Introductory Statistics Notes

Jamie DeCoster

Department of Psychology University of Alabama 348 Gordon Palmer Hall Box 870348 Tuscaloosa, AL 35487-0348

Phone: (205) 348-4431 Fax: (205) 348-8648

August 1, 1998

These were compiled from Jamie DeCoster's introductory statistics class at Purdue University. Textbook references refer to Moore's *The Active Practice of Statistics*. CD-ROM references refer to Velleman's *ActivStats*. If you wish to cite the contents of this document, the APA reference for them would be

DeCoster, J. (1998). Introductory Statistics Notes. Retrieved <month, day, and year you downloaded this file> from http://www.stat-help.com/notes.html

For help with data analysis visit http://www.stat-help.com

ALL RIGHTS TO THIS DOCUMENT ARE RESERVED.

Contents

Ι	Understanding Data	2
1	Introduction	3
2	Data and Measurement	4
3	The Distribution of One Variable	5
4	Measuring Center and Spread	7
5	Normal Distributions	10
II	Understanding Relationships	12
6	Comparing Groups	13
7	Scatterplots	14
8	Correlation	15
9	Least-Squares Regression	16
9	Association vs. Causation	18
Π	I Generating Data	19
10) Sample Surveys	20
11	Designed Experiments	21
I	V Experience with Random Behavior	23
12	2 Randomness	24
13 Intuitive Probability		25
14	14 Conditional Probability	
15	15 Random Variables	
16	16 Sampling Distributions	

V Statistical Inference	29
17 Estimating With Confidence	30
18 Confidence Intervals for a Mean	31
19 Testing Hypotheses	32
20 Tests for a Mean	34
VI Topics in Inference	36
21 Comparing Two Means	37
22 Inference for Proportions	39
23 Two-Way Tables	41
24 Inference for Regression	42
25 One-Way Analysis of Variance	45

Part I

Understanding Data

Introduction

- It is important to know how to understand statistics so that we can make the proper judgments when a person or a company presents us with an argument backed by data.
- *Data* are numbers with a context. To properly perform statistics we must always keep the meaning of our data in mind.
- You will spend several hours every day working on this course. You are responsible for material covered in lecture, as well as the contents of the textbook and the CD-ROM. You will have homework, CD-ROM, and reading assignments every day. It is important not to get behind in this course. A good work schedule would be:
 - Review the notes from the previous day's lecture, and take care of any unfinished assignments.
 - Attend the lecture.
 - $\circ~$ Attend the lab section.
 - Do your homework. You will want to plan on staying on campus for this, as your homework will often require using the CD-ROM.
 - Do the CD-ROM assignments.
 - $\circ\,$ Do the Reading assignments.

This probably seems like a lot of work, and it is. This is because we need to cover 15 weeks of material in 4 weeks during Maymester. Completing the course will not be easy, but I will try to make it as good an experience as I can.

Data and Measurement

- Statistics is primarily concerned with how to summarize and interpret *variables*. A variable is any characteristic of an object that can be represented as a number. The values that the variable takes will vary when measurements are made on different objects or at different times.
- Each time that we record information about an object we observe a *case*. We might include several different variables in the same case. For example, we might measure the height, weight, and hair color of a group of people in an experiment. We would have one case for each person, and that case would contain that person's height, weight, and hair color values. All of our cases put together is called our *data set*.
- Variables can be broken down into two types:
 - *Quantitative variables* are those for which the value has numerical meaning. The value refers to a specific amount of some quantity. You can do mathematical operations on the values of quantitative variables (like taking an average). A good example would be a person's height.
 - *Categorical variables* are those for which the value indicates different groupings. Objects that have the same value on the variable are the same with regard to some characteristic, but you can't say that one group has "more" or "less" of some feature. It doesn't really make sense to do math on categorical variables. A good example would be a person's gender.
- Whenever you are doing statistics it is very important to make sure that you have a practical understanding of the variables you are using. You should make sure that the information you have truly addresses the question that you want to answer.

Specifically, for each variable you want to think about who is being measured, what about them is being measured, and why the researcher is conducting the experiment. If the variable is quantitative you should additionally make sure that you know what units are being used in the measurements.

The Distribution of One Variable

- The pattern of variation of a variable is called its *distribution*. If you examined a large number of different objects and graphed how often you observed each of the different values of the variable you would get a picture of the variable's distribution.
- *Bar charts* are used to display distributions of categorical variables. In a bar chart different groups are represented on the horizontal axis. Over each group a bar is drawn such that the height of the bar represents the number of cases falling in that group.
- A good way to describe the distribution of a quantitative variable is to take the following three steps:
 - 1. Report the center of the distribution.
 - 2. Report the general shape of the distribution.
 - 3. Report any significant deviations from the general shape.
- A distribution can have many different shapes. One important distinction is between *symmetric* and *skewed* distributions. A distribution is symmetric if the parts above and below its center are mirror images. A distribution is skewed to the right if the right side is longer, while it is skewed to the left if the left side is longer.
- Local peaks in a distribution are called *modes*. If your distribution has more than one mode it often indicates that your overall distribution is actually a combination of several smaller ones.
- Sometimes a distribution has a small number of points that don't seem to fit its general shape. These points are called *outliers*. It's important to try to explain outliers. Sometimes they are caused by data entry errors or equipment failures, but other times they come from situations that are different in some important way.
- Whenever you collect a set of data it is useful to plot its distribution. There are several ways of doing this.
 - For relatively small data sets you can construct a *stemplot*. To make a stemplot you
 - 1. Separate each case into a stem and a leaf. The stem will contain the first digits and the leaf will contain a single digit. You ignore any digits after the one you pick for your leaf. Exactly where you draw the break will depend on your distribution: Generally you want at least five stems but not more than twenty.
 - 2. List the stems in increasing order from top to bottom. Draw a vertical line to the right of the stems.
 - 3. Place the leaves belonging to each stem to the right of the line, arranged in ascending numerical order.

