## Contents

### 12 Comparing Two Means
- 12.1 The Distinction Between Paired T and 2-sample T ........................................... 1
- 12.2 Comparing Two Means in R ................................................................. 2
- 12.3 Formulas for 2-Sample T ........................................................................ 3
- 12.4 Where do these formulas come from? .................................................... 6
- 12.5 Exercises ....................................................................................... 10

### 13 Handling Violations of Assumptions
- 13.1 Diagnosis .................................................................................. 1
- 13.2 Prescription ............................................................................. 2

### 14 Designing Experiments
- 14.1 Power Calculations ....................................................................... 1

### 15 Comparing More Than Two Means Using ANOVA
- 15.1 The Basic ANOVA situation .............................................................. 1
- 15.2 Confidence Intervals for One Mean At a Time .................................. 5
- 15.3 Pairwise Comparison ...................................................................... 5
- 15.4 Example 15.4 ............................................................................... 8
- 15.5 Example 15.6: Walking Sticks ......................................................... 9

### 16 Correlation
- ........................................................................................................ 1

### 17 Regression
- 17.1 The Simple Linear Regression Model ............................................. 1
- 17.2 The Least Squares Method ............................................................... 1
- 17.3 Making Predictions ....................................................................... 2
- 17.4 How Good Are Our Predictions? .................................................... 3
- 17.5 Checking Assumptions .................................................................... 6
12.1 The Distinction Between Paired T and 2-sample T

- A paired t-test or interval compares two quantitative variables measure on the same observational units.
  - Data: two quantitative variables
  - Example: each swimmer swims in both water and syrup and we compare the speeds. We record speed in water (quantitative) and speed in syrup (quantitative) for each swimmer.

- A paired t-test is really just a 1-sample t-test after we take our two measurements and combine them into one, typically by taking the difference (most common) or the ratio.

- A 2-sample t-test or interval looks at one quantitative variable in two populations.
  - Data: one quantitative variable and one categorical variable (telling which group the subject is in).
  - Example: Some swimmers swim in water, some swim in syrup. We record the speed (quantitative) and what they swam in (categorical) for each swimmer.

An advantage of the paired design (when it is available) is the potential to reduce the amount of variability by comparing individuals to themselves rather than comparing individuals to individuals in an independent sample.

Paired tests can also be used for “matched pairs designs” where the observational units are pairs of people, animals, etc. For example, we could use a matched pairs design with married couples as observational units. We would have one quantitative variable for the husband and one for the wife.

In a 2-sample situation, we need two independent samples (or as close as we can get to that).
12.2 Comparing Two Means in R

12.2.1 Using R for Paired T

```r
> react <- read.csv('http://www.calvin.edu/~rpruim/data/calvin/react.csv')
> # Is there improvement from first try to second try with the dominant hand?
> t.test(react$dom1 - react$dom2)

One Sample t-test

data: react$dom1 - react$dom2
t = -1.6024, df = 19, p-value = 0.1256
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
   -2.3062108  0.3062108
sample estimates:
    mean of x
        -1

> # This gives slightly nicer labeling of the otherwise identical output
> t.test(react$dom1, react$dom2, paired=TRUE)

Paired t-test

data: react$dom1 and react$dom2
t = -1.6024, df = 19, p-value = 0.1256
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
   -2.3062108  0.3062108
sample estimates:
mean of the differences
        -1
```

12.2.2 Using R for 2-Sample T

```r
> # Is there improvement from first try to second try?
> t.test(dom1 ~ Sex.of.Subject, react)

Welch Two Sample t-test

data: dom1 by Sex.of.Subject
t = -0.5098, df = 17.998, p-value = 0.6164
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
   -4.096683  2.496683
sample estimates:
mean in group female  mean in group male
          30.0           30.8
```
Comparing Two Means

> # Is the first try with the dominant hand better if you get to do the other hand first?
> t.test(dom1 ~ Hand.Pattern, react)

Welch Two Sample t-test

data: dom1 by Hand.Pattern
t = -1.0936, df = 16.521, p-value = 0.2898
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-4.642056 1.477221
sample estimates:
mean in group dominant first mean in group other first
29.84615 31.42857

12.3 Formulas for 2-Sample T

12.3.1 Standard Error

The standard error formula for the (Welch) 2-sample t is

\[ SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \]

This should look familiar. If you think of it as

\[ SE = \sqrt{\left(\frac{s_1}{\sqrt{n_1}}\right)^2 + \left(\frac{s_2}{\sqrt{n_2}}\right)^2} \]

it is just a “Pythagorean” combination of the standard error formulas for the two samples put together. See Section 12.4 for more information about this Pythagorean identity.

12.3.2 Degrees of Freedom

If \( \sigma_1 \) and \( \sigma_2 \) are known, then the distribution of \( \bar{X} = \bar{Y} \) is normal and we can work out the mean and standard deviation using the rules in Section 12.4. When we substitute in \( s_1 \) and \( s_2 \), we have to work with t-distributions that are only approximately correct. Furthermore, the degrees of freedom is given by the following rather messy formula

\[ df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\left(\frac{s_1^2}{n_1-1}\right) + \frac{s_2^2}{n_2-1}} \]

We won’t ever compute this by hand, but you will see these values in R output.

You should know two important inequalities, however:

- \( df \geq \min(n_1 - 1, n_2 - 1) \)
- \( df \leq (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2 \)

We can use these bounds to give approximations when a computer is unavailable.
12.3.3 Confidence Intervals and Hypothesis Tests

The formulas are now just what we would expect. The confidence interval for the difference in the two means is given by

\[ x - y \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \]

And the test statistic for hypothesis tests is:

\[ t = \frac{x - y - (\text{hypothesized difference})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \]

12.3.4 Welch vs. Pooled

The method outlined above is sometime called the Welch 2-sample t. There is another version of the 2-sample t (emphasized in the book) which makes the assumption that the standard deviations are the same in the two populations. This requires us to estimate a single standard deviation from both groups. This estimate is sometimes called a “pooled” estimate because it pools the data from both samples.

The only slight advantage to the pooled method is that the formula for degrees of freedom is easier, so it was simpler to do on a hand-held calculator. (The formula for standard error is more complicated than ours because of the pooled estimate of the standard deviation, but it is still doable.)

The Welch two sample t has several advantages over the pooled variance approach, so we recommend that you always use the Welch method (which is the default in R).

1. If you have access to a computer, it is equally easy to do.
2. The method works even if the two groups have different standard deviations. The pooled method does not work well in those situations.
3. When the two standard deviations are the same, the pooled method is only very slightly better, so there is essentially no downside to the Welch method.
4. There are no particularly good ways to check whether the two standard deviations are the same. (This is one reason we skipped over some of that material in Chapter 11.)

12.3.5 Example 12.3: Horned Lizards

Here are some solutions to Example 12.3 (read it in the book).

First, we’ll do it by hand:
Comparing Two Means

> summary( horn.length ~ group, data=HornedLizards, fun=favstats)

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>median</th>
<th>IQR</th>
<th>mean</th>
<th>sd</th>
<th>var</th>
</tr>
</thead>
<tbody>
<tr>
<td>group</td>
<td>killed</td>
<td>30</td>
<td>22.25</td>
<td>2.700</td>
<td>21.987</td>
<td>2.709</td>
</tr>
<tr>
<td></td>
<td>living</td>
<td>154</td>
<td>24.55</td>
<td>3.000</td>
<td>24.281</td>
<td>2.631</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td>184</td>
<td>24.15</td>
<td>3.325</td>
<td>23.907</td>
<td>2.769</td>
</tr>
</tbody>
</table>

From this we get

\[
SE = \sqrt{\frac{2.71^2}{30} + \frac{2.63^2}{154}} = 0.5383
\]

So a 95% confidence interval for the difference in mean horn length between killed and living lizards can be calculated as follows.

