's LSD method, Tukey's HSD method, and Kruskal's method. The first two (along with the several others mentioned in the text) are based on the same assumptions as the analysis of variance procedure. Thus, if you cannot use this procedure, you should not use these multiple testing methods. The third method is based on the Kruskal-Wallis non-parametric test. As such, it makes the same assumptions of that test and shares its strengths and weaknesses.

7.8: End of Chapter Materials

7.8.1 R Functions

In this chapter, we were introduced to several R functions that will be useful in the future. These are listed here.

Packages:

agricolae This package includes many multiple-testing functions, in addition to functions related to agricultural research.

Statistics:

aov(formula) This performs the analysis of variance procedure on the data using the offered formula. If you do not attach the data, then you will add a `data=` parameter to this function call. Thus, for the rice data, the command that performs analysis of variance would be

```
aov(yield ~ variety, data=rice)
```

Note that the dependent variable is to the left of the tilde, and the grouping variable is (or grouping variable*s* are) to the right.

As the `aov` function returns a lot of information, you will want to store the results in a variable and then use `summary()` to summarize the data. You should also save the `aov` results in a variable as many multiple comparisons test require it as a parameter.

bartlett.test(formula) The Bartlett test determines if the groups have the same variance.

fligner.test(formula) The Fligner-Killeen test determines if the groups have the same variance. This test is most robust against departures from Normality (Conover, Johnson, and Johnson 1981). As such, it tests the equal-variance assumption without the results being dirtied by variations from Normality.

HSD.test(model, g) Tukey's Honestly Significant Difference test also determine which of the groups is significantly different from the others. It is conservative, although not as conservative as Fisher's LSD test. This test requires the `agricolae` package to be loaded, via `library(agricolae)`.

kruskal.test(formula) This performs the Kruskal-Wallis test, which is the non-parametric analogue of the analysis of variance procedure. As with `aov`, the `kruskal.test` function takes an optional `data=` parameter. Thus, if we had not attached the `gdpcap` data, we would use

```
kruskal.test(gdpcap ~ region, data=gdp)
```

to perform the Kruskal-Wallis test.

kruskal(y,x) The Kruskal method performs multiple testing of the several means among the groups. The measurements are `y`, and the groups are `x`. This test makes the same assumptions of the Kruskal-Wallis test (above). This test requires the `agricolae` package to be loaded, via `library(agricolae)`.

ks.test(x,p) The Kolmogorov-Smirnov test calculates the 'distance' between the supplied data and the stated distribution (or a second data set).

LSD.test(model,g) Fisher's Least Significant Difference *post hoc* test determines which of the groups is significantly different from the others in terms of the measurements. It requires that you have already performed the analysis of variance procedure on the data and that you supply the name of the grouping variable. The LSD test is more conservative than Tukey's HSD test, but not as conservative as the Bonferroni adjustment. This test requires the `agricolae` package to be loaded, via `library(agricolae)`.

shapiro.test(x) The Shapiro-Wilk test is used to quantify the degree of Normality in a group of data. The null hypothesis is that the data is Normally distributed. Thus, a p-value greater than $\alpha = 0.05$ signifies that there is not enough evidence to conclude the data is *not* Normally distributed.

TukeyHSD(model) Tukey's Honestly Significant Difference test also determine which of the groups is significantly different from the others. It is conservative, although not as conservative as Fisher's LSD test.

Probability:

pf(x) The F distribution is the probability distribution tied to the analysis of variance test. in `R`, to calculate the probability of observing a specific value, `x`, from an F distribution with degrees of freedom `df1` and `df2`,

you would use `pf(x, df1, df2, lower.tail=FALSE)`. The lower tail parameter specifies that you want the probability of getting a value more extreme than the `x`; the default is `lower.tail=TRUE`.

Graphics:

qqnorm(d) One of the graphical techniques used to determine normality of residuals is the Q-Q Plot. In `R`, the `qqplot` function is actually a more general tool that plots the quantiles of any two sets of numbers against each other. The `qqnorm` is dedicated to plotting residuals against the Normal distribution to determine Normality.

qqline(d) Unfortunately, the `qqnorm()` command does not add a reference line to the quantile-quantile plot. To add this line, you will need the `qqline(d)` command.

Programming:

library() Most analysis in `R` can be done using the basic functions. There are times, however, when additional functionality is needed or desired. Hundreds of statisticians and computer scientists have created packages of functions that extend the abilities of `R`. To use these packages, one must both have them installed in their `R` folder and loaded into memory. The former can be done with the command

```
utils:::menuInstallPkgs()
```

or through the provided menu:

```
Packages | Install Package(s)...
```

Once the package is in the `R` folder, you activate it using the `library()` command.

