Math 143: Introduction to Biostatistics

R Pruim

Spring 2011
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>0 What Is Statistics?</strong></td>
<td>1</td>
</tr>
<tr>
<td>0.1 Some Definitions of Statistics</td>
<td>1</td>
</tr>
<tr>
<td>0.2 A First Example: The Lady Tasting Tea</td>
<td>2</td>
</tr>
<tr>
<td>0.3 Coins and Cups</td>
<td>4</td>
</tr>
<tr>
<td><strong>1 Statistics and Samples</strong></td>
<td>1</td>
</tr>
<tr>
<td>1.1 Samples and Populations</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Data</td>
<td>2</td>
</tr>
<tr>
<td>1.3 Types of Statistical Studies</td>
<td>3</td>
</tr>
<tr>
<td><strong>2 Graphical Summaries</strong></td>
<td>1</td>
</tr>
<tr>
<td>2.1 Graphical Summaries – Important Ideas</td>
<td>1</td>
</tr>
<tr>
<td><strong>3 Numerical Summaries</strong></td>
<td>1</td>
</tr>
<tr>
<td>3.1 Summarizing Distributions of Quantitative Variables</td>
<td>1</td>
</tr>
<tr>
<td>3.2 Measures of Center</td>
<td>1</td>
</tr>
<tr>
<td>3.3 Measures of Spread</td>
<td>2</td>
</tr>
<tr>
<td>3.4 Shapes of Numerical Distributions</td>
<td>3</td>
</tr>
<tr>
<td>3.5 Summary Statistics and Transforming Data</td>
<td>4</td>
</tr>
<tr>
<td>3.6 Summarizing Categorical Variables</td>
<td>4</td>
</tr>
<tr>
<td>3.7 Relationships Between Two Variables</td>
<td>4</td>
</tr>
<tr>
<td><strong>5 Probability</strong></td>
<td>1</td>
</tr>
<tr>
<td>5.1 Key Definitions and Ideas</td>
<td>1</td>
</tr>
<tr>
<td>5.2 Computing Probabilities</td>
<td>1</td>
</tr>
<tr>
<td>5.3 Probability Axioms and Rules</td>
<td>2</td>
</tr>
<tr>
<td><strong>6 Hypothesis Testing</strong></td>
<td>1</td>
</tr>
<tr>
<td>6.1 The Four Step Process</td>
<td>1</td>
</tr>
<tr>
<td>6.2 Some Additional Details</td>
<td>4</td>
</tr>
<tr>
<td><strong>7 Inference for a Proportion</strong></td>
<td>1</td>
</tr>
<tr>
<td>7.1 Binomial Distributions</td>
<td>1</td>
</tr>
<tr>
<td>7.2 The Binomial Test</td>
<td>3</td>
</tr>
<tr>
<td>7.3 More Examples</td>
<td>4</td>
</tr>
</tbody>
</table>
A Getting Started with R

A.1 Welcome to R and RStudio ....................................................... 1
A.2 Using R as a Calculator .......................................................... 2
A.3 R packages ........................................................................... 3
A.4 Data ...................................................................................... 4
A.5 Summarizing Data ................................................................. 7
A.6 Getting Help ........................................................................ 18
A.7 Additional Notes on R Syntax ................................................ 19
A.8 Installing R ........................................................................ 20
0.1 Some Definitions of Statistics

This is a course primarily about statistics, but what exactly is statistics? In other words, what is this course about? Here are some definitions of statistics from other people:

- a collection of procedures and principles for gaining information in order to make decisions when faced with uncertainty (J. Utts [Utt05]),
- a way of taming uncertainty, of turning raw data into arguments that can resolve profound questions (T. Amabile [MA89]),
- the science of drawing conclusions from data with the aid of the mathematics of probability (S. Garfunkel [MA86]),
- the explanation of variation in the context of what remains unexplained (D. Kaplan [Kap09]),
- the mathematics of the collection, organization, and interpretation of numerical data, especially the analysis of a population’s characteristics by inference from sampling (American Heritage Dictionary [AmH82]).

While not exactly the same, these definitions highlight four key elements of statistics.

Data – the raw material

Data are the raw material for doing statistics. We will learn more about different types of data, how to collect data, and how to summarize data as we go along.

Information – the goal

The goal of doing statistics is to gain some information or to make a decision. Statistics is useful because it helps us answer questions like the following:

- Which of two treatment plans leads to the best clinical outcomes?

\[^1\] As we will see, the words statistic and statistics get used in more than one way. More on that later.
• Are men or women more successful at quitting smoking? And does it matter which smoking cessation program they use?

• Is my cereal company complying with regulations about the amount of cereal in its cereal boxes?

In this sense, statistics is a science – a method for obtaining new knowledge.

**Uncertainty – the context**

The tricky thing about statistics is the uncertainty involved. If we measure one box of cereal, how do we know that all the others are similarly filled? If every box of cereal were identical and every measurement perfectly exact, then one measurement would suffice. But the boxes may differ from one another, and even if we measure the same box multiple times, we may get different answers to the question *How much cereal is in the box?*

So we need to answer questions like *How many boxes should we measure?* and *How many times should we measure each box?* Even so, there is no answer to these questions that will give us absolute certainty. So we need to answer questions like *How sure do we need to be?*

**Probability – the tool**

In order to answer a question like *How sure do we need to be?*, we need some way of measuring our level of certainty. This is where mathematics enters into statistics. Probability is the area of mathematics that deals with reasoning about uncertainty.

**0.2 A First Example: The Lady Tasting Tea**

There is a famous story about a lady who claimed that tea with milk tasted different depending on whether the milk was added to the tea or the tea added to the milk. The story is famous because of the setting in which she made this claim. She was attending a party in Cambridge, England, in the 1920s. Also in attendance were a number of university dons and their wives. The scientists in attendance scoffed at the woman and her claim. What, after all, could be the difference?

All the scientists but one, that is. Rather than simply dismiss the woman’s claim, he proposed that they decide how one should *test* the claim. The tenor of the conversation changed at this suggestion, and the scientists began to discuss how the claim should be tested. Within a few minutes cups of tea with milk had been prepared and presented to the woman for tasting.

Let’s take this simple example as a prototype for a statistical study. What steps are involved?

1. Determine the question of interest.

   Just what is it we want to know? It may take some effort to make a vague idea precise. The precise questions may not exactly correspond to our vague questions, and the very exercise of stating the question precisely may modify our question. Sometimes we cannot come up with any way to answer the question we really want to answer, so we have to live with some other question that is not exactly what we wanted but is something we can study and will (we hope) give us some information about our original question.

   In our example this question seems fairly easy to state: Can the lady tell the difference between the two tea preparations? But we need to refine this question. For example, are we asking if she always correctly identifies cups of tea or merely if she does better than we could do ourselves (by guessing)?

2. Determine the population.
Just who or what do we want to know about? Are we only interested in this one woman or women in
general or only women who claim to be able to distinguish tea preparations?

3. Select measurements.

We are going to need some data. We get our data by making some measurements. These might be physical
measurements with some device (like a ruler or a scale). But there are other sorts of measurements too,
like the answer to a question on a form. Sometimes it is tricky to figure out just what to measure. (How
do we measure happiness or intelligence, for example?) Just how we do our measuring will have important
consequences for the subsequent statistical analysis. The recorded values of these measurements are called
variables (because the values vary from one individual to another).

In our example, a measurement may consist of recording for a given cup of tea whether the woman’s claim
is correct or incorrect.

4. Determine the sample.

Usually we cannot measure every individual in our population; we have to select some to measure. But
how many and which ones? These are important questions that must be answered. Generally speaking,
bigger is better, but it is also more expensive. Moreover, no size is large enough if the sample is selected
inappropriately.

Suppose we gave the lady one cup of tea. If she correctly identifies the mixing procedure, will we be
convinced of her claim? She might just be guessing; so we should probably have her taste more than one
cup. Will we be convinced if she correctly identifies 5 cups? 10 cups? 50 cups?

What if she makes a mistake? If we present her with 10 cups and she correctly identifies 9 of the 10, what
will we conclude? A success rate of 90% is, it seems, much better than just guessing, and anyone can
make a mistake now and then. But what if she correctly identifies 8 out of 10? 80 out of 100?

And how should we prepare the cups? Should we make 5 each way? Does it matter if we tell the woman
that there are 5 prepared each way? Should we flip a coin to decide even if that means we might end up
with 3 prepared one way and 7 the other way? Do any of these differences matter?

5. Make and record the measurements.

Once we have the design figured out, we have to do the legwork of data collection. This can be a time-
consuming and tedious process. In the case of the lady tasting tea, the scientists decided to present her
with ten cups of tea which were quickly prepared. A study of public opinion may require many thousands
of phone calls or personal interviews. In a laboratory setting, each measurement might be the result of a
carefully performed laboratory experiment.

6. Organize the data.

Once the data have been collected, it is often necessary or useful to organize them. Data are typically
stored in spreadsheets or in other formats that are convenient for processing with statistical packages.
Very large data sets are often stored in databases.

Part of the organization of the data may involve producing graphical and numerical summaries of the
data. These summaries may give us initial insights into our questions or help us detect errors that may
have occurred to this point.

7. Draw conclusions from data.

Once the data have been collected, organized, and analyzed, we need to reach a conclusion. Do we believe
the woman’s claim? Or do we think she is merely guessing? How sure are we that this conclusion is
correct?

Eventually we will learn a number of important and frequently used methods for drawing inferences from
data. More importantly, we will learn the basic framework used for such procedures so that it should
become easier and easier to learn new procedures as we become familiar with the framework.

8. Produce a report.

Typically the results of a statistical study are reported in some manner. This may be as a refereed article
in an academic journal, as an internal report to a company, or as a solution to a problem on a homework
assignment. These reports may themselves be further distilled into press releases, newspaper articles, advertisements, and the like. The mark of a good report is that it provides the essential information about each of the steps of the study.

As we go along, we will learn some of the standard terminology and procedures that you are likely to see in basic statistical reports and will gain a framework for learning more.

At this point, you may be wondering who the innovative scientist was and what the results of the experiment were. The scientist was R. A. Fisher, who first described this situation as a pedagogical example in his 1925 book on statistical methodology \cite{Fis25}. Fisher developed statistical methods that are among the most important and widely used methods to this day, and most of his applications were biological.

### 0.3 Coins and Cups

You might also be curious about how the experiment came out. How many cups of tea were prepared? How many did the woman correctly identify? What was the conclusion?

Fisher never says. In his book he is interested in the method, not the particular results. But let’s suppose we decide to test the lady with ten cups of tea. We’ll flip a coin to decide which way to prepare the cups. If we flip a head, we will pour the milk in first; if tails, we put the tea in first. Then we present the ten cups to the lady and have her state which ones she thinks were prepared each way.

It is easy to give her a score (9 out of 10, or 7 out of 10, or whatever it happens to be). It is trickier to figure out what to do with her score. Even if she is just guessing and has no idea, she could get lucky and get quite a few correct – maybe even all 10. But how likely is that?

Let’s try an experiment. I’ll flip 10 coins. You guess which are heads and which are tails, and we’ll see how you do.