> se <- sqrt( 2.71^2 /30 + 2.63^2/154 ); se
> t.star <- qt( .975, 29 ); t.star  # this will be a touch larger than from the messy formula
> me <- t.star* se; me;
> diff.in.means <- 21.987 - 24.281; diff.in.means
> diff.in.means - t.star * se  # lower bound of CI
> diff.in.means + t.star * se  # upper bound of CI

We can test the null hypotheses

- \( H_0 : \mu_{killed} - \mu_{living} = 0 \)
- \( H_a : \mu_{killed} - \mu_{living} \neq 0 \)

by computing the test statistic

\[
t = \frac{\bar{x} - \bar{y} - (\text{hypothesized difference})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{21.987 - 24.281}{0.5382} = -4.263
\]

and the p-value

> t <- (21.987 - 24.281 )/se; t
> 2 * pt( t, df=29 ) # this p-value will be slightly too large
> 2 * pt( t, df=29 + 153 ) # this p-value will be slightly too small
Or we can let R do all the work:

```r
> t.test( horn.length ~ group, data=HornedLizards )

Welch Two Sample t-test
data: horn.length by group
t = -4.2634, df = 40.372, p-value = 0.0001178
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -3.381912  -1.207092
sample estimates:
mean in group killed mean in group living
   21.98667     24.28117
```

Here is the pooled variance version (which should match what is in the book, but which we don’t recommend using):

```r
> t.test( horn.length ~ group, data=HornedLizards, var.equal=TRUE )

Two Sample t-test
data: horn.length by group
t = -4.3494, df = 182, p-value = 2.27e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -3.335402  -1.253602
sample estimates:
mean in group killed mean in group living
   21.98667     24.28117
```

### 12.4 Where do these formulas come from?

The formulas in chapters 10 – 12 all come from the same key facts.

If $X$ and $Y$ are random variables and $a$ is a number, then

- $\text{Mean}(aX) = a \text{Mean}(X)$
- $\text{Mean}(X \pm Y) = \text{Mean}(X) \pm \text{Mean}(Y)$
- $\text{Var}(aX) = a^2 \text{Var}(X)$
  - $\text{SD}(aX) = |a| \text{SD}(X)$
- $\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y)$, provided $X$ and $Y$ are independent.
  - $\text{SD}(X \pm Y) = \sqrt{\text{SD}(X)^2 + \text{SD}(Y)^2}$

Furthermore, if $X$ and $Y$ are normally distributed and independent, then $aX + b$ and $X + Y$ are also normally distributed.
Example: Alice and Bob

Suppose Alice and Bob play a game where there scores are largely independent of each other (as in bowling, or darts, where players alternate and don’t interfere with the other person’s play). Furthermore, we know that their performance is approximately normally distributed with

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>150</td>
<td>15</td>
</tr>
<tr>
<td>Bob</td>
<td>145</td>
<td>20</td>
</tr>
</tbody>
</table>

1. What is the probability that Alice scores at least 170?

   ```r
   > 1 - pnorm(170, mean=150, sd=15)
   [1] 0.09121122
   ```

2. What is the probability that Bob scores at least 170?

   ```r
   > 1 - pnorm(170, mean=145, sd=20)
   [1] 0.1056498
   ```

3. If Alice plays two games and adds her scores, what is the probability that she scores at least 300 (for the two games together)?

   ```r
   > 1 - pnorm(300, mean=150+150, sd=sqrt(15^2 + 15^2))
   [1] 0.5
   ```

   Alternatively, just notice that the mean is 300, so the probability must be 50%.

4. If Bob plays two games and adds his scores, what is the probability that he scores at least 300 (for the two games together)?

   ```r
   > 1 - pnorm(300, mean=145+145, sd=sqrt(20^2 + 20^2))
   [1] 0.3618368
   ```

5. If they are a team, what is the probability that their total score is at least 300?

   ```r
   > 1 - pnorm(300, mean=150+150+150 - 145-145-145, sd=sqrt(15^2+15^2+15^2 + 20^2+20^2+20^2))
   [1] 0.4207403
   ```

6. What is the probability that Alice beats Bob in a single game?

   The trick here is to think about $A - B$. When $A - B$ is positive, Alice wins.

   ```r
   > 1 - pnorm(0, mean=150-145, sd=sqrt(15^2 + 20^2))
   [1] 0.5792597
   ```

7. What is the probability that Alice beats Bob in a match where they play 3 games and add the scores?

   This time, think about $A_1 + A_2 + A_3 - B_1 - B_2 - B_3$.

   ```r
   > 1 - pnorm(0, mean=150+150+150 - 145-145-145, sd=sqrt(15^2+15^2+15^2 + 20^2+20^2+20^2))
   [1] 0.6354828
   ```
Example: Sample Mean

The sample mean is \( \bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \). If we have a random sample from a population with mean \( \mu \) and standard deviation \( \sigma \), then

- Mean \( (X_1 + X_2 + \cdots + X_n) = \mu + \mu + \cdots + \mu = n\mu \)
- So Mean \( \left( \frac{X_1 + X_2 + \cdots + X_n}{n} \right) = \frac{n\mu}{n} = \mu \)
- Var \( (X_1 + X_2 + \cdots + X_n) = \sigma^2 + \sigma^2 + \cdots + \sigma^2 = n\sigma^2 \)
- So Var \( \left( \frac{X_1 + X_2 + \cdots + X_n}{n} \right) = \frac{n\sigma^2}{n^2} \)
- and SD \( \left( \frac{X_1 + X_2 + \cdots + X_n}{n} \right) = \sqrt{\frac{n\sigma^2}{n^2}} = \frac{\sigma}{\sqrt{n}}. \)

Example: Binomial

The results for the binomial distribution are similar using the fact that if \( X_i \) is the number of success in a single trial with success probability \( p \), then

- Mean \( (X_i) = p \)
- Var \( (X_i) = p(1 - p) \)
- SD \( (X_i) = \sqrt{p(1 - p)} \)

So for \( n \) trials we get \( X = X_1 + X_2 + \cdots + X_n \), and

- Mean \( (X) = p + p + \cdots + p = np \)
- Var \( (X) = p(1 - p) + p(1 - p) + \cdots + p(1 - p) = n(p(1 - p)) \)
- SD \( (X) = \sqrt{np(1 - p)} \)

Example: 2-Sample t

Suppose we have two populations that are normal with means \( \mu_1 \) and \( \mu_2 \) and standard deviations \( \sigma_1 \) and \( \sigma_2 \). Let \( \bar{X} \) and \( \bar{Y} \) be the sample means for group 1 and group 2, using samples of size \( n_1 \) and \( n_2 \), then

- Mean \( (\bar{X}) = \mu_1 \)
- SD \( (\bar{X}) = \frac{\sigma_1}{\sqrt{n_1}} \)
- Mean \( (\bar{Y}) = \mu_2 \)

Last Modified: May 5, 2011
• $SD(Y) = \frac{\sigma_2}{\sqrt{n_2}}$

• $Mean(X - Y) = \mu_1 - \mu_2$

• $SE(X - Y) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
12.10 Comparing Two Means

12.5 Exercises

1. $A$ and $B$ are independent normal random variables with means and standard deviations as indicated in the table below.

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td>$B$</td>
<td>18</td>
<td>4</td>
</tr>
</tbody>
</table>

a) What is are the mean and standard deviation of $3A$?
b) What is are the mean and standard deviation of $2B$?
c) What is are the mean and standard deviation of $A + B$?
d) What is are the mean and standard deviation of $A - B$?
e) What is are the mean and standard deviation of $3A - 2B$?

2. $A$ and $B$ are independent normal random variables with means and standard deviations as indicated in the table below.