7.8.2 Exercises and Extensions

This section offers suggestions on things you can practice from this chapter. Save the scripts in your Chapter 7 folder. For each of the following problems, please save the associated R script in the chapter folder as ext0x.R, where x is the problem number.

Summary:

1. How is the analysis of variance procedure an extension of the two-sample t-test? How is it different?
2. What are three characteristics of the F distribution? How many parameters does it take? What are those parameters?
3. What is the strength of the Bonferroni adjustment? What is the weakness? Why did Fisher and Tukey develop other adjustment methods?
4. What does the R-function qp give?
5. Why should a researcher prefer the analysis of variance procedure over the Kruskal-Wallis test? Why might a researcher still need to use the Kruskal-Wallis test?
6. Why should one not use the Kruskal-Wallis test *and* Fisher's LSD test?

Data:

7. The crime datafile consists of several variables. Those of interest to this question are census9 (region of the US) and vcrime90 (the violent crime rate in 1990). Are the nine regions of the country the same with respect to crime rate? If not, rank the regions from highest to lowest according to statistically significant differences in violent crime rate. Again, make sure you perform the appropriate checks of assumptions and that you answer the question appropriately and include appropriate graphs to support your conclusions. Include an appropriate box-and-whiskers plot.
8. Redo the previous problem using census4 in lieu of census9. The new variable divides the country into just four regions. Again, make sure you perform the appropriate checks of assumptions and that you answer the question appropriately and include appropriate graphs to

support your conclusions. Include an appropriate box-and-whiskers plot.

9. Redo the previous problem using `vcrime00`, the violent crime rate in 2000. Again, make sure you perform the appropriate checks of assumptions and that you answer the question appropriately and include appropriate graphs to support your conclusions. Include an appropriate box-and-whiskers plot.

10. According to the `crime` datafile, does the 1990 property crime rate (`pcrime90`) differ across the four census regions? If so, which region has the highest average property crime rate in 1990? Provide an appropriate graphic to illustrate your findings.

11. Redo the previous problem. This time, use the property crime rate in 2000 (`pcrime00`) in lieu of that in 1990.

12. The `crime` datafile also contains information about the use of the citizen's initiative during the 1990s (`inituse`). According to the data, do the four census regions use the citizen's initiative at a different rate? If so, which region tends to use it most? Make sure you provide necessary evidence, explanation, and graphics.

13. Continuing our use of the `crime` datafile, do the four census regions have the same average level of conservatism (`conserve`)? If not, which region is most conservative? Which region is least conservative? As always, provide necessary evidence, explanation, and graphics.

14. According to the `crime` datafile, do the four census regions tend to have the same proportion of time that the two branches of government are held by the same party (`unifGOVT`)? As always, provide necessary evidence, explanation, and graphics.

15. Again, according to the `crime` datafile, do the four census regions tend to have the same level of professionalism in the state legislature, as compared to that of the US Congress (`profleg`)? As always, provide necessary evidence, explanation, and graphics.

Monte Carlo:

16. One of the assumptions of the analysis of variance procedure is that the several groups are Normally distributed. Let us see how much we can relax that assumption. Create three variables with the same mean, variance, and distribution, but with that distribution being non-Normal. Use Monte Carlo to determine if the p-values from the analysis of variance test are Uniformly distributed.

7.8.3 Applied Research

This section offers some applied research works that are connected with the topics in this chapter.

- Liam Downey and Brian Hawkins. (2008) "Race, Income, and Environmental Inequality in the United States." *Sociological Perspectives* **51**(4): 759–81.
- Rebecca C. Fauth, Tama Leventhal, and Jeanne Brooks-Gunn. (2008) "Seven Years Later: Effects of a Neighborhood Mobility Program on Poor Black and Latino Adults' Well-being." *Journal of Health and Social Behavior* **49**(2): 119–30.
- Kishore S. Gawande, Pravin Krishna, and Marcelo Olarreaga. (2009) "What Governments Maximize and Why: The View from Trade." *International Organization* **63**(3): 491–531.
- Kenneth L. Leonard, Melkiory C. Masatu, and Alexandre Vialou. (2007) "Getting Doctors to Do Their Best: The Roles of Ability and Motivation in Health Care Quality." *The Journal of Human Resources* **42**(3): 682–700.
- Linda D. Molm, David R. Schaefer, and Jessica L. Collett. (2007) "The Value of Reciprocity." *Social Psychology Quarterly* **70**(2): 199–217.
- Garrick L. Percival, Martin Johnson, and Max Neiman. (2009) "Representation and Local Policy: Relating County-Level Public Opinion to Policy Outputs." *Political Research Quarterly* **62**(1): 164–77.
- Markus H. Schafer. (2011) "Ambiguity, Religion, and Relational Context: Competing Influences on Moral Attitudes?" *Sociological Perspectives* **54**(1): 59–82.
- Robb Willer, Ko Kuwabara, and Michael W. Macy. (2009) "The False Enforcement of Unpopular Norms." *American Journal of Sociology* **115** (2): 451–90.