:  

Comparing with your classmates, we will undoubtedly see that some of you did better and others worse.

Now let’s suppose the lady gets 9 out of 10 correct. That’s not perfect, but it is better than we would expect for someone who was just guessing. On the other hand, it is not impossible to get 9 out of 10 just by guessing. So here is Fisher’s great idea: Let’s figure out how hard it is to get 9 out of 10 by guessing. If it’s not so hard to do, then perhaps that’s just what happened, so we won’t be too impressed with the lady’s tea tasting ability. On the other hand, if it is really unusual to get 9 out of 10 correct by guessing, then we will have some evidence that she must be able to tell something.

But how do we figure out how unusual it is to get 9 out of 10 just by guessing? We’ll learn another method later, but for now, let’s just flip a bunch of coins and keep track. If the lady is just guessing, she might as well be flipping a coin.

So here’s the plan. We’ll flip 10 coins. We’ll call the heads correct guesses and the tails incorrect guesses. Then we’ll flip 10 more coins, and 10 more, and 10 more, and . . . . That would get pretty tedious. Fortunately, computers are good at tedious things, so we’ll let the computer do the flipping for us using a tool in the mosaic package, which you can install using

```
> install.packages("mosaic", repos="http://R-Forge.R-project.org")
```

The \texttt{rflip()} function can flip one coin.
What Is Statistics?

> require(mosaic)
> rflip()

Flipping 1 coins [ Prob(Heads) = 0.5 ] ...

T

Result: 0 heads.

or a number of coins

> rflip(10)

Flipping 10 coins [ Prob(Heads) = 0.5 ] ...

H H H T H H H T T

Result: 7 heads.

and show us the results.

Typing `rflip(10)` a bunch of times is almost as tedious as flipping all those coins. But it is not too hard to tell R to do() this a bunch of times.

> do(2) * rflip(10)

[[1]]

Flipping 10 coins [ Prob(Heads) = 0.5 ] ...

H H H T H H T T H

Result: 7 heads.

[[2]]

Flipping 10 coins [ Prob(Heads) = 0.5 ] ...

H T T T T H T H H

Result: 4 heads.

And if we switch to a “quiet” version of `nflip()`, we will get much terser output – just the counts of the number of heads (i.e., correct guesses).

> do(10000) * nflip(10) # "quiet version" of rflip()

[1] 5 4 7 4 7 7 5 5 5 7 6 7 3 4 5 6 5 6 6 5 6 5 4 6 5 4 5 3 7 7 5 7 3 6 4 3 6 4 4 6
[43] 5 7 5 6 3 8 5 6 4 3 7 5 2 6 3 7 4 5 5 6 6 5 7 5 4 4 7 6 6 6 5 4 4 2 4 8 3 6 7 3
[85] 6 7 6 5 5 4 6 4 7 7 6 4 7 4 5 4

Let’s get R to do() it for us 10,000 times and make a table of the results.

> results <- do(10000) * nflip(10)
> table(results)

results
0 1 2 3 4 5 6 7 8 9 10
5 102 467 1203 2048 2470 2035 1140 415 108 7
You might be surprised to see that the number of correct guesses is exactly 5 (half of the 10 tries) only 25% of the time. But most of the results are quite close to 5 correct. 67% of the results are 4, 5, or 6, for example. And 90% of the results are between 3 and 7 (inclusive). But getting 8 correct is a bit unusual, and getting 9 or 10 correct is even more unusual.

So what do we conclude? It is possible that the lady could get 9 or 10 correct just by guessing, but it is not very likely (it only happened in about 1.2% of our simulations). So one of two things must be true:

- The lady got unusually “lucky”, or
- The lady is not just guessing.

Although Fisher did not say how the experiment came out, others have reported that the lady correctly identified all 10 cups! [Sal01]
1.1 Samples and Populations

population the collection of animals, plants, objects, etc. that we want to know about

sample the (smaller) set of animals, plants, objects, etc. about which we have data

parameter a number that describes a population or model.

statistic a number that describes a sample.

Much of statistics centers around this question:

What can we learn about a population from a sample?

Estimation

Often we are interested in knowing (approximately) the value of some parameter. A statistic used for this purpose is called an estimate. For example, if you want to know the mean length of the tails of lemurs (that’s a parameter), you might take a sample of lemurs and measure their tails. The mean length of the tails of the lemurs in your sample is a statistic. It is also an estimate, because we use it to estimate the parameter.

Statistical estimation methods attempt to

- reduce bias, and
- increase precision.

bias the systematic tendency of sample estimates to either overestimate or underestimate population parameters; that is, a tendency to be off in a particular direction.

precision the measure of how close estimates are to the thing being estimated (called the estimand).
Sampling

**Sampling** is the process of selecting a sample. Statisticians use random samples

- to avoid (or at least reduce) **bias**, and
- so they can quantify **sampling variability** (the amount samples differ from each other), which in turn allows us to quantify precision.

The simplest kind of random sample is called a **simple random sample** (aren’t statisticians clever about naming things?). A simple random sample is equivalent to putting all individuals in the population into a big hat, mixing thoroughly, and selecting some out of the hat to be in the sample. In particular, in a simple random sample, *every individual has an equal chance to be in the sample*, in fact, every subset of the population of a fixed size has an equal chance to be in the sample.

Other sampling methods include

- **convenience sampling** using whatever individuals are easy to obtain
- **volunteer sampling** using people who volunteer to be in the sample
- **systematic sampling** sampling done in some systematic way (every tenth unit, for example).
- **stratified sampling** sampling separately in distinct sub-populations (called *strata*).

### 1.2 Data

Imagine data as a 2-dimensional structure (like a spreadsheet).

- Rows correspond to **observational units** (people, animals, plants, or other objects we are collecting data about).
- Columns correspond to **variables** (measurements collected on each observational unit).

**Variable terminology**

- **categorical variable** a variable that places observation units into one of two or more categories (examples: color, sex, case/control status, species, etc.)
- **quantitative variable** a variable that records measurements along some scale (examples: weight, height, age, temperature) or counts something (examples: number of siblings, number of colonies of bacteria, etc.)
- **response variable** a variable we are trying to predict or explain
- **explanatory variable** a variable used to predict or explain a response variable
Distributions

The distribution of a variable answers two questions:

- What values can the variable have?
- With what frequency does each value occur?

The frequency may be described in terms of counts, proportions (often called relative frequency), or densities (more on densities later).

A distribution may be described using a table (listing values and frequencies) or a graph (e.g., a histogram).

1.3 Types of Statistical Studies

Statisticians use the word experiment to mean something very specific. In an experiment, the researcher determines the values of one or more (explanatory) variables, typically by random assignment. If there is no such assignment by the researcher, the study is an observational study.
2.1 Graphical Summaries – Important Ideas

2.1.1 Patterns and Deviations from Patterns

The goal of a statistical plot is to help us see

- potential patterns in the data, and
- deviations from those patterns.

2.1.2 Different Plots for Different Kinds of Variables

Graphical summaries can help us see the distribution of a variable or the relationships between two (or more) variables. The type of plot used will depend on the kinds of variables involved. There is a nice summary of these on page 48. You can use `demo()` to see how to get R to make the plots in this section.

Later, when we do statistical analysis, we will see that the analysis we use will also depend on the kinds of variables involved, so this is an important idea.

2.1.3 Side-by-side Plots and Overlays Can Reveal Importance of Additional Factors

The `lattice` graphics plots make it particularly easy to generate plots that divide the data into groups and either produce a panel for each group (using `|`) or display each group in a different way (different colors or symbols, using the `groups` argument). These plots can reveal the possible influence of additional variables – sometimes called covariates.

2.1.4 Area = (relative) frequency

Many plots are based on the key idea that our eyes are good at comparing areas. Plots that use area (e.g., histograms, mosaic plots, bar charts, pie charts) should always obey this principle

\[ \text{Area} = \text{(relative) frequency} \]
Plots that violate this principle can be deceptive and distort the true nature of the data.

An Example: Histogram with unequal bin widths

It is possible to make histograms with bins that have different widths. But in this case it is important that the height of the bars is chosen so that area (NOT height) is proportional to frequency. Using height instead of area would distort the picture.

Correct:

```r
> histogram(~Sepal.Length, data=iris, breaks=c(4,5.5,5.75,6,6.5,7,8,9), type='density')
```

![Histogram with correct bin width and density](image)

The `type='density'` is important. It tells R to use a scale such that the area (height $\times$ width) of the rectangles is equal to the relative frequency. For example, the bar from 5.0 to 5.5 has width $\frac{1}{2}$ and height about 0.36, so the area is 0.18, which means approximately 18% of the sepal lengths are between 5.0 and 5.5.

Incorrect:

```r
> histogram(~Sepal.Length, data=iris, breaks=c(4,5.5,5.75,6,6.5,7,8,9))
```

![Histogram with incorrect bin width and frequency](image)

Notice how different this looks. Now the heights are equal to the relative frequency, but this makes the wider bars have too much area.
3.1 Summarizing Distributions of Quantitative Variables

Important Note

Numerical summaries are a convenient way to describe a distribution, but remember that numerical summaries do not tell you everything there is to know about a distribution any more than know a few facts (e.g., age, employment, sex) tell you everything there is to know about a person.

Notation

In statistics $n$ (or sometimes $N$) almost always means the number of observations (i.e., the number of rows in a data frame).

If $y$ is a variable in a data set with $n$ observational units, we can denote the $n$ values of $y$ as

- $y_1, y_2, y_3, \ldots, y_n$ (in the original order of the data).
- $y_{(1)}, y_{(2)}, y_{(3)}, \ldots, y_{(n)}$ (in sorted order from smallest to largest).

The text uses capital letters for this, but it is more common to use lower case letters for particular data sets. (Capitals get used for something called random variables.)

The symbol $\sum$ represents summation (adding up a bunch of values).

3.2 Measures of Center

Measures of center attempt to give us a sense of what is a typical value for the distribution.

The mean of $y = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\text{sum of values}}{\text{number of values}}$

The median of $y$ is the “middle” number (after putting the numbers in increasing order)
• The mean is the “balancing point” of the distribution.
• The median is the 50th percentile: half of the distribution is below the median, half is above.
• If the distribution is symmetric, then the mean and median are the same.
• In a skewed distribution, the mean is pulled toward the tail of the distribution (relative to the median).
• *A few very large or very small values can change the mean a lot*, so the mean is **sensitive to outliers** and is a better measure of center when the distribution is symmetric than when it is skewed.
• The median is a **resistant measure** – it is not affected much by a few very large or very small values.