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td>$B$</td>
<td>18</td>
<td>4</td>
</tr>
</tbody>
</table>

Calculate the following probabilities.

a) $\Pr(A \geq 20)$
 b) $\Pr(B \geq 20)$
 c) $\Pr(A \leq 10)$
 d) $\Pr(A + B \geq 30)$
 e) $\Pr(B - A \geq 0)$
 f) $\Pr(B - A \geq 8)$

3. The Quincy Quintuplets like to play an online game called Quest. All five of them have scores that are normally distributed with a mean of 123 and a standard deviation of 45. They like to compete as a team. The team score is the sum of the scores for the five team members.

a) What is the probability that the Quincy Quintuplets score at least 500?
b) What is the probability that the Quincy Quintuplets score at least 600?
c) What is the probability that the Quincy Quintuplets score at less than 400?
When they are feeling mean, the Quincy Quintuplets like to play an online game called MeanQuest. MeanQuest is harder, and all five of them have scores that are normally distributed with a mean of 87 and a standard deviation of 15. They like to compete as a team. In MeanQuest the team score is the mean of the scores for the five team members.

a) What is the probability that Quinn (the youngest of the quintuplets) scores more than 100?

b) What is the probability that the Quincy Quintuplets score more than 100 as a team?

c) What is the probability that Quinn (the youngest of the quintuplets) scores less than 50?

d) What is the probability that the Quincy Quintuplets score less than 50 as a team?
The \( t \)-procedures (1-sample \( t \), 2-sample \( t \), and paired \( t \)) all rely on an assumption that the distribution of the variable(s) is normal in the population.

13.1 Diagnosis

13.1.1 Histograms

Histograms of our data can reveal potential problems in the population distribution, but

1. It is sometimes hard to judge the shape of a histogram by eye.
2. When data sets are small, we don’t have very much information about the population.

The `xhistogram()` function can add an overlay of a normal distribution to help with issue 1:

\[
\text{> xhistogram( } \sim \text{ biomass.ratio, MarineReserve, density=TRUE, n=8 )}
\]

But normal quantile plots work even better.
13.1.2 Normal Quantile Plots

Normal quantile plots address the first problem of histograms. (We can’t really do anything about the second problem.)

A normal quantile plot is a scatter plot comparing the values in our data set to the values we would expect to see in a standard normal distribution based on quantiles. For example, the median of our data will be paired with 0, the median of the standard normal distribution. The 80th percentile in our data will be paired with the 80th percentile of a standard normal distribution, namely:

\[
> \text{qnorm}(.80) \\
[1] 0.8416212
\]

and so on. If the points of the normal quantile plot fall roughly along a straight line, then the data are behaving as we would expect for a normal distribution. If they deviate (especially systematically) from a line, that is an indication of problems. We can make a normal quantile plot using \texttt{qqmath()} or \texttt{xqqmath()}.

\[
> \text{xqqmath( ~ biomass.ratio, MarineReserve) }
\]

This plot indicates that the largest biomass ratios are too large. That is, our data appear to be skewed to the right.

13.2 Prescription

When this is not the case we have three choices:

1. Ignore the problem.
2. Transform the data.
3. Use a different analysis method.
13.2.1 Solution 1: Ignore the Problem

Some violations of assumptions don’t matter much. (When this happens, we will say that a statistical method is robust.) The $t$-procedures generally more robust when the sample size is larger.

13.2.2 Solution 2: Transform the Data

Sometimes a simple transformation of the data will make things more normal.

While not perfect, this is certainly an improvement over the untransformed data.

An analysis of the log transformed data is pretty simple, but we must keep the transformation in mind as we do it.

Suppose our original null hypothesis is that the mean biomass ratio is 1.

- $H_0 : \mu = 1$

Then our new null hypothesis is that the mean of the log transformed biomass ratios is 0 (because $\log(1) = 0$).

- $H_0 : \mu' = 0$

```
> t.test ( log(MarineReserve$biomass.ratio) )

One Sample t-test

data: log(MarineReserve$biomass.ratio)
t = 7.3968, df = 31, p-value = 2.494e-08
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.3470180 0.6112365
sample estimates:
mean of x 0.4791272
```
The confidence interval given is for the log of biomass ratio. If we want a confidence interval for the biomass ratio, we need to “back-transform”:

```r
> exp( interval (t.test ( log(MarineReserve$biomass.ratio) ) ) )
Method: One Sample t-test

mean of x
0.4791272

95% confidence interval:
0.347018 0.6112365

[1] 1.414842 1.842708
attr(, "conf.level")
[1] 0.95
```

So the 95% confidence interval for the biomass ratio is $(1.415, 1.843)$.

See pages 330–331 and page 345 for other commonly used transformations.

### 13.2.3 Solution 3: Use another Analysis Method

**Sign Test as an Alternative to 1-sample t or Paired t**

A sign test works by converting a quantitative variable to a categorical variable with two levels (over and under, or + and −, whence the name sign test). The sign test uses a 1-proportion test (`binom.test()` or `prop.test()`) to see if 50% of the data fall on either side of a hypothesized median. It is very often used in place of paired $t$-tests, in which case the categorical variable indicates which of the two values is larger.

These tests are easy to do in R because R can work with logical values (`TRUE` and `FALSE`):

```r
> SexualSelection$polyandrous.species > SexualSelection$monandrous.species
[1] TRUE FALSE TRUE TRUE TRUE FALSE FALSE FALSE TRUE FALSE FALSE TRUE TRUE TRUE TRUE
> table(SexualSelection$polyandrous.species > SexualSelection$monandrous.species)
FALSE TRUE
  7   18
> binom.test(SexualSelection$polyandrous.species > SexualSelection$monandrous.species)

  Exact binomial test

data: SexualSelection$polyandrous.species > SexualSelection$monandrous.species
number of successes = 7, number of trials = 25, p-value = 0.04329
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.1207167 0.4938768
sample estimates:
probability of success
 0.28
```

Or alternatively,
> SexualSelection$difference > 0
[1] TRUE FALSE TRUE TRUE TRUE TRUE FALSE FALSE FALSE TRUE FALSE FALSE TRUE
[15] TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
> table( SexualSelection$difference > 0 )
character(0)
> binom.test( SexualSelection$difference > 0 )

   Exact binomial test

data: SexualSelection$difference > 0
number of successes = 7, number of trials = 25, p-value = 0.04329
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.1207167 0.4938768
sample estimates:
  probability of success
  0.28

An Alternative for the 2-sample t

The replacement test goes by two names: The Mann-Whitney U-Test and the Wilcoxon Rank Sum Test. We will use Example 13.5 to illustrate. (See the text for a description of the study.)

First we take a look at the data graphically:

> histogram( ~time.to.mating | treatment, SagebrushCrickets )

> xqmath( ~time.to.mating | treatment, SagebrushCrickets )
Neither population appears to be normally distributed, and the sample sizes are pretty small (13 in one group and 11 in the other). So we will try a method that doesn’t assume the populations are normally distributed.

1. State Hypotheses

   - $H_0$: the two populations have the same distribution.
   - $H_0$: the two populations have different distributions.

   Stated this way, if we reject the null hypothesis, we can’t conclude anything particular about how the distributions differ (mean? standard deviation? shape?)

   In typical applications, we assume that the two distributions have the same shape (same standard deviation and skew, etc.) but are perhaps shifted versions of each other. The test has better power under this assumption than in the general case.

   If we make this extra assumption, then the hypotheses become the same as for the 2-sample $t$-test.

   - $H_0: \mu_{\text{fed}} = \mu_{\text{starved}}$ The two groups (starved or fed) have the same mean time until mating.
   - $H_0: \mu_{\text{fed}} \neq \mu_{\text{starved}}$ The two groups (starved or fed) have the different mean time until mating.

2. Compute Test Statistic

   The test statistic (usually called $U$ or $W$) is a little trickier to compute that others we have seen. Fortunately, R will calculate it for us, as we’ll see shortly.

   The idea is to imagine a competition between the two groups (teams). In this competition, each player (a measurement) plays one match against each player on the opposing team. The winner of a match is the player with the larger measurement. And the team score is the total number of wins.