7.8.4 References and Additional Readings

This section provides a list of statistical works. Those works cited in the chapter are here. Also here are works that complement the chapter's topics.

- M. S. Bartlett. (1937) "Properties of sufficiency and statistical tests." *Proceedings of the Royal Statistical Society Series A*. **160**(901): 268–82.
- S. G. Carmer and M. R. Swanson. (1973) "An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods." *Journal of the American Statistical Association*. **68**(341): 66–74.
- Indra M. Chakravarti, Radha G. Laha, and J. Roy. (1967) *Handbook of Methods of Applied Statistics*, Volume I. New York: John Wiley and Sons.
- William J. Conover. (1999) *Practical Nonparametric Statistics*. New York: John Wiley and Sons.
- William J. Conover, Mark E. Johnson, and Myrle M. Johnson. (1981) "A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data." *Technometrics*. **23**(4):351–61.
- R. D'Agostino and M. Stephens. (1986) *Goodness-of-Fit Techniques*. New York: Marcel Dekker.
- Olive Jean Dunn. (1958) "Estimation of the Means of Dependent Variables." *The Annals of Mathematical Statistics*. **29**(4): 1095–1111.
- Ronald A. Fisher. (1918) "The Correlation Between Relatives on the Supposition of Mendelian Inheritance." *Philosophical Transactions of the Royal Society of Edinburgh*. **52**(Part I):399–433.
- Ronald A. Fisher. (1921) "On the 'Probable Error' of a Coefficient of Correlation Deduced from a Small Sample." *Metron*. **1**(4):3–32.
- Ronald A. Fisher. (1925) *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.
- Ronald A. Fisher. (1948) *The Design of Experiments*. Edinburgh: Oliver and Boyd.
- Rudolf J. Freund and William J. Wilson. (2003) *Statistical Methods*, Second Edition. New York: Academic Press.

- Sture Holm. (1979) "A simple sequentially rejective multiple test procedure." *Scandinavian Journal of Statistics* **6**(2): 65–70.

- Andrei N. Kolmogorov. (1933) "*Sulla Determinazione Empirica di una Legge di Distributione*." *Giornale dell'Istituto Italiano degli Attuari*. **4**(1): 1–11.

- Clyde Y. Kramer. (1956) "Extension of multiple range tests to group means with unequal numbers of replications." *Biometrics* **12**(3): 307–310.

- William Kruskal and W. Allen Wallis. (1952)"Use of ranks in one-criterion variance analysis." *Journal of the American Statistical Association*. **47**(260): 583–621.

- Frank J. Massey, Jr. (1951) "The Kolmogorov-Smirnov Test for Goodness of Fit." *Journal of the American Statistical Association*. **46**(253): 68-78.

- Edward Paulson. (1952) "On the Comparison of Several Experimental Categories with a Control." *The Annals of Mathematical Statistics*. **23**(2): 239–246.

- Samuel S. Shapiro and Martin B. Wilk. (1965) "An analysis of variance test for normality (complete samples)." *Biometrika*. **52**(3–4): 591–611.

- Sidney Siegel and N. John Castellan, Jr. (1988) *Nonparametric Statistics for the Behavioral Sciences*, Second Edition. New York: McGraw-Hill.

- Nicolai V. Smirnov. (1939) "*Sur les écarts de la Courbe de Distribution Empirique*." *Recueil Mathématique (Matematicheskii Sbornik)*. **6**(48): 3–26.

- George W. Snedecor. (1934) *Calculation and Interpretation of Analysis of Variance and Covariance*. Boston: Collegiate Press.

- George W. Snedecor and William G. Cochran. (1989) *Statistical Methods*, Eighth Edition. Ames, IA: Iowa State University Press.

- John W. Tukey. (1949) "Comparing Individual Means in the Analysis of Variance." *Biometrics* **5**(2): 99–114.