### 3.3 Measures of Spread

\[
\text{variance of } y = s_y^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2
\]

\[
\text{standard deviation of } y = s_y = \sqrt{s_y^2} = \text{square root of variance}
\]

\[
\text{interquartile range } = \text{IQR} = Q_3 - Q_1 = \text{difference between first and third quartiles}
\]

• Roughly, the standard deviation is the “average deviation from the mean”. (That’s not exactly right because of the squaring involved.)
• The mean and standard deviation are especially useful for describing **normal distributions** and other unimodal, symmetric distributions that are roughly “bell-shaped”. (We’ll learn more about normal distributions later.)
• Like the mean, the variance and standard deviation are sensitive to outliers and less suited for summarizing skewed distributions.
• The book shows examples of computing the variance and standard deviation by hand. This is mostly important so we understand how these measures work, but we will typically let R do the calculations for us.

```R
> mean(faithful$eruptions)
[1] 3.487783
> sd(faithful$eruptions)
[1] 1.141371
> var(faithful$eruptions)
[1] 1.302728

> median(faithful$eruptions)
[1] 4
> IQR(faithful$eruptions)
[1] 2.2915
> favstats(faithful$eruptions)
    median    IQR    mean     sd     var
  4.000000 2.291500 3.487783 1.141371 1.302728
```
> summary( Sepal.Length ~ Species, iris, fun=favstats)  # favorite stats for each species

Sepal.Length  N=150

+-------+----------+---+------+-----+--------+---------+---------+
|      |  N | median |IQR | mean |sd | var  |
+-------+----------+---+------+-----+--------+---------+---------+
|Species|setosa | 50|5.0 |0.400|5.006000|0.3524897|0.1242490|
| |versicolor| 50|5.9 |0.700|5.936000|0.5161711|0.2664327|
| |virginica | 50|6.5 |0.675|6.588000|0.6358796|0.4043429|
+-------+----------+---+------+-----+--------+---------+---------+
|Overall|150|5.8 |1.300|5.843333|0.8280661|0.6856935|
+-------+----------+---+------+-----+--------+---------+---------+

3.4 Shapes of Numerical Distributions

If we make a histogram of our data, we can describe the overall shape of the distribution. Keep in mind that the shape of a particular histogram may depend on the choice of bins. Choosing too many or too few bins can hide the true shape of the distribution. (When in doubt, make more than one histogram.)

Here are some words we use to describe shapes of distributions.

**symmetric**  The left and right sides are mirror images of each other.

**skewed**  The distribution stretches out farther in one direction than in the other. (We say the distribution is skewed toward the long tail.)

**uniform**  The heights of all the bars are (roughly) the same. (So the data are equally likely to be anywhere within some range.)

**unimodal**  There is one major “bump” where there is a lot of data.

**bimodal**  There are two “bumps”.

**outlier**  An observation that does not fit the overall pattern of the rest of the data.

None of these measures (especially the mean and median) is a particularly good summary if the distribution is not unimodal.

> favstats(faithful$eruptions)

<table>
<thead>
<tr>
<th>median</th>
<th>IQR</th>
<th>mean</th>
<th>sd</th>
<th>var</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.000000</td>
<td>2.291500</td>
<td>3.487783</td>
<td>1.141371</td>
<td>1.302728</td>
</tr>
</tbody>
</table>

> histogram(~ eruptions, data=faithful, n=20, +
  main="Old Faithful Eruption Times", xlab="duration (min)"")

![Old Faithful Eruption Times](chart.png)
3.5 Summary Statistics and Transforming Data

<table>
<thead>
<tr>
<th>statistic</th>
<th>original data</th>
<th>original data + $k$</th>
<th>original data × $c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>$\bar{y}$</td>
<td>$\bar{y} + k$</td>
<td>$\bar{y} \cdot c$</td>
</tr>
<tr>
<td>standard deviation</td>
<td>$s$</td>
<td>$s$</td>
<td>$s \cdot c$</td>
</tr>
<tr>
<td>variance</td>
<td>$s^2$</td>
<td>$s^2$</td>
<td>$s^2 \cdot c^2$</td>
</tr>
<tr>
<td>median</td>
<td>$m$</td>
<td>$m + k$</td>
<td>$m \cdot c$</td>
</tr>
<tr>
<td>interquartile range</td>
<td>IQR</td>
<td>IQR + $k$</td>
<td>IQR × $c$</td>
</tr>
</tbody>
</table>

3.6 Summarizing Categorical Variables

The most common summary of a categorical variable is the proportion:

$$\hat{p} = \frac{\text{number in one category}}{n}$$

Proportions can be expressed as fractions, decimals or percents. For example, if there are 10 observations in one category and $n = 50$ observations in all, then

$$\hat{p} = \frac{10}{25} = \frac{2}{5} = 0.40 = 40\%$$

If we code our categorical variable using 1 for observations in “the one category” and 0 for observations in any other category, then a proportion is a sample mean.

$$\frac{1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0}{25} = \frac{10}{25}$$

3.7 Relationships Between Two Variables

It is also possible to give numerical summaries of the relationship between two variables. The most common one is the correlation coefficient, which we will learn about later.

Entering Cross Tables via Excel or Google

I’ve written a little function to help convert data frames that are holding a contingency table into an actual contingency table. This can be used to get a contingency table from Google or Excel into RStudio. (See page A.15 for a different way to do this.) To use this function, you must first load it from my shared space.

```r
> source('shared/rpruim@calvin.edu/Math143/utilities.R')
```
Now we can convert a data frame into a cross table.

```r
> df # a data frame that really contains a contingency table
   X AAA BBB
  1 XXX 15  23
  2 YYY 20  17

> xt <- as.xtabs(df, "row variable", "col variable"); xt # convert df to an xtabs object

   col variable
row variable AAA BBB
   XXX 15  23
   YYY 20  17
```

Now you can use `mosaic()` to make a mosaic plot of the contingency table.

```r
> mosaic(xt)
```

![Mosaic Plot](image-url)
5.1 Key Definitions and Ideas

random process A repeatable process that has multiple unpredictable potential outcomes.

Although we sometimes use language that suggests that a particular result is random, it is really the process that is random, not its results.

outcome A potential result of a random process.

sample space The set of all potential outcomes of a random process.

event A subset of the sample space. That is, a set of outcomes (possibly all or none of the outcomes).

trial One repetition of a random process.

mutually exclusive events. Events that cannot happen on the same trial.

probability A numerical value between 0 and 1 assigned to an event to indicate how often the event occurs (in the long run).

random variable A random process that results in a numerical value.

Examples:

- Roll a die and record the number.
- Roll two dice and record the sum.
- Flip 100 coins and count the number of heads.
- Sample 1000 people and ask them if they approve of the job the president is doing.

Note: Statisticians use capital letters for random variables.

probability distribution The distribution of a random variable. (Remember that a distribution describes what values? and with what frequency?)

5.2 Computing Probabilities

There are two main ways to calculate the probability of an event \( A \), denoted \( \Pr(A) \).
5.2 Probability

5.2.1 Empirical Method: Just Do It

Since random processes are repeatable, we could simply repeat the process over and over and keep track of how often the event \( A \) occurs. For example, we could flip a coin 10,000 times and see what fraction are heads.\(^1\)

\[
\text{Empirical Probability} = \frac{\text{number of times } A \text{ occurred}}{\text{number of times random process was repeated}}
\]

Now that we have computers, there is a new wrinkle on empirical probability. If we can simulate our random process in the computer, then we can repeat the process many times very quickly. We did this on the first day with our Lady Tasting Tea example.

5.2.2 Theoretical Method: Just Think About It

The theoretical method combines

1. Some basic facts that are true about every probability situation,
2. Some assumptions about the particular situation at hand, and
3. Mathematical reasoning (arithmetic, algebra, logic, etc.).

5.3 Probability Axioms and Rules

5.3.1 The Three Axioms

Let \( S \) be the sample space and let \( A \) and \( B \) be events.

1. Probability is between 0 and 1: \( 0 \leq \Pr(A) \leq 1 \).
2. The probability of the sample space is 1: \( \Pr(S) = 1 \).
3. Additivity: If \( A \) and \( B \) are mutually exclusive, then \( \Pr(A \text{ or } B) = \Pr(A) + \Pr(B) \).

5.3.2 Other Probability Rules

These rules all follow from the axioms.

The Addition Rule

If events \( A \) and \( B \) are mutually exclusive, then

\[
\Pr(A \text{ or } B) = \Pr(A) + \Pr(B) .
\]

More generally,

\[
\Pr(A \text{ or } B) = \Pr(A) + \Pr(B) - \Pr(A \text{ and } B) .
\]

\(^1\)This has actually been done a couple of times in history, including once by a mathematician who was a prisoner of war and had lots of time on his hands.
The Complement Rule

\[ \Pr(\text{not } A) = 1 - \Pr(A) \]

The Equally Likely Rule

If the sample space consists of \( n \) equally likely outcomes, then the probability of an event \( A \) is given by

\[ \Pr(A) = \frac{\text{number of outcomes in } A}{n} = \frac{|A|}{|S|}. \]

One of the most common mistakes in probability is to apply this rule when the outcomes are not equally likely.

Examples

1. Coin Toss: \( \Pr(\text{heads}) = \frac{1}{2} \) if heads and tails are equally likely.
2. Rolling a Die: \( \Pr(\text{even}) = \frac{3}{6} = \frac{1}{2} \) if the die is fair (each of the six numbers equally likely to occur).
3. Sum of two Dice: the sum is a number between 2 and 12, but these numbers are NOT equally likely. But there are 36 equally likely combinations of two dice:

\[
\begin{array}{cccccc}
1,1 & 2,1 & 3,1 & 4,1 & 5,1 & 6,1 \\
1,2 & 2,2 & 3,2 & 4,2 & 5,2 & 6,2 \\
1,3 & 2,3 & 3,3 & 4,3 & 5,3 & 6,3 \\
1,4 & 2,4 & 3,4 & 4,4 & 5,4 & 6,4 \\
1,5 & 2,5 & 3,5 & 4,5 & 5,5 & 6,5 \\
1,6 & 2,6 & 3,6 & 4,6 & 5,6 & 6,6 \\
\end{array}
\]

Let \( X \) be the sum of two dice.

- \( \Pr(X = 3) = \frac{2}{36} = \frac{1}{18} \)
- \( \Pr(X = 7) = \frac{6}{36} = \frac{1}{6} \)
- \( \Pr(\text{doubles}) = \frac{6}{36} = \frac{1}{6} \)

4. Punnet Squares

\[
\begin{array}{c|ccc}
 & A & a \\
\hline
A & AA & Aa \\
\hline
a & Aa & aa \\
\end{array}
\]

In an \( Aa \times Aa \) cross, if \( A \) is the dominant allele, then the probability of the dominant phenotype is \( \frac{3}{4} \), and the probability of the recessive phenotype is \( \frac{1}{4} \).
5.3.3 Conditional Probability and Independence

**Example.** Q. Suppose a family has two children and one of them is a boy. What is the probability that the other is a girl?

A. We’ll make the simplifying assumption that boys and girls are equally likely (which is not exactly true). Under that assumption, there are four equally likely families: BB, BG, GB, and GG. But only three of these have at least one boy, so our sample space is really \{BB, BG, GB\}. Of these, two have a girl as well as a boy. So the probability is 2/3 (see Figure 5.1).