   Let’s illustrate with a smaller version of the SagebrushCrickets data set.
```r
> set.seed(12345); Crickets <- sample(SagebrushCrickets,7)
> Crickets
   treatment time.to.mating orig.row
1     fed        22.8         18
2     fed        72.1         21
3     fed        22.6         17
4     fed        39.0         19
5   starved       29.0         10
6   starved        9.0          4
7   starved       13.0          6
```

Now let’s divide up the teams:

```r
> team1 <- subset(Crickets, treatment=='fed')$time.to.mating
> team2 <- subset(Crickets, treatment=='starved')$time.to.mating
> team1
[1] 22.8 72.1 22.6 39.0
> team2
[1] 29 9 13
```

and then calculate the number of wins for each player on team 1.

<table>
<thead>
<tr>
<th>team 1 player</th>
<th>wins</th>
<th>team 2 players defeated</th>
</tr>
</thead>
<tbody>
<tr>
<td>22.8</td>
<td>2</td>
<td>9 13</td>
</tr>
<tr>
<td>72.1</td>
<td>3</td>
<td>29 9 13</td>
</tr>
<tr>
<td>22.6</td>
<td>2</td>
<td>9 13</td>
</tr>
<tr>
<td>39</td>
<td>3</td>
<td>29 9 13</td>
</tr>
</tbody>
</table>

Our test statistic is

\[ W = 10 = \text{the number of wins for team 1} \]

3. Determine p-value

The p-value is computed by imagining many competitions like the one above but where we randomly shuffle the measurements first.
13.8 Handling Violations of Assumptions

We see that the probability of getting a 10 or more wins or 2 or fewer wins is about 0.229.

The good news is that R can do the entire test in a single command.

```r
> wilcox.test(time.to.mating ~ treatment, Crickets)

    Wilcoxon rank sum test

data:  time.to.mating by treatment
W = 10, p-value = 0.2286
alternative hypothesis: true location shift is not equal to 0
```

Notice that the test statistic ($W$ in R) and p-value match our calculations.

4. Draw a Conclusion.

Once we have the p-value, the interpretation is the same as always. But let’s look at the full data set now.

Our test statistic will be much larger because there are a lot more games now:

```r
> summary(time.to.mating ~ treatment, SagebrushCrickets)

time.to.mating  N=24

+-------------------------+-------+-------------------------+
|                          | N | time.to.mating          |
|-------------------------+---+-------------------------+
| treatment              |   |                         |
| fed                    | 13| 35.98462                |
| starved                | 11| 17.72727                |
| Overall                | 24| 27.61667                |
+-------------------------+-------+-------------------------+

So there are $13 \cdot 11 = 143$ games.
> wilcox.test(time.to.mating ~ treatment, SagebrushCrickets)

   Wilcoxon rank sum test

data:  time.to.mating by treatment
W = 88, p-value = 0.3607
alternative hypothesis: true location shift is not equal to 0

The first “team” won 88 of these (and team 2 won the other 55), but the p-value indicates that if the null hypothesis is true and the two populations have the same distributions, then the difference in wins for the two teams would be at least this large more than 1/3 of the time just by random chance. So there is no reason to reject our null hypothesis in this case.

Why is it called the Rank Sum Test?

It may have occurred to you that we don’t need to know the values, we only need to know their ranks, to determine which player wins a match.

Let’s go back to our small subset and compute the ranks:

> Crickets$rank <- rank(Crickets$time.to.mating)
> Crickets

<table>
<thead>
<tr>
<th>treatment</th>
<th>time.to.mating</th>
<th>orig.row</th>
<th>rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>fed 18</td>
<td>22.8</td>
<td>18</td>
<td>4</td>
</tr>
<tr>
<td>fed 21</td>
<td>72.1</td>
<td>21</td>
<td>7</td>
</tr>
<tr>
<td>fed 17</td>
<td>22.6</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>fed 19</td>
<td>39.0</td>
<td>19</td>
<td>6</td>
</tr>
<tr>
<td>starved 10</td>
<td>29.0</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>starved 4</td>
<td>9.0</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>starved 6</td>
<td>13.0</td>
<td>6</td>
<td>2</td>
</tr>
</tbody>
</table>

> team1 <- sort(subset(Crickets, treatment=='fed')$rank); team1
[1] 3 4 6 7
> team2 <- sort(subset(Crickets, treatment=='starved')$rank); team2
[1] 1 2 5

Based on ranks, our table becomes

<table>
<thead>
<tr>
<th>team 1 player</th>
<th>wins</th>
<th>team 2 players defeated</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
<td>1 2</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1 2</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>1 2 5</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>1 2 5</td>
</tr>
</tbody>
</table>

| 20 | 10 |

The sum of ranks is not quite the same as the total number of wins, but there is a relationship. A player’s rank is the same as the number of players it would defeat (including the players on its own team and itself). If we subtract away those extra “wins”, we get the value we are after:
13.10 Handling Violations of Assumptions

<table>
<thead>
<tr>
<th>team 1 player</th>
<th>extra</th>
<th>wins</th>
<th>team 2 players defeated</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1 2</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
<td>1 2</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>3</td>
<td>1 2 5</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>3</td>
<td>1 2 5</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

The general formula is

\[ \text{sum of ranks for team 1} - \frac{n_1(n_1 + 1)}{2} \]

which in our case gives

\[ 20 - \frac{4 \cdot 5}{2} = 20 - 10 = 10 . \]

We can try this out on the larger data set too:

```r
> SagebrushCrickets$rank <- rank(SagebrushCrickets$time.to.mating)
> summary( rank ~ treatment, SagebrushCrickets, fun=function(x) c(ranksum=sum(x)) ) # fun = sum works too

rank  N=24
+---------+-------+---+-------+
|         |       |N  |ranksum|
+---------+-------+---+-------+
|treatment|fed    |13 |179   |
|         |starved|11 |121   |
+---------+-------+---+-------+
|Overall   |       |24 |300   |
+---------+-------+---+-------+
```

So \( W \) is given by

\[ W = 179 - \frac{13 \cdot 14}{2} = 179 - 91 = 88 . \]

Or, from the perspective of the starved crickets:

\[ W = 121 - \frac{11 \cdot 12}{2} = 121 - 66 = 55 . \]

In the Mann-Whitney variation, we always take the larger of these two values of \( W \) (and call it \( U \)). This is what the book describes, but not the way \( R \) does things (since it really isn’t necessary).
This chapter is quite nicely laid out. Here I’ll just highlight a few extras from class:

1. We used the Polio Vaccine Trials as a case study for the design of an experiment. I’ve posted a link to a web article with details of the study for your reference.

2. I demonstrated how to use power.t.test() and power.prop.test() to determine sample sizes for studies comparing two proportions or two means.

## 14.1 Power Calculations

Each of these functions allows you to specify all but one of a number of design elements and solves for the missing one. Here’s one example: Suppose we want to estimate the difference in two population means (using 2-sample $t$). The precision of this estimate depends on several ingredients:

1. Our desired level of confidence for the interval.

2. The sizes of our two samples. ($n_1$ and $n_2$)

3. The standard deviations of the two populations. ($\sigma_1$ and $\sigma_2$, which will in turn affect our sample standard deviations $s_1$ and $s_2$)

4. The difference between the two means. ($\delta = \mu_1 - \mu_2$)

power.t.test() makes the simplifying assumptions that the two sample sizes are equal ($n_1 = n_2 = n$) and the two population standard deviations are equal ($\sigma_1 = \sigma_2 = \sigma$).

Power calculations always require that we make some assumptions about things we don’t know in advance. We can determine the power of a test with samples of size $n = 20$ assuming that $\sigma = .4$ and $\delta = .1$ as follows:
> power.t.test(n=20, delta=.1, sd=.4, sig.level=0.05)

    Two-sample t test power calculation

    n = 20
delta = 0.1
    sd = 0.4
    sig.level = 0.05
    power = 0.1171781
    alternative = two.sided

NOTE: n is number in *each* group

This shows that we would have just over a 10% chance of detecting a difference in populations means of $\delta = 0.1$ using a confidence interval at the 95% level or a hypothesis test with $\alpha = 0.05$. If we wanted to have higher power, we can let R estimate the sample size for our desired power:

> power.t.test(delta=.1, sd=.4, sig.level=0.05, power=.80)

    Two-sample t test power calculation

    n = 252.1281
delta = 0.1
    sd = 0.4
    sig.level = 0.05
    power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group

For the case of the polio vaccine trials, if we hoped to detect a difference in polio rate of 50 in 100,000 for unvaccinated children and 25 in 100,000 for vaccinated children, we can determine the sample size required to achieve 90% power using `power.prop.test()`.

> power.prop.test( p1=50/100000, p2=25/100000, power=.90 )

    Two-sample comparison of proportions power calculation

    n = 126039.7
    p1 = 5e-04
    p2 = 0.00025
    sig.level = 0.05
    power = 0.9
    alternative = two.sided

NOTE: n is number in *each* group

This explains why such a large sample was required. The researchers wanted to be very confident of reaching a clear conclusion, but the rate of polio was relatively low even without vaccination. Allowing for the possibility of a lower than average rate of polio in the year of the study (polio rates varied substantially from year to year), would require even larger samples.
Given these calculations, a study with 200,000 vaccinated and 200,000 unvaccinated children doesn’t seem at all unreasonable.
15.1 The Basic ANOVA situation

- Two variables: categorical explanatory and quantitative response
  - Can be used in either experimental or observational designs.
- Main Question: Does the population mean response depend on the (treatment) group?
  - $H_0$: the population group means are all the equal ($\mu_1 = \mu_2 = \cdots \mu_k$)
  - $H_a$: the population group means are not all equal
- If categorical variable has only 2 values, we already have a method: 2-sample $t$-test
  - ANOVA allows for 3 or more groups (sub-populations)
- $F$ statistic compares within group variation (how different are individuals in the same group?) to between group variation (how different are the different group means?)
- ANOVA assumes that each group is normally distributed with the same (population) standard deviation.
  - Check normality with histograms, normal quantile plots, etc.
  - Check equal standard deviation using 2:1 ratio rule (largest standard deviation at most twice the smallest standard deviation).

15.1.1 An Example: Jet Lag

This example is described on page 394.
Question: Are these differences significant? Or would we expect sample differences this large by random chance even if (in the population) the mean amount of shift is equal for all three groups?

Whether differences between the groups are significant depends on

1. the difference in the means
2. the amount of variation within each group
3. the sample sizes

The p-value listed in this output is the p-value for our null hypothesis that the mean population response is the same in each treatment group.
15.1.2 The ANOVA test statistic

The $F$ statistic is a bit complicated to compute. We'll generally let the computer handle that for us. But is useful to see one small example.

A Small Dataset

Suppose we have three groups

- Group 1: 5.3, 6.0, 6.7 $\bar{y}_1 = 6.00$
- Group 2: 5.5, 6.2, 6.4, 5.7 $\bar{y}_2 = 5.95$
- Group 3: 7.5, 7.2, 7.9 $\bar{y}_3 = 7.53$

Computing $F$

The spreadsheet below shows how $F$ is calculated. (There is a bit of round-off error in a few of these numbers; compare with the ANOVA table below.)

<table>
<thead>
<tr>
<th>data</th>
<th>group</th>
<th>mean</th>
<th>data - group mean</th>
<th>squared</th>
<th>group mean - overall mean</th>
<th>squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.3</td>
<td>1</td>
<td>6.0</td>
<td>-0.70</td>
<td>0.490</td>
<td>-0.4</td>
<td>0.194</td>
</tr>
<tr>
<td>6.0</td>
<td>1</td>
<td>6.0</td>
<td>0.00</td>
<td>0.000</td>
<td>-0.4</td>
<td>0.194</td>
</tr>
<tr>
<td>6.7</td>
<td>1</td>
<td>6.0</td>
<td>0.70</td>
<td>0.490</td>
<td>-0.4</td>
<td>0.194</td>
</tr>
<tr>
<td>5.5</td>
<td>2</td>
<td>5.95</td>
<td>-0.45</td>
<td>0.203</td>
<td>-0.5</td>
<td>0.240</td>
</tr>
<tr>
<td>6.2</td>
<td>2</td>
<td>5.95</td>
<td>0.25</td>
<td>0.063</td>
<td>-0.5</td>
<td>0.240</td>
</tr>
<tr>
<td>6.4</td>
<td>2</td>
<td>5.95</td>
<td>0.45</td>
<td>0.203</td>
<td>-0.5</td>
<td>0.240</td>
</tr>
<tr>
<td>5.7</td>
<td>2</td>
<td>5.95</td>
<td>-0.25</td>
<td>0.063</td>
<td>-0.5</td>
<td>0.240</td>
</tr>
<tr>
<td>7.5</td>
<td>3</td>
<td>7.53</td>
<td>-0.03</td>
<td>0.001</td>
<td>1.1</td>
<td>1.188</td>
</tr>
<tr>
<td>7.2</td>
<td>3</td>
<td>7.53</td>
<td>-0.33</td>
<td>0.109</td>
<td>1.1</td>
<td>1.188</td>
</tr>
<tr>
<td>7.9</td>
<td>3</td>
<td>7.53</td>
<td>0.37</td>
<td>0.137</td>
<td>1.1</td>
<td>1.188</td>
</tr>
<tr>
<td>TOTAL</td>
<td></td>
<td></td>
<td></td>
<td>1.757</td>
<td>5.106</td>
<td></td>
</tr>
<tr>
<td>TOTAL/df</td>
<td></td>
<td></td>
<td></td>
<td>0.25095714</td>
<td>2.55275</td>
<td></td>
</tr>
</tbody>
</table>

overall mean: 6.44; $F = \frac{2.5528}{0.25025} = 10.21575$

```r
> anova( lm ( response ~ group, small ) )
Analysis of Variance Table

Response: response
    Df Sum Sq Mean Sq  F value   Pr(>F)
  group  2 5.1273 2.56367 10.216 0.008394 **
Residuals 7 1.7567 0.25095
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```
Here are the details. The $F$ statistic is a ratio of two quantities. The numerator measures the amount of variation between groups (G). The denominator measures the amount of variation within groups (E).

**Sum of Squares (SS):**

- \( SSG = \sum_{obs} (\bar{y}_i - \bar{y})^2 = \sum_{obs} (\text{group mean} - \text{overall mean})^2 = 5.106 \)
- \( SSE = \sum_{obs} (y_{ij} - \bar{y}_i)^2 = \sum_{obs} (\text{value} - \text{group mean})^2 = 1.757 \)
- \( SST = SSG + SSE = \sum_{obs} (y_{ij} - \bar{y})^2 = \sum_{obs} (\text{value} - \text{overall mean})^2 = 6.863 \)

**Degrees of Freedom:**

- \( DFG = 2 = 1 \text{ less than number of groups} \)
- \( DFE = 7 = \text{total sample size} - \text{number of groups} \)
- \( DFT = 7 + 2 = 9 = \text{total sample size} - 1 \)

**Mean Squares (MS):** \( MS = SS/DF \)

- \( MSG = SSG/2 = 5.106/2 = 2.55275 \)
- \( MSE = SSE/7 = 1.757/7 = 0.2509 \)
- \( MST = SST/9 = 6.863/9 = 0.763 \)

\( F = MSG/MSE = 2.5528/0.25025 = 10.216 \)

- If the group means are all the same, then \( F \) is usually a little above or below 1.
- A large value of \( F \) is evidence against \( H_0 \), since it indicates that there is more difference between groups than within groups.
- The \( F \) distribution has two degrees of freedom parameters
  - numerator: one less than number of groups
  - denominator: observations minus number of groups

**Proportion of Variation Explained**

The ratio

\[ R^2 = \frac{SSG}{SST} \]

measures the proportion of the total variation that is explained by the grouping variable (treatment).
15.2 Confidence Intervals for One Mean At a Time

We can construct a confidence interval for any of the means by just taking a subset of the data and using `t.test()`, but there are some problems with this approach. Most importantly,

We were primarily interested in comparing the means across the groups. Often people will display confidence intervals for each group and look for “overlapping” intervals. But this is not the best way to look for differences.