![Figure 5.1](#)

GG  GB  BG  BB  probability = 2/3

Figure 5.1: Illustrating the sample space for Example 5.3.3

We can also think of this in a different way. In our original sample space of four equally likely families,

\[
\begin{align*}
\Pr(\text{at least one girl}) &= 3/4, \\
\Pr(\text{at least one girl and at least one boy}) &= 2/4, \text{ and} \\
\frac{2/4}{3/4} &= 2/3;
\end{align*}
\]

so 2/3 of the time when there is at least one boy, there is also a girl. We will denote this probability as \Pr(\text{at least one girl | at least one boy}). We’ll read this as “the probability that there is at least one girl given that there is at least one boy”. See Figure 5.2 and Definition 5.3.3.

![Figure 5.2](#)

Figure 5.2: A Venn diagram illustrating the definition of conditional probability. \Pr(A \mid B) is the ratio of the area of the football shaped region that is both shaded and striped \((A \cap B)\) to the area of the shaded circle \((B)\).

Let \(A\) and \(B\) be two events such that \(\Pr(B) \neq 0\). The **conditional probability** of \(A\) given \(B\) is defined by

\[
\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}.
\]

If \(\Pr(B) = 0\), then \(\Pr(A \mid B)\) is undefined.

**Example.** A class of 5th graders was asked to give their favorite color. The most popular color was blue. The table below contains a summary of the students’ responses:
Q. Suppose we randomly select a student from this class. Let $L$ be the event that a child’s favorite color is blue. Let $B$ be the event that the child is a boy, and let $G$ be the event that the child is a girl. Express each of the following probabilities in words and determine their values:

- $\Pr(L)$,
- $\Pr(L | B)$,
- $\Pr(B)$,
- $\Pr(L | G)$,
- $\Pr(G | L)$,
- $\Pr(B | G)$.

A. The conditional probabilities can be computed in two ways. We can use the formula from the definition of conditional probability directly, or we can consider the condition event to be a new, smaller sample space and read the conditional probability from the table.

- $\Pr(L) = \frac{17}{34} = \frac{1}{2}$ because 17 of the 34 kids prefer blue.
  This is the probability that a randomly selected student prefers blue.

- $\Pr(L | B) = \frac{10/34}{18/34} = \frac{10}{18}$ because 10 of the 18 boys prefer blue.
  This is the probability that a randomly selected boy prefers blue.

- $\Pr(B | L) = \frac{10/34}{17/34} = \frac{10}{17}$ because 10 of the 17 students who prefer blue are boys.
  This is the probability that a randomly selected student who prefers blue is a boy.

- $\Pr(L | G) = \frac{7/34}{16/34} = \frac{7}{16}$ because 7 of the 16 girls prefer blue.
  This is the probability that a randomly selected girl prefers blue.

- $\Pr(G | L) = \frac{7/34}{17/34} = \frac{7}{17}$ because 7 of the 17 kids who prefer blue are girls.
  This is the probability that a randomly selected kid who prefers blue is a girl.

- $\Pr(B | G) = \frac{0}{16/34} = 0$ because none of the girls are boys.
  This is the probability that a randomly selected girl is a boy.

One important use of conditional probability is as a tool to calculate the probability of an intersection.

Let $A$ and $B$ be events with non-zero probability. Then

\[
\Pr(A \cap B) = \Pr(A) \cdot \Pr(B | A) = \Pr(B) \cdot \Pr(A | B).
\]

This follows directly from the definition of conditional probability by a little bit of algebra and can be generalized to more than two events.
Example. Q. If you roll two standard dice, what is the probability of doubles? (Doubles is when the two numbers match.)

A. Let $A$ be the event that we get a number between 1 and 6 on the first die. So $\Pr(A) = 1$. Let $B$ be the event that the second number matches the first. Then the probability of doubles is $\Pr(A \cap B) = \Pr(A) \cdot \Pr(B \mid A) = 1 \cdot \frac{1}{6} = \frac{1}{6}$ since regardless of what is rolled on the first die, 1 of the 6 possibilities for the second die will match it.

Example. Q. A 5-card hand is dealt from a standard 52-card deck. What is the probability of getting a flush (all cards the same suit)?

A. Imagine dealing the cards in order. Let $A_i$ be the event that the $i$th card is the same suit as all previous cards. Then

\[
\Pr(\text{flush}) = \Pr(A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5) \\
= \Pr(A_1) \cdot \Pr(A_2 \mid A_1) \cdot \Pr(A_3 \mid A_1 \cap A_2) \cdot \Pr(A_4 \mid A_1 \cap A_2 \cap A_3) \cdot \Pr(A_5 \mid A_1 \cap A_2 \cap A_3 \cap A_4) \\
= \frac{1}{52} \cdot \frac{12}{51} \cdot \frac{11}{50} \cdot \frac{10}{49} \cdot \frac{9}{48}.
\]

Example. Q. Suppose a test correctly identifies diseased people 99% of the time and correctly identifies healthy people 98% of the time. Furthermore assume that in a certain population, one person in 1000 has the disease. If a random person is tested and the test comes back positive, what is the probability that the person has the disease?

A. We begin by introducing some notation. Let $D$ be the event that a person has the disease. Let $H$ be the event that the person is healthy. Let $+$ be the event that the test comes back positive (meaning it indicates disease – probably a negative from the perspective of the person tested). Let $-$ be the event that the test is negative.

- $\Pr(D) = 0.001$, so $\Pr(H) = 0.999$.
- $\Pr(+) \mid D) = 0.99$, so $\Pr(- \mid D) = 0.01$.

$\Pr(+) \mid D)$ is called the **sensitivity** of the test. (It tells how sensitive the test is to the presence of the disease.)
- $\Pr(- \mid H) = 0.98$, so $\Pr(+) \mid H) = 0.02$.

$\Pr(- \mid H)$ is called the **specificity** of the test.

- $\Pr(D \mid +) = \frac{\Pr(D \cap +)}{\Pr(+)} = \frac{\Pr(D) \cdot \Pr(+) \mid D) \Pr(D \cap +) + \Pr(H \cap +)}{\Pr(D \cap +) + \Pr(H \cap +)} = \frac{0.001 \cdot 0.99}{0.001 \cdot 0.99 + 0.999 \cdot 0.02} = 0.0472.$

A tree diagram is a useful way to visualize these calculations.

This low probability surprises most people the first time they see it. This means that if the test result of a random person comes back positive, the probability that that person has the disease is less than 9%, even though the test is “highly accurate”. This is one reason why we do not routinely screen an entire population for a rare disease – such screening would produce many more false positives than true positives.
Of course, if a doctor orders a test, it is usually because there are some other symptoms. This changes the \textit{a priori} probability that the patient has the disease.
hypothesis A statement that can be true or false

statistical hypothesis A hypothesis about the parameters of a population or random process

hypothesis testing Using data to assess statistical hypotheses

6.1 The Four Step Process

We will see many hypothesis tests over the rest of the semester. The details differ, but they all following the same 4 step outline:

1. State the Null and Alternative Hypotheses.
2. Compute a Test Statistic.
3. Compute a p-value.
4. Draw a conclusion.

Let’s illustrate with Example 6.2 (about the “handedness” of toads).

Step 1: State the Null and Alternative Hypotheses

Population of interest: all European *bufo bufo* toads.

Parameter of interest: \( p \) = the proportion of toads that exhibit right handed behavior.

Here are some example hypotheses about the parameter \( p \):

\( H_1: \ p = 0.5 \)
Toads are equally likely to be left- or right-“handed”.

\( H_2: \text{ If } p > 0.5 \)
More toads are right-handed than left-handed.
6.2 Hypothesis Testing

$H_3$: If $p < 0.5$

More toads are left-handed than right-handed.

$H_4$: $p \neq 0.5$

Toads are not equally likely to be left-handed or right-handed (but we aren’t making any claim about which “hand” is more likely to be preferred).

Hypothesis testing makes an important distinction between two hypotheses:

- **null hypothesis** a specific claim about the value of a parameter. It is made for purposes of argument – we are looking to see whether we have enough evidence in the data to reject the null hypothesis. Abbreviation: $H_0$

- **alternative hypothesis** typically a more general hypothesis representing what happens if the null hypothesis is false. Abbreviation: $H_a$ or $H_A$.

Students often get confused about which is the null and which is the alternative. If it helps you think about the following courtroom analogy. The null hypothesis is that the defendant is innocent, and that is the operating assumption under which the proceedings take place. Of course, the plaintiff does not believe this hypothesis, so the plaintiff brings data hoping to show that the null hypothesis of innocence is false. The jury will assess the data and render a verdict of “guilty” or “not guilty”:

- **guilty**: the evidence is strong enough to reject the assumption of innocence.
  - The data are not consistent (within reasonable doubt) with an innocent defendant.

- **not guilty**: the evidence is not strong enough to reject the assumption of innocence.
  - The data could reasonably have happened even if the defendant was innocent.

In our example,

$H_0$: $p = 0.50$

$H_a$: $p \neq 0.50$ (Or possibly $p > 0.50$ or $p < 0.50$; more on that choice later.)

**Step 2: Compute a Test Statistic**

In statistical hypothesis test, we summarize all the evidence with a single number called a **test statistic**. The toad researchers tested a sample of 18 toads and found 14 of them to be right-handed. So $x = 14$ is a good test statistic. If the null hypothesis is true, we would expect roughly half ($\approx 9$) of the toads to be right-handed. The farther our count is from 9, the stronger the evidence against the null hypothesis.

Note: We could also have used $\hat{p} = \frac{14}{18}$. We’ll return to this idea later.

**Step 3: Compute a p-value**

Since it would be possible to see 14 out of 18 toads be right-handed just by chance, we can’t know for sure if $H_0$ is true or false. What we really want to know is this:

Assuming the null hypothesis is true (since that is our working assumption),

What are the chances of seeing so many right-handed toads in a random sample of 18?
This is a probability question. To answer it we need to compare our test statistic to the null distribution

**null distribution** the distribution of the test statistic assuming the null hypothesis is true.

**p-value** the probability of obtaining a test statistic that is at least as extreme as the test statistic calculated from our data, *assuming the null hypothesis is true.*

**Empirical p-values**

If we can simulate what happens when the null hypothesis is true, we can calculate approximate p-values using a computer to perform the simulations many times.

```r
> rflip(18)
Flipping 18 coins [ Prob(Heads) = 0.5 ] ... 
H H H T T T H H T T H T T T T T T
Result: 8 heads.
> rflips <- do(100000) * nflip(18)
> xtabs(~ rflips) / 100000
rflips
   0   1   2   3   4   5   6   7   8   9  10
0.00001 0.00005 0.00059 0.00291 0.01161 0.03284 0.07078 0.12340 0.16595 0.18572 0.16625
   11  12  13  14  15  16  17  18
0.12136 0.06954 0.03340 0.01177 0.00307 0.00065 0.00009 0.00001
> histogram(~ rflips, breaks=-.5 + (0:19),
+ xlab="number of heads",
+ scales=list(x=list(at=0:18))
+ )
```

From this we see that the probability of having 14 or more right-handed toads (assuming toads are equally likely to be left- and right-handed) is approximately 0.01521 = 1.521%.