Nevertheless, you will sometimes see graphs showing multiple confidence intervals and labeling them to indicate which means appear to be different from which. (See the solution to problem 15.3 for an example.)

15.3 Pairwise Comparison

We really want to compare groups in pairs, and we have a method for this: 2-sample $t$. But we need to make a couple adjustments to the two-sample $t$.

1. We will use a new formula for standard error that makes use of all the data (even from groups not involved in the pair).
2. We also need to adjust the critical value to take into account the fact that we are (usually) making multiple comparisons.

15.3.1 The Standard Error

$$SE = \sqrt{MSE \left( \frac{1}{n_i} + \frac{1}{n_j} \right)} = \sqrt{MSE} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

where $n_i$ and $n_j$ are the sample sizes for the two groups being compared. Basically, $\sqrt{MSE}$ is taking the place of $s$ in our usual formula. The degrees of freedom for this estimate is

$$DFE = \text{total sample size} - \text{number of groups}.$$ 

Ignoring the multiple comparisons issue, we can now to confidence intervals or hypothesis tests just as before.

- confidence interval:
  $$\bar{y}_i - \bar{y}_j \pm t_*SE$$

- test statistic (for $H_0: \mu_1 - \mu_2 = 0$):
  $$t = \frac{\bar{y}_i - \bar{y}_j}{SE}.$$
15.3.2 The Multiple Comparisons Problem

Suppose we have 5 groups in our study and we want to make comparisons between each pair of groups. That’s \( 4 + 3 + 2 + 1 = 10 \) pairs. If we made 10 independent 95% confidence intervals, the probability that all of the cover the appropriate parameter is 0.599:

\[
> .95^{10} \\
[1] 0.5987369
\]

So we have family-wide error rate of nearly 40%.

We can correct for this by adjusting our critical value. Let’s take a simple example: just two 95% confidence intervals. The probability that both cover (assuming independence) is

\[
> .95^2 \\
[1] 0.9025
\]

Now suppose we want both intervals to cover 95% instead of 90.2% of the time. We could get this by forming two 97.5% confidence intervals.

\[
> \sqrt{.95}; \\
[1] 0.9746794 \\
> .975^2 \\
[1] 0.950625
\]

This means we need a larger value for \( t_\ast \) for each interval.

The ANOVA situation is a little bit more complicated because

- There are more than two comparisons.
- The different comparisons are not independent (because they all come from the same data set.)

We will briefly describe two ways to make an adjustment for multiple comparisons.

15.3.3 Bonferroni Corrections – An Easy Over-adjustment

Bonferroni’s idea is simple: Simple divide the desired family-wise error rate by the number of tests or intervals. This is an over-correction, but it is easy to do, and is used in many situations where a better method is not known or a quick estimate is desired.

Here is a table showing a few Bonferroni corrections for looking at all pairwise comparisons.

<table>
<thead>
<tr>
<th>number of groups</th>
<th>number of pairs of groups</th>
<th>family-wise error rate</th>
<th>individual error rate</th>
<th>confidence level for determining ( t_\ast )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
<td>.05</td>
<td>0.017</td>
<td>0.983</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>.05</td>
<td>0.008</td>
<td>0.992</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>.05</td>
<td>0.005</td>
<td>0.995</td>
</tr>
</tbody>
</table>
15.3.4 Tukey’s Honest Significant Differences

Tukey’s Honest Significant Differences is a more exact method specifically for the situation of making all pairwise comparisons in an ANOVA situation. R can compute Tukey’s Honest Significant Differences (see Section 15.4 of the text for more information).

```
> TukeyHSD ( aov ( shift ~ treatment, JetLagKnees ) )
Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = shift ~ treatment, data = JetLagKnees)

$treatment
  diff      lwr      upr   p adj
eyes-control -1.2427  -2.1682  -0.3171  0.0079
knee-control -0.0269  -0.9525  0.8986  0.9969
knee-eyes  1.2157   0.2598   2.1716  0.0117

> plot( TukeyHSD ( aov ( shift ~ treatment, JetLagKnees ) ) )

95% family–wise confidence level

Differences in mean levels of treatment
```

Tukey’s method adjusts the confidence intervals, making them a bit wider, to give them the desired family-wide error rate. Tukey’s method also adjusts p-values (making them larger), so that when the means are all the same, there is only a 5% chance that a sample will produce any p-values below 0.05.

In this example we see that the eye group differs significantly from control group and also from the knee group, but that the knee and control groups are not significantly different. (We can tell this by seeing which confidence intervals contain 0 or by checking which adjusted p-values are less than 0.05.)
15.3.5 Other Adjustments

There are similar methods for testing other sets of multiple comparisons. Testing “one against all the others” goes by the name of Dunnet’s method, for example.

15.4 Example 15.4

Below is R code for some other examples that appear in the text. See the text for a description of the studies and for the questions being asked.

The data for the carbon transfer example on page 407 are not available, so I made fake data that exhibit the same behavior (see Table 15.1).

<table>
<thead>
<tr>
<th>shade</th>
<th>carbon.transfer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 deep</td>
<td>14.39</td>
</tr>
<tr>
<td>2 deep</td>
<td>14.64</td>
</tr>
<tr>
<td>3 deep</td>
<td>20.75</td>
</tr>
<tr>
<td>4 deep</td>
<td>12.39</td>
</tr>
<tr>
<td>5 deep</td>
<td>29.49</td>
</tr>
<tr>
<td>6 part</td>
<td>2.22</td>
</tr>
<tr>
<td>7 part</td>
<td>10.71</td>
</tr>
<tr>
<td>8 part</td>
<td>4.38</td>
</tr>
<tr>
<td>9 part</td>
<td>10.65</td>
</tr>
<tr>
<td>10 part</td>
<td>13.49</td>
</tr>
<tr>
<td>11 none</td>
<td>1.78</td>
</tr>
<tr>
<td>12 none</td>
<td>3.66</td>
</tr>
<tr>
<td>13 none</td>
<td>3.91</td>
</tr>
<tr>
<td>14 none</td>
<td>8.79</td>
</tr>
<tr>
<td>15 none</td>
<td>7.91</td>
</tr>
</tbody>
</table>

Table 15.1: Data from a carbon transfer experiment

We can use `xyplot()` to inspect the data:

```r
> xyplot(carbon.transfer ~ shade, carbon)
```
And then compute the $F$-statistic and p-value for our test that the mean amount of carbon exchange is the same for each shade group.

```r
> anova(lm(carbon.transfer ~ shade, carbon))
Analysis of Variance Table

Response: carbon.transfer
  Df Sum Sq Mean Sq  F value    Pr(>F)  
shade  2 470.70 235.352  8.7842 0.004468 **
Residuals 12 321.51  26.793
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Since the p-value is small, we follow up with Tukey’s pairwise comparisons.