Similarly, the probability of having 14 or more left-handed toads (i.e., 4 or fewer right-handed toads) is approximately 0.01526 = 1.526%.

\[
p\text{-value} = \Pr(X \geq 14) \approx 0.01521 \quad \text{if } H_a \text{ is } p > 0.50
\]
\[
p\text{-value} = \Pr(X \geq 14 \text{ or } X \leq 4) \approx 0.01521 + 0.01526 = 0.03047 \quad \text{if } H_a \text{ is } p \neq 0.50
\]
\[
p\text{-value} = \Pr(X \leq 14) \approx 0.99625 \quad \text{if } H_a \text{ is } p < 0.50
\]
6.4 Hypothesis Testing

Theoretical P-values

If we can work out the probabilities for the null distribution theoretically, we can compute p-values without simulations. Table 6.2-1 shows these probabilities for this situation. We’ll learn how to calculate these probabilities soon. The (approximate) values from our simulations are pretty close to the theoretical (exact) values.

**Step 4: Draw a conclusion**

The smaller the p-value, the stronger the evidence against the null hypothesis. (Think of the p-value as a measure of the reasonableness of the doubt. The smaller the p-value, the less reasonable our doubt.) In this case, our (two-sided) p-value is \( \approx 0.03 \). What does this mean? One of the following:

1. The null hypothesis is false.
2. The null hypothesis is true, but we got rather unusual data.

Seeing such a large difference in the number of left- and right-handed toads would happen only about 3% of the time if there really are equal numbers of left- and right-handed toads.

We can’t know which if these is the case, but when the p-value is small enough, then we reject the null hypothesis based on our data because our data would be very unlikely if the null hypothesis were true.

6.2 Some Additional Details

6.2.1 Errors in Hypothesis Testing

There are two types of error we can make when making a decision about a null hypothesis:

<table>
<thead>
<tr>
<th></th>
<th>( H_0 ) is true</th>
<th>( H_0 ) is false</th>
</tr>
</thead>
<tbody>
<tr>
<td>reject ( H_0 )</td>
<td></td>
<td>Type I error</td>
</tr>
<tr>
<td>don’t reject ( H_0 )</td>
<td>good case</td>
<td>Type II error</td>
</tr>
</tbody>
</table>

6.2.2 How small is small?

Here is a table that can help you interpret p-values.

<table>
<thead>
<tr>
<th>p-value</th>
<th>interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 0.20 \leq ) p-value</td>
<td>no evidence against the null hypothesis</td>
</tr>
<tr>
<td>( 0.10 \leq ) p-value</td>
<td>perhaps enough evidence to warrant collecting more data</td>
</tr>
<tr>
<td>( 0.05 \leq ) p-value</td>
<td>suggestive evidence against the null hypothesis (probably want to confirm with additional data)</td>
</tr>
<tr>
<td>( 0.01 \leq ) p-value</td>
<td>evidence against the null hypothesis</td>
</tr>
<tr>
<td>( 0.001 \leq ) p-value</td>
<td>strong evidence against the null hypothesis</td>
</tr>
<tr>
<td>p-value ( \leq 0.001 )</td>
<td>very strong evidence against the null hypothesis</td>
</tr>
</tbody>
</table>
Often researchers will decide in advance what threshold (denoted $\alpha$ and called the significance level) will be used for the boundary between rejecting and not rejecting the null hypothesis. The most commonly used threshold is $\alpha = 0.05$. Another common threshold is $\alpha = 0.01$.

The choice of $\alpha$ determines the probability of type I error, so it should be selected in consideration of the consequences of making such an error.

### 6.2.3 What if the p-value is large?

When the p-value is large, our data are consistent with the null hypothesis (in the sense that they are not that unlikely when the null hypothesis is true). But this does not imply that the null hypothesis is indeed true. Especially when sample sizes are small, we may not have enough power to tell one way or the other.

For example, if we tested 4 toads and found 3 to be right handed, or p-value would be 0.625.

<table>
<thead>
<tr>
<th>number of right-handed toads</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>probability (assuming $p = 0.50$)</td>
<td>0.0625</td>
<td>0.25</td>
<td>0.375</td>
<td>0.25</td>
<td>0.0625</td>
</tr>
</tbody>
</table>

This is consistent with a null hypothesis of equal numbers of left- and right-handed toads, but it is also consistent with many other hypothesis. Of course it is also consistent with other hypotheses, like $p = .75$ or even $p = .85$. There just isn’t enough data say much about the value of $p$.

Soon we will learn about confidence intervals, which will give another way to express what we can learn about $p$ from our sample data.

### 6.2.4 Which Alternative?

Because 1-sided p-values are smaller than 2-sided p-values, it can be tempting to use them. To avoid this temptation, some statisticians recommend always using a two-sided alternative. Backing off this extreme position, here are two questions you must answer whenever you want to use a 1-sided alternative.

Can/should I use a 1-sided alternative?

1. Can I justify the choice of a 1-sided alternative without reference to the data?
   
   If the answer is No, you must use a 2-sided alternative.

2. If the data are extreme in the “wrong direction”, what will I conclude?
   
   If you feel like you should reject the null hypothesis in this case, you must use a 2-sided alternative.

In the case of the toads, if we don’t know whether we should expect more left-handed or more right-handed toads before we collect that data, we must use a 2-sided alternative. Even supposing we suspected more right-handed toads (based on analogy to humans or other animals, perhaps) we must ask what we would do if we find many more left-handed toads in our data. If we would find that scientifically interesting (and somewhat surprising), we should be doing a 2-sided test.

### 6.2.5 Statistical Significance and Scientific Significance

When the p-value is below our threshold for rejecting the null hypothesis, we say that our results are statistically significant. This simply means that data at least as extreme as our data are unlikely to occur by chance. It does not necessarily mean that anything interesting is going on biologically.
7.1 Binomial Distributions

7.1.1 The Binomial Model

We will say that $X$ is a binomial random variable if:

1. The random process consists of $n$ trials ($n$ is known in advance).
2. Each trial has two outcomes (traditionally called success and failure).
3. The probability of success ($p$) is the same for each trial.
4. The trials are independent.
5. The random variable $X$ counts the number of successes.

We will abbreviate this $X \sim \text{Binom}(n, p)$

The binomial model is a model of randomly sampling from a large population and recording the result of a two-level categorical variable for each observational unit.

7.1.2 Binomial Distributions in $\mathbb{R}$

Let $X \sim \text{Binom}(18, 0.5)$, then

- \text{rbinom(100, 18, 0.5)} simulates a $\text{Binom}(18, 0.5)$-random variable 100 times.

\begin{verbatim}
> rbinom(100,18,0.5)
[1] 6 11 7 11 11 7 9 11 9 11 11 6 12 11 11 2 5 9 11 6 11 9 6 12 7 6 10 11 5 9 11 10 14 8 6 8 11 8 12 10 7 10 10 9 7 9 5 10 6 8 5 7 7 8 6
[29] 7 10 11 9 9 8 9 12 9 11 6 8 7 10 12 7 6 10 14 8 7 7 5 12 10 11 11 11
[57] 9 8 9 9 13 6 12 10 9 9 10 6 10 9 8 8
[85]
\end{verbatim}

- \text{dbinom(14, 18, 0.5)} computes $\Pr(X = 14)$ when $X \sim \text{Binom}(18, 0.5)$.
7.2 Inference for a Proportion

```r
> dbinom(14,18,0.5)
[1] 0.01167297

• `pbinom(14, 18, 0.5)` computes \( \Pr(X \leq 14) \) when \( X \sim \text{Binom}(18, 0.5) \).

> pbinom(14,18,0.5)
[1] 0.996231

We can use `pbinom()` to compute the p-value for our toad study data. Recall that the 2-sided p-value is \( \Pr(X \geq 14 \text{ or } X \leq 4) = \Pr(X \leq 4) + \Pr(X \geq 14) \).

> pbinom(4,18,0.5) + 1 - pbinom(13,18,0.5)  # p-value from the toads study
[1] 0.03088379
```

7.1.3 The Binomial Formula

If \( X \sim \text{Binom}(n,p) \), then

\[
\Pr(X = x) = \Pr(x \text{ successes}) = \binom{n}{x} p^x (1-p)^{n-x}
\]

where

\[
\binom{n}{x} = \frac{n!}{x!(n-x)!}
\]

\( \binom{n}{x} \) is read “\( n \) choose \( x \)” and counts the number of ways to select a subset of size \( x \) from a set of size \( n \).

**Example.** Q. Compute \( \binom{5}{2} \).

A. Let the 5 choices be A, B, C, D, and E. Here is the list of subsets of size 2:

AB AC AD AE BC BD BE CD CE DE

There are ten, which matches our formula: \( \binom{5}{2} \frac{5!}{2!3!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 3 \cdot 2 \cdot 1} = 10 \).

**Example.** Let \( X \sim \text{Binom}(5, 0.25) \), then \( \Pr(X = 2) = \binom{5}{2}(0.25)^2(0.75)^3 = 10(0.25)^2(0.75)^3 = 0.2637 \).

```r
> dbinom(2,5,.25)
[1] 0.2636719

> factorial(5) / (factorial(2) * factorial(3)) * 0.25^2 * 0.75^3
[1] 0.2636719

> choose(5,2) * 0.25^2 * 0.75^3
[1] 0.2636719
```
7.2 The Binomial Test

The test that uses the binomial distribution to give a p-value for a test with hypotheses

- $H_0 : p = p_0$
- $H_a : p \neq p_0$ (or $p < p_0$ or $p > p_0$)

is called the binomial test.

R automates the entire process for us if we provide a summary of the data, the value $p_0$, and whether we want a 1-sided or 2-sided alternative:

```r
> binom.test(14,18,p=0.5) # 2-sided by default

Exact binomial test

data: 14 and 18
number of successes = 14, number of trials = 18, p-value = 0.03088
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval: 
0.5236272 0.9359080
sample estimates:
probability of success
0.7777778
```

There is more information in this output than we need at this point, but notice the statement of the alternative hypothesis (so know we have done a 2-sided test) and the p-value.

If we want to do a 1-sided test, R will happily do that for us as well:

```r
> binom.test(14,18,p=0.5,alternative='greater') # 1-sided test

Exact binomial test

data: 14 and 18
number of successes = 14, number of trials = 18, p-value = 0.01544
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
0.5611172 1.0000000
sample estimates:
probability of success
0.7777778
```

```r
> binom.test(14,18,p=0.5,alternative='less') # the other 1-sided test

Exact binomial test

data: 14 and 18
number of successes = 14, number of trials = 18, p-value = 0.9962
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
0.0000000 0.9203046
sample estimates:
probability of success
0.7777778
```

The word 'exact' in the output indicates that this method is not using simulations or approximations but is calculating the p-value exactly from the sampling distribution. Later we will learn some approximate methods for this same situation.
7.3 More Examples

Example. Here is R code for Example 7.2 in the text.

```r
> binom.test(10,25,p=.061)

  Exact binomial test

data: 10 and 25
number of successes = 10, number of trials = 25, p-value = 9.94e-07
alternative hypothesis: true probability of success is not equal to 0.061
95 percent confidence interval:
  0.2112548 0.6133465
sample estimates:
  probability of success
        0.4
```

Note that this p-value doesn’t match the one in the book. That is because R is using a more accurate method for two-sided tests. See the footnote on page 160.
A.1 Welcome to R and RStudio

R is a system for statistical computation and graphics. We will use R in this course for several reasons:

1. R is open-source and freely available for Mac, PC, and Linux machines.

2. R is user-extensible and user extensions can easily be made available to others.

3. R is commercial quality. It is the package of choice for many statisticians and those who use statistics frequently.

4. R is becoming very popular with biologists, especially in certain sub-disciplines, like genetics. Articles in research journals such as Science often include links to the R code used for the analysis and graphics presented.