```r
> TukeyHSD(aov(carbon.transfer ~ shade, carbon))

Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = carbon.transfer ~ shade, data = carbon)

$shade

  diff lwr  upr     p adj
none-deep  -13.12   -21.853767   -4.386233 0.0045715
part-deep  -10.04   -18.773767   -1.306233 0.0245930
part-none  3.08    -5.653767    11.813767 0.6260735
```

We have significant differences between the deep shade and the other two groups, but we cannot conclude that the partial shade and no shade groups differ based on this sample.

### 15.5 Example 15.6: Walking Sticks

This example illustrates a couple additional points about ANOVA using data on the femur length of walking sticks.
Comparing More Than Two Means Using ANOVA

> head(WalkingStickFemurs)

<table>
<thead>
<tr>
<th>specimen</th>
<th>femur length</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.26</td>
</tr>
<tr>
<td>2</td>
<td>0.26</td>
</tr>
<tr>
<td>3</td>
<td>0.23</td>
</tr>
<tr>
<td>4</td>
<td>0.19</td>
</tr>
<tr>
<td>5</td>
<td>0.25</td>
</tr>
<tr>
<td>6</td>
<td>0.23</td>
</tr>
</tbody>
</table>

Numerically coded categorical variables

The following seems like the right thing to do, but it doesn’t work right.

> anova( lm( femur.length ~ specimen, data=WalkingStickFemurs ) )

Analysis of Variance Table

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>specimen</td>
<td>1</td>
<td>0.0102</td>
<td>0.0102</td>
<td>8.4668</td>
<td>0.0054</td>
</tr>
<tr>
<td>Residuals</td>
<td>48</td>
<td>0.0578</td>
<td>0.0012</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notice that the degrees of freedom are wrong. There should be 24 degrees for freedom for the 25 groups (the specimens), and 25 (50 – 25) degrees of freedom for the residuals. The problem is that our categorical variable (specimen) is coded with numbers (1–25), so R is treating it like a quantitative variable.

The solution is to tell R that the variable is really categorical:

> anova( lm( femur.length ~ factor(specimen), data=WalkingStickFemurs ) )

Analysis of Variance Table

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>factor(specimen)</td>
<td>24</td>
<td>0.0591</td>
<td>0.0025</td>
<td>6.9209</td>
<td>4.077e-06</td>
</tr>
<tr>
<td>Residuals</td>
<td>25</td>
<td>0.0089</td>
<td>0.0004</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Now the degrees of freedom what we expect.

Random Effects

The other wrinkle in this example is that our 25 walking stick specimens are not the point of the study. They are merely a random sample from a larger population of walking sticks. A design like this leads to a random effects model. This is in contrast to the fixed effects model of our previous example.

The interpretation of a random effects model is slightly different, but in this simple situation, the $F$-statistic and p-value are the same. (In more complicated situations, the random effects and fixed effects situations need to be analyzed differently.)
We only covered the first section of this chapter where the correlation coefficient is defined. Highlights of this section:

1. The correlation coefficient measures the **strength** and **direction** of a linear relationship.

2. The correlation coefficient is on a scale from $-1$ to $1$.
   - Values near 1 indicate a strong positive association. Values near $-1$ indicate a strong negative association.
   - Values near 0 indicate a weak or non-existent association.

3. The correlation coefficient does not check that there is a linear relationship.
   - If the relationship follows some other pattern, then the correlation coefficient may be completely confused about it. Use a scatter plot to see if there is indeed a linear relationship.

4. The correlation coefficient can be computed in R using `cor(x,y)` where $x$ and $y$ are the variables.

5. The formula for the correlation coefficient is based on adding products of z-scores (and adjusting for sample size):

$$ r = \frac{1}{n - 1} \sum_{i} \frac{x_i - \overline{x}}{s_x} \cdot \frac{y_i - \overline{y}}{s_y} $$
17.1 The Simple Linear Regression Model

\[ Y = \alpha + \beta x + \epsilon \quad \text{where} \quad \epsilon \sim \text{Norm}(0, \sigma). \]

In other words:

- The mean response for a given predictor value \( x \) is given by a linear formula
  \[ \text{mean response} = \alpha + \beta x \]
- The distribution of all responses for a given predictor value \( x \) is normal.
- The standard deviation of the responses is the same for each predictor value.

17.2 The Least Squares Method

We want to determine the best fitting line to the data. The usual method is the method of least squares which chooses the line that has the \textit{smallest possible sum of squares of residuals}, where residuals are defined by

\[ \text{residual} = \text{observed} - \text{predicted} \]

For a line with equation \( y = a + bx \), this would be

\[ e_i = y_i - (a + bx) \]

Simple calculus (that you don’t need to know) allows us to compute the best \( a \) and \( b \) possible. These best values define the least squares regression line. We always compute these values using software, but it is good to note that the least squares line satisfies two very nice properties.

1. The point \((\bar{x}, \bar{y})\) is on the line.
   
   This means that \( \bar{y} = a + b\bar{x} \) (and \( a = \bar{y} - b\bar{x} \))

2. The slope of the line is \( b = r \frac{s_y}{s_x} \)
Since we have a point and the slope, it is easy to compute the equation for the line if we know $\bar{x}$, $s_x$, $\bar{y}$, $s_y$, and $r$.

Fortunately, statistical software packages do all this work for us. For example, consider the lion noses example from the book.

```r
> lm( age ~ proportion.black, data=LionNoses)
Call:
  lm(formula = age ~ proportion.black, data = LionNoses)
Coefficients:
(Intercept) proportion.black
     0.879       10.647
```

This means that the equation of the least squares regression line is

$$\hat{y} = 0.879 + 10.647x$$

We use $\hat{y}$ to indicate that this is not an observed value of the response variable but an estimated value (based on the linear equation given).

R can add a regression line to our scatter plot if we ask it to.

```r
> xyplot( age ~ proportion.black, data=LionNoses, type=c('p','r') )
```

We see that the line does run roughly “through the middle” of the data but that there is also a good deal of variability above and below the line.

**Explanatory and Response Variables Matter**

It is important that the explanatory variable be the “x” variable and the response variable be the “y” variable when doing regression.

### 17.3 Making Predictions

We can use our least squares regression line to make predictions. For example, if a lion has a nose that is 30% black, we estimate the age to be
\[ \hat{\text{age}} = 0.879 + 10.647 \cdot 0.30 = 4.0731 \]

As it turns out, the 17th lion in our data set has a nose that was 30% black.

```
> LionNoses[17,]
  age proportion.black
17 4.3 0.3
```

That lion had an age of 4.3 years, so the residual for this lion is

\[ \text{observed} - \text{predicted} = 4.3 - 4.07 = 0.23 \]

The observed age is 0.23 years larger that the regression line predicts.

### 17.4 How Good Are Our Predictions?

How well have we estimated the slope and intercept? How accurately can we predict a lion’s age from the blackness of its nose? We would like more than just point estimates for these values, we would like confidence intervals and p-values. Since the regression model is based on normal distributions, these will look very much like they do for a 1-sample \( t \). The only things that change are the standard error formulas and the degrees of freedom.

The degrees of freedom is \( n - 2 \) for simple linear regression because we estimate two parameters (\( \alpha \) and \( \beta \)). The standard error formulas are bit messier, but computer software does the work of computing them for us.

#### 17.4.1 Estimates for Slope and Intercept

```
> lion.model <- lm( age ~ proportion.black, data=LionNoses)
> summary(lion.model)

Call:
  lm(formula = age ~ proportion.black, data = LionNoses)

Residuals:
     Min      1Q  Median      3Q     Max
-2.5449 -1.1117 -0.5285  0.9635  4.3421