5. R is very powerful. Furthermore, it is gaining new features every day. New statistical methods are often available first in R.

RStudio provides access to R in a web browser. The URL is

http://beta.rstudio.org/

You should be able to log in using your Calvin ID (myname@calvin.edu) and KnightVision password. Once you have logged in, you will see something like Figure A.1.

Notice that RStudio divides its world into four panels. Several of the panels are further subdivided into multiple tabs. The console panel is where we type commands that R will execute.
A.2 Using R as a Calculator

R can be used as a calculator. Try typing the following commands in the console panel.

\[
\begin{align*}
\text{> } 5 + 3 & \quad \text{[1] } 8 \\
\text{> } 15.3 \times 23.4 & \quad \text{[1] } 358.02 \\
\text{> } \sqrt{16} & \quad \text{[1] } 4
\end{align*}
\]

You can save values to named variables for later reuse:

\[
\begin{align*}
\text{> } \text{product} = 15.3 \times 23.4 & \quad \text{# save result} \\
\text{> } \text{product} & \quad \text{# show the result} \\
\text{[1] } 358.02 \\
\text{> } \text{product} & \quad \text{# <- is assignment operator, same as =} \\
\text{[1] } 358.02 \\
\text{> } 15.3 \times 23.4 \rightarrow \text{newproduct} & \quad \text{# -> assigns to the right} \\
\text{> } \text{newproduct} & \quad \text{# half of the product} \\
\text{[1] } 179.01 \\
\text{> } \log(\text{product}) & \quad \text{# (natural) log of the product} \\
\text{[1] } 5.880589 \\
\text{> } \log10(\text{product}) & \quad \text{# base 10 log of the product} \\
\text{[1] } 2.553907 \\
\text{> } \log(\text{product},\text{base}=2) & \quad \text{# base 2 log of the product} \\
\text{[1] } 8.483896
\end{align*}
\]

The semi-colon can be used to place multiple commands on one line. One frequent use of this is to save and print a value all in one go:
Three Things to Know About R

1. R is case-sensitive
   If you mis-capitalize something in R it won’t do what you want.

2. Functions in R use the following syntax:

   ```r
   > functionname( argument1, argument2, ... )
   ```

   - The arguments are always *surrounded by (round) parentheses* and *separated by commas*.
   - Some functions (like `data()`) have no required arguments, but you still need the parentheses.
   - If you type a function name without the parentheses, you will see the code for that function – which probably isn’t what you want at this point.

3. TAB completion and arrows can improve typing speed and accuracy.
   If you begin a command and hit the TAB key, R will show you a list of possible ways to complete the command. If you hit TAB after the opening parenthesis of a function, it will show you the list of arguments it expects. The up and down arrows can be used to retrieve past commands.

A.3 R packages

In addition to its core features, R provides many more features through a (large) number of packages. To use a package, it must be installed (one time), and loaded (each session). A number of packages are already available in RStudio. The Packages tab in RStudio will show you the list of installed packages and indicate which of these are loaded.

Here are some packages we will use frequently in this course:

- `lattice` (for graphics; this will always be installed in R)
- `Hmisc` (a package with some nice utilities; available on CRAN)
- `vcd` (a package for visualizing categorical data; available on CRAN)
- `fastR` (a package with some nice utilities; available on CRAN)
- `abd` (a package with data and utilities for our textbook; available on CRAN)
A.4 Getting Started with R

There may be others that we use from time to time as well. You should install \texttt{Hmisc}, \texttt{vcd}, \texttt{fastR}, and \texttt{abd} (in that order) the first time you use \texttt{R} so that they are always available to you. You can install these packages by clicking on the “Install Package” button and following the directions or by using the following commands:

\begin{verbatim}
> install.packages('Hmisc')        # note the quotation marks
> install.packages('vcd')
> install.packages('fastR')
> install.packages('abd')
\end{verbatim}

Once these are installed, you can load them by checking the box in the Packages tab or by using the commands

\begin{verbatim}
> require(lattice)
> require(Hmisc)
> require(fastR)
> require(abd)
\end{verbatim}

A.4 Data

A.4.1 Data in Packages

Many packages contain data sets. You can see a list of all data sets in all loaded packages using

\begin{verbatim}
> data()
\end{verbatim}

The \texttt{abd} package contains data sets from our text, \textit{The Analysis of Biological Data}. The \texttt{findData()} function can help you determine the correct name for the data set you are looking for.

\begin{verbatim}
> findData('human')       # all data sets with 'human' in the name
     name       chapter   type number sub
24 HumanGeneLengths    4 Example 1
43 HumanBodyTemp       11 Example 3
\end{verbatim}

\begin{verbatim}
> findData(2)            # all data sets in chapter 2
     name       chapter   type number sub
 1 TeenDeaths    2 Example 1 a
 2 DesertBirds   2 Example 1 b
 3 SockeyeFemales   2 Example 1 c
 4 GreatTitMalaria 2 Example 3
 5 Hemoglobin     2 Example 4
 6 Guppies        2 Example 5 a
 7 Lynx           2 Example 5 b
 8 EndangeredSpecies 2 Problem 6
 9 ShuttleDisaster 2 Problem 10
10 Convictions     2 Problem 16
11 ConvictionsAndIncome 2 Problem 17
12 Fireflies       2 Problem 18
13 NeotropicalTrees 2 Problem 23
\end{verbatim}

Typically you can use data sets by simply typing their names. But if you have already used that name for something or need to refresh the data after making some changes you no longer want, you can explicitly load the data using the \texttt{data()} function with the name of the data set you want.

\begin{verbatim}
> data(iris)
\end{verbatim}
A.4.2 Data Frames

Data sets are usually stored in a special structure called a **data frame**.

Data frames have a 2-dimensional structure.

- Rows correspond to **observational units** (people, animals, plants, or other objects we are collecting data about).
- Columns correspond to **variables** (measurements collected on each observational unit).

We’ll talk later about how to get your own data into R. For now we’ll use some data that comes with R and is all ready for you to use. The *iris* data frame contains 5 **variables** measured for each of 150 iris plants (the observational units). The *iris* data set is included with the default R installation. (Technically, it is located in a package called **datasets** which is always available.)

There are several ways we can get some idea about what is in the *iris* data frame.

```
> str(iris)
'data.frame': 150 obs. of 5 variables:
$ Sepal.Length: num 5.1 4.9 4.7 4.6 5.4 5.4 5.6 4.4 4.4 4.9 ...  
$ Sepal.Width : num 3.5 3.0 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...  
$ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.4 1.5 1.5 ...  
$ Petal.Width : num 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.2 0.1 ...  
$ Species : Factor w/ 3 levels "setosa","versicolor",...:
```

```
> summary(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
  Min. :4.300 Min. :2.000 Min. :1.000 Min. :0.100 setosa :50
  1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0.300 versicolor:50
  Median :5.800 Median :3.000 Median :4.350 Median :1.300 virginica :50
  Mean :5.843 Mean :3.057 Mean :3.758 Mean :1.199
  3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800
```

```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
  1 5.1         3.5         1.4       0.2 setosa
  2 4.9         3.0         1.4       0.2 setosa
  3 4.7         3.2         1.3       0.2 setosa
  4 4.6         3.1         1.5       0.2 setosa
  5 5.0         3.6         1.4       0.2 setosa
  6 5.4         3.9         1.7       0.4 setosa
```

In interactive mode, you can also try

```
> View(iris)
```

to see the data or

```
> ?iris
```
to get the documentation about for the data set.
Access to an individual variable in a data frame uses the $ operator in the following syntax:

```r
> dataframe$variable
```

For example,

```r
> iris$Sepal.Length
[1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9 5.4 4.8 4.8 4.3 5.8 5.7 5.4 5.1 5.7 5.1 5.4
[22] 4.4 5.0 5.1 4.8 5.1 4.6 5.3 5.0 7.0 6.4 6.9 5.5 6.5 5.7 6.3 4.9 6.6 5.2 5.0 5.9 6.0
[43] 6.1 5.6 6.7 5.6 5.8 6.2 5.6 5.9 6.1 6.3 6.1 6.4 6.6 6.8 6.7 6.0 5.7 5.5 5.5 5.8 6.0
[64] 5.4 6.0 6.7 6.3 5.6 5.5 6.1 5.8 5.0 5.6 5.7 5.7 6.2 5.1 5.7 6.3 5.8 7.1 6.3 6.5
[85] 7.6 4.9 7.3 6.7 7.2 6.5 6.4 6.8 5.7 5.8 6.4 6.5 7.7 7.7 6.0 6.9 5.6 7.7 6.3 6.7 7.2
[106] 6.2 6.1 6.4 7.2 7.4 7.9 6.4 6.3 6.1 7.7 6.3 6.4 6.0 6.9 6.7 6.9 5.8 6.8 6.7 6.7 6.3
[127] 6.5 6.2 5.9
```

shows the contents of the `Sepal.Length` variable. But this isn’t very useful for a large data set. We would prefer to compute numerical or graphical summaries. We’ll do that shortly.

### A.4.3 Using Your Own Data

RStudio will help you import your own data. To do so use the “Import Dataset” button in the **Workspace** tab. You can load data from text files, from the web, or from google spreadsheets.

**From Google.** The easiest of these is the Google spreadsheets option: Just click, select your spreadsheet, choose a name, and you’re done.

**From Excel,** you need to follow a 3-step process:

1. Save your Excel worksheet as a csv file.
2. Upload (in the **Files** tab) your csv file to the server, where you can create folders and store files in your personal account. (To share things with others: Put the files in your Public folder. Read the `AboutPublic.txt` file in that folder for directions.)
3. Now import “from a text file” in the **Workspace** tab.

In either case, be sure to do the following:

- Choose good variables names.
- Put your variables names in the first row.
- Use each subsequent row for one observational unit.
- Give the resulting data frame a good name.
### A.5 Summarizing Data

#### A.5.1 A Few Numerical Summaries

R includes functions that compute a wide range of numerical and graphical summaries. Most of the numerical summaries already familiar to you have obvious names. Here are a few examples. (If you don’t know what some of these – like IQR – are, don’t worry; we’ll be discussing them soon.)

```r
> mean(iris$Sepal.Length)
[1] 5.843333
> median(iris$Sepal.Length)
[1] 5.8
> sd(iris$Sepal.Length)
[1] 0.8280661
> quantile(iris$Sepal.Length)
   0%  25%  50%  75% 100% 
4.3  5.1  5.8  6.4  7.9
```

The `favstats()` function in the `abd` package computes several numerical summaries all at once.

```r
> require(abd) # if you haven't already loaded the abd package
> favstats(iris$Sepal.Length)
                       median    IQR     mean    sd     var
setosa       50 5.000000 0.400000 5.006000 0.3524897 0.1242490
versicolor   50 5.936000 0.700000 5.936000 0.5161711 0.2664327
virginica    50 6.588000 0.675000 6.588000 0.6358796 0.4043429
```

Here’s something a little fancier.

```r
> summary(Sepal.Length ~ Species, data=iris, fun=favstats)

  Sepal.Length N=150
+-------+----------+---+-----+-----+--------+---------+---------+
|      |        |N  |median|IQR  |mean   |sd      |var      |
+-------+----------+---+-----+-----+--------+---------+---------+
|Species|setosa   | 50|5.00 |0.400|5.006000|0.3524897|0.1242490|
|       |versicolor| 50|5.90 |0.700|5.936000|0.5161711|0.2664327|
|       |virginica | 50|6.50 |0.675|6.588000|0.6358796|0.4043429|
+-------+----------+---+-----+-----+--------+---------+---------+
|Overall|         |150|5.80 |1.300|5.843333|0.8280661|0.6856935|
```

(Note: This requires installation of two add-on packages: `Hmisc` and `abd`.)

#### A.5.2 Lattice Graphics

There are several ways to make graphs in R. I like to use a system called lattice graphics. The first step for using lattice is to load the lattice package using the check box in the Packages tab or using the following command:

```r
> require(lattice)
```

Lattice plots make use of a formula interface:

```r
> plotname( y ~ x | z, data=dataname, groups=grouping_variable, ...)
```
• Our most common plots have the following names:
  - `histogram`
  - `bwplot`
  - `xyplot`

• `x` is the name of the variable that is plotted along the horizontal (`x`) axis.

• `y` is the name of the variable that is plotted along the vertical (`y`) axis. (For some plots, this slot is empty because R computes these values from the values of `x`.)

• `z` is a conditioning variable used to split the plot into multiple subplots called **panels**.

• `grouping_variable` is used to display different groups differently (different colors or symbols, for example) within the same panel.

• ... There are many additional arguments to these functions that let you control just how the plots look. (But we’ll focus on the basics for now.)

### A.5.3 Histograms: `histogram()`

For a histogram, the `y` component of the formula is empty since we let R compute the heights of the bars for us.

```
> histogram(~Sepal.Length, data=iris)
```

![Histogram of Sepal.Length](image)

We can use a conditional variable to give us separate histograms for each species.

```
> histogram(~Sepal.Length|Species, data=iris)
```

![Separated histograms by species](image)
In lattice lingo, the three subplots are called panels and the labels at the top are called strips. (Strips can be placed on the left side if you prefer.)

### A.5.4 Boxplots: `bwplot()`

Here is a bad way to make side-by-side boxplots.

```r
> bwplot(~Sepal.Length|Species, data=iris)
```

![Boxplots for Sepal.Length by Species](image)

This is better, but the species names run into each other.

```r
> bwplot(Sepal.Length~Species, data=iris)
```

![Boxplots for Sepal.Length by Species](image)

We could fix that run-together text by using abbreviated names or rotating the labels 45 or 90 degrees. Instead of those solutions, we can also just reverse the roles of the horizontal and vertical axes.

```r
> bwplot(Species ~ Sepal.Length, data=iris)
```

![Boxplots for Species by Sepal.Length](image)
A.5.5 Scatterplots: \texttt{xyplot()}

Scatterplots are made with \texttt{xyplot()}. The formula interface is very natural for this. Just remember that the “\textit{y} variable” comes first. (Its label is also farther left on the plot, if that helps you remember.)

\begin{verbatim}
> xyplot(Sepal.Length ~ Sepal.Width, data=iris)
\end{verbatim}

![Scatterplot of Sepal.Length vs Sepal.Width](image1)

Again, we can use conditioning to make a panel for each species.

\begin{verbatim}
> xyplot(Sepal.Length ~ Sepal.Width|Species, data=iris)
\end{verbatim}

![Scatterplot panel for each species](image2)

Even better (for this example), we can use the \texttt{groups} argument to indicate the different species using different symbols on the same panel.

\begin{verbatim}
> xyplot(Sepal.Length ~ Sepal.Width, groups=Species, data=iris)
\end{verbatim}

![Scatterplot with different symbols for each species](image3)
A.5.6 Saving Your Plots

There are several ways to save plots, but the easiest is probably the following:

1. In the Plots tab, click the “Export” button.
2. Copy the image to the clipboard using right click.
3. Go to your Word document and paste in the image.
4. Resize or reposition your image in Word as needed.

A.5.7 A Few Bells and Whistles

There are lots of arguments that control how these plots look. Here are just a few examples.

auto.key

It would be useful to have a legend for the previous plot. auto.key=TRUE turns on a simple legend. (There are ways to have more control, if you need it.)

\[
\text{\texttt{xyplot(Sepal.Length ~ Sepal.Width, groups=Species, data=iris,}}
\text{\hspace{1cm} + auto.key=TRUE)}}
\]

alpha, cex

Sometimes it is nice to have elements of a plot be partly transparent. When such elements overlap, they get darker, showing us where data are “piling up.” Setting the \texttt{alpha} argument to a value between 0 and 1 controls the degree of transparency: 1 is completely opaque, 0 is invisible. The \texttt{cex} argument controls “character expansion” and can be used to make the plotting “characters” larger or smaller by specifying the scaling ratio.

\[
\text{\texttt{xyplot(Sepal.Length ~ Sepal.Width, groups=Species, data=iris,}}
\text{\hspace{1cm} + auto.key=list(columns=3),}}
\text{\hspace{1cm} + alpha=.5,}}
\text{\hspace{1cm} + cex=1.3)}}
\]
A.12 Getting Started with R

You can add a title or subtitle, or change the default labels of the axes.

```r
> xyplot(Sepal.Length ~ Sepal.Width, groups=Species, data=iris,
+     main="Some Iris Data",
+     sub="(R. A. Fisher analyzed this data in 1936)",
+     xlab="sepal width (cm)",
+     ylab="sepal length (cm)",
+     alpha=.5,
+     auto.key=list(columns=3))
```

```
Some Iris Data

setosa ● versicolor ○ virginica ●
```

```r
trellis.par.set()
```

Default settings for lattice graphics are set using `trellis.par.set()`. Don’t like the default font sizes? You can change to a 7 point (base) font using

```r
> trellis.par.set(fontsize=list(text=7)) # base size for text is 7 point
```

Nearly every feature of a lattice plot can be controlled: fonts, colors, symbols, line thicknesses, colors, etc. Rather than describe them all here, we’ll mention only that groups of these settings can be collected into a theme. `show.settings()` will show you what the theme looks like.
> trellis.par.set(theme=col.whitebg())  # a theme in the lattice package
> show.settings()

> trellis.par.set(theme=col.abd())  # a theme in the abd package
> show.settings()
> trellis.par.set(theme=col.fastR())  
> show.settings()  

> trellis.par.set(theme=col.fastR(bw=TRUE))  
> show.settings()
A.5.8 Tabulating Categorical Data

The Trematodes data set contains data from an experiment to see if fish infected with a certain parasite are more likely to be eaten by birds (because they swim closer to the surface of the water). Highly infected, lightly infected, and uninfected fish were placed in an outdoor tank that was open to foraging birds. For each fish, the researchers recorded its infection status and whether it had been eaten.

\[
\text{head(Trematodes, 3)}
\]

```r
infection.status eaten
1 uninfected yes
2 uninfected no
2.1 uninfected no
```

Making Frequency and Contingency Tables with `xtabs()`

Categorical variables are often summarized in a table. R can make a table for a categorical variable using `xtabs()`.

\[
\text{xtabs(~infection.status, Trematodes)}
\]

```r
infection.status
  high  light  uninfected
  46     45     50
```

\[
\text{xtabs(~eaten, Trematodes)}
\]

```r
eaten
  no yes
  93  48
```

We can make a cross-table (also called a contingency table or a two-way table) summarizing this data with `xtabs()`. This is often a more useful view of data with two categorical variables.

\[
\text{xtabs(~ infection.status + eaten, Trematodes)}
\]

```r
eaten
  infection.status no yes
    high   9  37
    light  35 10
  uninfected 49  1
```

Entering Tables by Hand

Because categorical data is so easy to summarize in a table, often the frequency or contingency tables are given instead. You can enter these tables manually as follows:

\[
\text{myinfection <- c( high=46, light=46, uninfected=50); myinfection}
\]

```r
myinfection <- c(high=46, light=46, uninfected=50)
```

> myinfection <- c(high=46, light=46, uninfected=50)

```r
high light uninfected
  46  46     50
```
> mycrosstable <- rbind( # bind row-wise
  high = c(no=9, yes=37), # note labels
  light = c(35,10), # don’t need to relabel columns
  uninfected = c(49,1) # note commas throughout
)

> mycrosstable

<table>
<thead>
<tr>
<th></th>
<th>no</th>
<th>yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
<td>9</td>
<td>37</td>
</tr>
<tr>
<td>light</td>
<td>35</td>
<td>10</td>
</tr>
<tr>
<td>uninfected</td>
<td>49</td>
<td>1</td>
</tr>
</tbody>
</table>

Replacing `rbind()` with `cbind()` will allow you to give the data column-wise instead.

### A.5.9 Graphing Categorical Data

The `lattice` function `barchart` can display these tables as barcharts.

```r
> barchart(xtabs(~infection.status, Trematodes))
```

![Barchart](image)

```r
> barchart(xtabs(~infection.status, Trematodes), horizontal=FALSE) # vertical bars
```

If the data are already summarized, we can make our bar charts without using `xtabs()` (Figure A.2).

```r
> barchart(cause~deaths,TeenDeaths,horizontal=TRUE)
```

Just as bar charts and pie charts are used to display the distribution of one categorical variable, mosaic plots can do the same for cross tables. `mosaic()` (from the `vcd` package) is not a `lattice` plot, but it does use a similar formula interface.

```r
> require(vcd) # load the visualing categorical data package
> mosaic(~ infection.status + eaten, Trematodes)
```

Alternatively, we can send `mosaic()` the output of `xtabs()`:

Last Modified: February 23, 2011
> mosaic(xtabs(~ infection.status + eaten, Trematodes))

Or we can send our own hand-made table (although the output isn’t quite as nice without some extra effort we won’t discuss just now):

> mosaic(mycrosstable)

Barcharts can also be used to display two-way tables. First we convert the cross-table to a data frame. Then we can use this data frame for plotting.

> trem <- as.data.frame( xtabs(~ infection.status + eaten, data=Trematodes) ); trem

<table>
<thead>
<tr>
<th>infection.status</th>
<th>eaten</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
<td>no</td>
<td>9</td>
</tr>
<tr>
<td>light</td>
<td>no</td>
<td>35</td>
</tr>
<tr>
<td>uninfected</td>
<td>no</td>
<td>49</td>
</tr>
<tr>
<td>high</td>
<td>yes</td>
<td>37</td>
</tr>
<tr>
<td>light</td>
<td>yes</td>
<td>10</td>
</tr>
<tr>
<td>uninfected</td>
<td>yes</td>
<td>1</td>
</tr>
</tbody>
</table>

> barchart( Freq ~ infection.status, groups=eaten, data=trem)
A.6 Getting Help

A.6.1 ?