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              0.8790     0.5688   1.545 0.133
proportion.black         10.6471     1.5095   7.053 7.68e-08 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.669 on 30 degrees of freedom
Multiple R-squared: 0.6238, Adjusted R-squared: 0.6113
F-statistic: 49.75 on 1 and 30 DF,  p-value: 7.677e-08
```

Notice the \( t \) statistics, standard errors and p-values in the table. We can use these to compute our confidence intervals and p-values.
A 95% confidence interval for the slope \( \beta \) is given by

\[
b \pm t_\ast SE
\]

\[
10.647 \pm 2.042 \cdot 1.5095
\]

\[
10.647 \pm 3.083
\]

using

\[
> \text{qt(.975, df=30)} # there are 32 lions so df= 32-2 = 30
\]

\[
[1] 2.042272
\]

to get the value of \( t_\ast \).

We can test the hypothesis that \( \beta = 0 \) even more easily since that is the hypothesis test for which the \( t \) statistic and p-value are listed in the summary output.

\[
t = \frac{b - 0}{SE} = \frac{10.647}{1.5095} = 7.053
\]

And the p-value is

\[
> 2 \cdot (1 - \text{pt}( 7.053, df=30) )
\]

\[
[1] 7.685428e-08
\]

just as the summary output reports.

To test \( H_0 : \beta = 1 \) (or any number other than 0) we need to do the work manually, but since the degrees of freedom and standard error are provided, it is straightforward to do.

Tests and intervals for the intercept work the same way but are usually much less interesting.

## 17.4.2 What Does the Slope Tell Us?

The slope is usually more interesting because it tells useful things about the relationship between the explanatory and response variable.

1. If the slope is 0, then our model becomes

\[
Y = a + \epsilon \quad \epsilon \sim \text{Norm}(0, \sigma)
\]

That is, \( Y \sim \text{Norm}(a, \sigma) \) and knowing the explanatory variable does not help us predict the response.

The test of whether \( \beta = 0 \) is therefore called the model utility test: it tells us whether the model is useful.

2. If the slope is not 0, then the slope tells us how much the average response changes for each unit increase in the predictor.

The intercept on the other hand, tells us what happens when the predictor is 0, which may be very uninteresting, especially if the predictor can’t be 0. (Suppose your predictor is your pulse, for example.)

## 17.4.3 Predictions for the Mean and Individual Response

R can compute two kinds of confidence intervals for the response for a given value.

1. The predict() function can plug new values into the linear regression equation for us. Here is our point estimate again for a lion with a nose that is 30% black.

\[
> \text{predict(lion.model, new=data.frame(proportion.black=.30))}
\]

\[
1
4.073142
\]
2. A confidence interval for the mean response for a given explanatory value can be computed by adding `interval='confidence'`.

```r
> predict(lion.model, new=data.frame(proportion.black=.30), interval='confidence')

fit lwr  upr
1  4.073142 3.466804 4.67948
```

3. An interval for an individual response (called a prediction interval to avoid confusion with the confidence interval above) can be computed by adding `interval='prediction'` instead.

```r
> predict(lion.model, new=data.frame(proportion.black=.30), interval='prediction')

fit  lwr  upr
1  4.073142 0.6115538 7.53473
```

Prediction intervals are

(a) much wider than confidence intervals  
(b) much less robust against violations of the normality assumption  
(c) a little bit wider than \( \hat{y} \pm 2SE \)

where \( SE \) is the “residual standard error” reported in the summary output.

```r
> 4.07 - 1.669 * 2
[1] 0.732
> 4.07 + 1.669 * 2
[1] 7.408
```

The prediction interval is a little wider because it takes into account the uncertainty in our estimated slope and intercept as well as the variability of responses around the true regression line.

The figure below shows the confidence (dotted) and prediction (dashed) intervals as bands around the regression line.

```r
> require(fastR)
> xyplot( age ~ proportion.black, data=LionNoses, panel=panel.lmbands, cex=.4, alpha=.5)
```

As the graph illustrates, the intervals are narrow near the center of the data and wider near the edges of the data. It is not safe to extrapolate beyond the data (without additional information), since there is no data to let us know whether the pattern of the data extends.
17.5 Checking Assumptions

17.5.1 Don’t Fit a Line If a Line Doesn’t Fit

When doing regression you should always look at the data to see if a line is a good fit. If it is not, it may be that a suitable transformation of one or both of the variables may improve things.

Anscombe’s Data

Anscombe illustrated the importance of looking at the data by concocting an interesting data set.

Notice how similar the numerical summaries are for these for pairs of variables

```
> summary( lm( y1 ~ x1, anscombe) )
Call:
  lm(formula = y1 ~ x1, data = anscombe)
Residuals:
      Min       1Q   Median       3Q      Max
-1.92127 -0.45577 -0.04136  0.70941  1.83882
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.0001    1.1247   2.667  0.02573 *
x1          0.5001    0.1179   4.241  0.00217 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared: 0.6665, Adjusted R-squared: 0.6295
F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217

> summary( lm( y2 ~ x2, anscombe) )
Call:
  lm(formula = y2 ~ x2, data = anscombe)
Residuals:
      Min       1Q   Median       3Q      Max
-1.9009 -0.7609  0.1291  0.9491  1.2691
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.001      1.125    2.667  0.02576 *
x2          0.500      0.118    4.239  0.00218 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared: 0.6662, Adjusted R-squared: 0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.00217
```
> summary(lm(y3 ~ x3, anscombe))
Call:
  lm(formula = y3 ~ x3, data = anscombe)

Residuals:
     Min     1Q   Median     3Q    Max
-1.1586 -0.6146 -0.2303  0.1540  3.2411

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.0025     1.1245   2.670  0.02562 *
x3          0.4997     0.1179   4.239  0.00218 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared: 0.6663,    Adjusted R-squared: 0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002176

> summary(lm(y4 ~ x4, anscombe))
Call:
  lm(formula = y4 ~ x4, data = anscombe)

Residuals:
     Min     1Q   Median     3Q    Max
-1.751 -0.831  0.000  0.809  1.839

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.0017     1.1239   2.671  0.02559 *
x4          0.4999     0.1178   4.243  0.00216 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared: 0.6667,    Adjusted R-squared: 0.6297
F-statistic: 18 on 1 and 9 DF,  p-value: 0.002165

But the plots reveal that very different things are going on.
17.5.2 Outliers in Regression

Outliers can be very influential in regression, especially in small data sets, and especially if they occur for extreme values of the explanatory variable. Outliers cannot be removed just because we don’t like them, but they should be explored to see what is going on (data entry error? special case? etc.)

Some researchers will do “leave-one-out” analysis, or “leave some out” analysis where they refit the regression with each data point left out once. If the regression summary changes very little when we do this, this means that the regression line is summarizing information that is shared among all the points relatively equally. But if removing one or a small number of values makes a dramatic change, then we know that that point is exerting a lot of influence over the resulting analysis (a cause for caution). See page 483 for an example of leave-one-out analysis.

17.5.3 Checking the Residuals

Residuals should be checked to see that the distribution looks approximately normal and that that standard deviation remains consistent across the range of our data (and across time).

```r
> xqqmath(~ resid(lion.model) )
```

![](image1.png)

```
> xyplot( resid(lion.model) ~ proportion.black, data=LionNoses )
```

![](image2.png)

Similar plots (and some others as well) can also be made with

```r
> plot(lion.model)
```
In this case things look pretty good, but not as good as they could be. The normal-quantile plot shows a slight curve, and the residuals seem to be a bit larger on the right side. A logarithmic transformation offers some improvement.

### 17.5.4 Transformations

Transformations of one or both variables can change the shape of the relationship (from non-linear to linear, we hope) and also the distribution of the residuals. In biological applications, a logarithmic transformation is often useful.

Consider the data from Problem 17.2 comparing zoo mortality rates (in %) to the minimal home ranges (in the wild) of various animal species.

```r
> xyplot( mortality ~ homerange, ZooMortality, type=c('p','r') )
```

![Graph showing mortality vs. homerange](image.png)

Clearly a line is not a good fit, and there is one very influential observation. Even if we remove the influential observation, things don’t look great.

```r
> xyplot( mortality ~ homerange, ZooMortality, subset=homerange<500, type=c('p','r') )
```

![Graph showing mortality vs. homerange after removing influential observation](image.png)

A logarithmic transformation of `homerange` takes care of most of our troubles.

```r
> xyplot( mortality ~ log(homerange), ZooMortality, type=c('p','r') )
```

This example won’t work quite right if you try to repeat it because in problem 17.2 the transformation was already done for you. I untransformed it so we could go through the whole process.
> zoo.model <- lm( mortality ~ log(homerange), ZooMortality )
> plot(zoo.model, which=1)

> plot(zoo.model, which=2)

> plot(zoo.model, which=2)

> plot(zoo.model, which=2)