To get help on a specific function or data set, simply precede its name with a ?:

> ?col.whitebg()

A.6.2 apropos()

If you don’t know the exact name of a function, you can give part of the name and R will find all functions that match. Quotation marks are mandatory here.

> apropos('hist') # must include quotes, single or double.

[1] "event.history"       "hist"
[3] "hist.data.frame"     "hist.default"
[5] "hist.FD"             "hist.scott"
[7] "histbackback"        "histochart"
[9] "histogram"           "histogram"
[11] "history"             "histSpike"
[13] "ldahist"             "loadhistory"
[15] "panel.histogram"     "panel.xhistogram"
[17] "pmfhistogram"        "prepanel.default.histogram"
[19] "savehistory"         "truehist"
[21] "xhistogram"

A.6.3 ??

If that fails, you can do a broader search using ??, which will find matches not only in the names of functions and data sets, but also in the documentation for them. Quotations marks are optional here.

> ??histogram # any of these will work
> ??"histogram"
> ??'histogram'

A.6.4 Demos

The abd package also includes demos for some of the sections of the book. These demos show you how to make the plots or perform the analyses in that section, so we don’t have to repeat them all here. This also makes it easy to copy and paste the commands within RStudio if you need to do something similar.

Here’s how demo() works.
> demo(sec2.2)
> demo(sec2.2)
> # figure 2.2-1
> cumfreq(~count,DesertBirds, xlab='Species Abundance')

You can get a list of all available demos using

> demo()

### A.7 Additional Notes on R Syntax

#### A.7.1 Text and Quotations

For the most part, text in R must be enclosed in either single or double quotations. It usually doesn’t matter which you use, unless you want one or the other type of quotation mark inside your text. Then you should use the other type of quotation mark to mark the beginning and the end.

> text1 <- "Mary didn’t come" # apostrophe inside requires double quotes around text
> text2 <- 'Do you use "scare quotes"?' # this time we flip things around

If you omit quotes, you will often see error messages telling you that R can’t find an object because R will look for a function, data set or other object with that name instead of treating your text as text.

> text3 <- blah
> Error: object 'blah' not found

#### A.7.2 Functions

Functions in R use the following syntax:

> functionname( argument1, argument2, ... )

- The arguments are always surrounded by (round) parentheses and separated by commas.
  - Some functions (like `col.whitebg()`) have no arguments, but you still need the parentheses.
- Most arguments have names, but you don’t need to use the names if you give the arguments in the correct order.
  
  If you use names, you can give the arguments out of order. The following do the same thing,
> xyplot(Sepal.Length ~ Sepal.Width, data=iris, groups=Species)
> xyplot(Sepal.Length ~ Sepal.Width, iris, groups=Species)
> xyplot(Sepal.Length ~ Sepal.Width, groups=Species, iris)

But these do not work

> xyplot(Sepal.Length ~ Sepal.Width, Species, iris)
> xyplot(Sepal.Length ~ Sepal.Width, iris, Species)

The first fails because the second argument is data, so iris needs to be in the second position if it is not named. The second fails because groups is not the third argument. (There are many other arguments between data and groups.) The documentation for functions shows the correct order of arguments.

- Typically, we will not use names for the first argument or two (these tend to be very important arguments that have to be there) but will use names for the rest (these are often optional arguments that can be present or not, depending on whether we want the default behavior or something special).

A.8 Installing R

A.8.1 RStudio in the cloud

Our primary version of R will be the online version of RStudio. You should have an RStudio account at [http://beta.rstudio.org/](http://beta.rstudio.org/) using your Calvin student email address and password. RStudio is a brand new (and quite nice) interface to R that runs in a web browser. This has the advantage that you don’t have to install or configure anything. Just login and you are good to go. Furthermore, RStudio will “remember” what you were doing so that each time you login (even on a different machine) you can pick up right where you left off. This is “R in the cloud” and works a bit like GoogleDocs for R.

If you find bugs or have suggestions for RStudio, let me know. It is in rapid development at the moment, and I can pass along your feedback to the developers. You are among the first to use RStudio, and this is your chance to improve the software!

This should be all you need for this course. But if you prefer to have a stand-alone version (because you study somewhere without an internet connection, for example), read on.

A.8.2 RStudio on Your Desktop/Laptop

There is also a stand-alone version of the RStudio that you can install on your desktop or laptop machine. I’ll send you email with instructions indicating where to find it. (It is currently in beta and has not been released to the general public.)

A.8.3 Getting R from CRAN

CRAN is the Comprehensive R Archive Network [http://cran.r-project.org/](http://cran.r-project.org/). You can download free versions of R for PC, Mac, and Linux from CRAN. (If you use the RStudio stand-alone version, you also need to install R this way first.) All the instructions for downloading and installing are on CRAN. Just follow the appropriate instructions for your platform.
Exercises

For these problems, create a single Word document containing all of your work. Be sure to put a nice header at the top indicating (a) your name, (b) the course (Math 143B), (c) my name (Pruim), (d) and the assignment number (Lab 1).

1 Calculate the natural logarithm (log base $e$) and base 10 logarithm of 12,345. Copy-and-paste your R code into your Word document. Be sure to use a fixed-width font (like Courier) when you display R code. This will (1) make it clear that you are displaying R input or output, and (2) keep things aligned properly.

2 Install and load the following packages: Hmisc, vcd, fastR, and abd. Make sure lattice is also loaded (no need to install it, it is already installed).

3 Enter the following small data set in an Excel or Google spreadsheet and import the data into RStudio. You can import directly from Google. From Excel, save the file as a csv and import that (as a text file) into RStudio. Name the data frame JunkData. Copy and paste the results of

   > JunkData

into your report. (Remember to use a fixed width font like Courier for all R input and output.)

4 What is the average (mean) width of the sepals in the iris data set?

5 Determine the average (mean) sepal width for each of the three species in the iris data set.

6 The Jordan8687 data set (in the fastR package) contains the number of points Michael Jordan scored in each game of the 1986–87 season.

   a) Make a histogram of this data. Add an appropriate title.

   b) How would you describe the shape of the distribution?

   c) In approximately what percentage of his games, did Michael Jordan score less than 20 points? More than 50? (You may want to add breaks=seq(0,70,by=5) to your command to neaten up the bins.)
7 Cuckoos lay their eggs in the nests of other birds. Is the size of cuckoo eggs different in different host species nests? The cuckoo data set (in fastR) contains data from a study attempting to answer this question.

   a) When were these data collected? (Use ?cuckoo to get information about the data set.)
   b) What are the units on the length measurements?
   c) Make side-by-side boxplots of the length of the eggs by species.
   d) Calculate the mean length of the eggs for each host species.
   e) What do you think? Does it look like the size is differs among the different host species? Refer to your R output as you answer this question. (We’ll learn formal methods to investigate this later in the semester.)

8 The utilities2 data set in the fastR package contains a number of variables about the utilities bills at a residence in Minnesota over a number of years. Since the number of days in a billing cycle varies from month to month, variables like gasbillpday (elecbillpday, etc.) contain the gasbill (electric bill, etc.) divided by the number of days in the billing cycle.

   a) Make a scatter plot of gasbillpday vs. monthsSinceY2K using the command

   > xyplot(gasbillpday ~ monthsSinceY2K, data=utilities2, type='l') # the letter l

   What pattern(s) do you see?
   b) What does type='l' do? Make your plot with and without it. Which is easier to read in this situation?
   c) What happens if we replace type='l' with type='b'?
   d) Make a scatter plot of gasbillpday by month. What do you notice?
   e) Make any other plot you like using this data. Include both a copy of your plot and a discussion of what you can learn from it.

9 The table below is from a study of nighttime lighting in infancy and eyesight (later in life).

<table>
<thead>
<tr>
<th></th>
<th>no myopia</th>
<th>myopia</th>
<th>high myopia</th>
</tr>
</thead>
<tbody>
<tr>
<td>darkness</td>
<td>155</td>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>nightlight</td>
<td>153</td>
<td>72</td>
<td>7</td>
</tr>
<tr>
<td>full light</td>
<td>34</td>
<td>36</td>
<td>3</td>
</tr>
</tbody>
</table>

   a) Do you think this was an experiment or an observational study? Why?
   b) Recreate the table in RStudio. Copy and paste the results into your Word document.
   c) What percent of the subjects slept with a nightlight as infants?

   There are several ways to do this. You could use R as a calculator to do the arithmetic. You can save some typing if you use the function prop.table(). See ?prop.table for documentation. If you just want row and column totals added to the table, see mar.table() in the vcd package.
   d) Make a mosaic plot for this data. What does this plot reveal?
R Examples

The commands below are illustrated with the data sets iris and Trematode. To apply these in other situations, you will need to substitute the name of your data frame and the variables in it.

```
answer <- 42          # Store the value 42 in a variable named answer.

log(123); log10(123); sqrt(123)  # Take natural logarithm, base 10 logarithm, or square root of 123.

x <- c(1,2,3)         # Make a variable containing values 1, 2, and 3 (in that order).

data(iris)            # (Re)load the data set iris.

findData(2)           # Find abd data in chapter 2.

summary(iris$Sepal.Length) # Summarize the distribution of the Sepal.Length variable in the iris data frame.

summary(iris)         # Summarize each variable in the iris data frame.

str(iris)             # A different way to summarize the iris data frame.

head(iris)            # First few rows of the data frame iris.

require(Hmisc)
require(abd)          # Load packages. (This can also be done by checking boxes in the Packages tab.)

summary(Sepal.Length~Species,data=iris,fun=favstats) # Compute favorite statistics of Sepal.Length for each Species. [requires Hmisc and abd]

histogram(~Sepal.Length|Species,iris) # Histogram of Sepal.Length conditioned on Species.

bwplot(Sepal.Length~Species,iris) # Boxplot of Sepal.Length conditioned on Species.

xyplot(Sepal.Length~Sepal.Width|Species,iris) # Scatterplot of Sepal.Length by Sepal.Width with separate panels for each Species.

xtabs(~infection.status,Trematodes) # Frequency table of the variable infection.status.

barchart(xtabs(~infection.status,Trematodes)) # Make a barchart from the table.

xtabs(~infection.status + eaten,Trematodes) # Cross tabulation of infection.status and eaten.

mosaic(~infection.status + eaten,Trematodes)
```
Make a mosaic plot.

```r
xtData <- as.data.frame(xtabs(~eaten + infection.status,Trematodes))
```

Save cross table information as `xtData`.

```r
barchart(Freq~infection.status,data=xtData,groups=eaten)
```

Use `xtData` to make a segmented bar chart.

```r
sum(x); mean(x); median(x);
var(x); sd(x); quantile(x)
```

Sum, mean, median, variance, standard deviation, quantiles of `x`. 