- To compare two distributions you can construct *back-to-back stemplots*. The basic procedure is the same as for stemplots, except that you place lines on the left and the right side of the stems. You then list out the leaves from one distribution on the right, and the leaves from the other distribution on the left.
- For larger distributions you can build a *histogram*. To make a histogram you
 - 1. Divide the range of your data into classes of equal width. Sometimes there are natural divisions, but other times you need to make them yourself. Just like in a stemplot, you generally want at least five but less than twenty classes.
 - 2. Count the number of cases in each class. These counts are called the *frequencies*.
 - 3. Draw a plot with the classes on the horizontal axis and the frequences on the vertical axis.
- *Time plots* are used to see how a variable changes over time. You will often observe cycles, where the variable regularly rises and falls within a specific time period.

Measuring Center and Spread

- In this section we will discuss some ways of generating mathematical summaries of distributions. In these equations we will make use of some stastical notation.
 - \circ We will always use *n* to refer to the total number of cases in our data set.
 - When referring to the distribution of a variable we will use a single letter (other than n). For example we might use h to refer to height. If we want to talk about the value of the variable in a specific case we will put a subscript after the letter. For example, if we wanted to talk about the height of person 5 in our data set we would call it h_5 .
 - Several of our formulas involve *summations*, represented by the \sum symbol. This is a shorthand for the sum of a set of values. The basic form of a summation equation would be

$$x = \sum_{i=a}^{b} f(i).$$

To determine x you calculate the function f(i) for each value from i = a to i = b and add them together. For example,

$$\sum_{i=2}^{4} i^2 = 2^2 + 3^2 + 4^2 = 29.$$

• Some important properties of summation:

$$\sum_{i=1}^{n} k = nk, \text{ where } k \text{ is a constant}$$

$$(4.1)$$

$$\sum_{i=1}^{n} (Y_i + Z_i) = \sum_{i=1}^{n} Y_i + \sum_{i=1}^{n} Z_i$$
(4.2)

$$\sum_{i=1}^{n} (Y_i + k) = \sum_{i=1}^{n} Y_i + nk, \text{ where } k \text{ is a constant}$$

$$(4.3)$$

$$\sum_{i=1}^{n} (kY_i) = k \sum_{i=1}^{n} Y_i, \text{ where } k \text{ is a constant}$$

$$(4.4)$$

- When people want to give a simple description of a distribution they typically report measures of its central tendency and spread.
- People most often report the *mean* and the *standard deviation*. The main reason why they are used is because of their relationship to the *normal distribution*, an important distribution that will be discussed in the next lesson.

• The mean represents the average value of the variable, and can be calculated as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$
(4.5)

- In statistics our sums are almost always over all of the cases, so we will typically not bother to write out the limits of the summation and just assume that it is performed over all the cases. So we might write the formula for the mean as $\frac{1}{n} \sum x_i$.
- The standard deviation is a measure of the average difference of each case from the mean, and can be calculated as

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}.$$
(4.6)

- The term $x_i \bar{x}$ is the deviation of an case from the mean. We square it so that all the deviation scores will come out positive (otherwise the sum would always be zero). We then take the square root to put it back on the proper scale.
- You might expect that we would divide the sum of our deviations by n instead of n-1, but this is not the case. The reason has to do with the fact that we use \bar{x} , an approximation of the mean, instead of the true mean of the distribution. Our values will be somewhat closer to this approximation than they will be to the true mean because we calculated our approximation from the same data. To correct for this underestimate of the variability we only divide by n-1 instead of n. Statisticians have proven that this correction gives us an unbiased estimate of the true standard deviation.
- While the mean and standard deviation provide some important information about a distribution, they may not be meaningful if the distribution has outliers or is strongly skewed. People will therefore often report the *median* and the *interquartile range* of a variable.
 - These are called *resistant* measures of central tendency and spread because they are not strongly affected by skewness or outliers.
 - $\circ~$ The formal procedure to calculate the median is:
 - 1. Arrange the cases in a list ordered from lowest to highest.
 - 2. If the number of cases is odd, the median is the center obsertation in the list. It is located $\frac{n+1}{2}$ places down the list.
 - If the number of cases is even, the median is the average of the two cases in the center of the list. These are located ⁿ/₂ and ⁿ⁺²/₂ places down the list.
 - The median is the 50th percentile of the list. This means that it is greater than 50% of the cases in the list. The interquartile range is the difference between the cases in the 25th and the 75th percentiles. To calculate the interquartile range you
 - 1. Calculate the value of the 25th percentile (also called the first quartile). This is the median of the cases below the location of the 50th percentile.
 - 2. Calculate the value of the 75th percentile (also called the third quartile). This is the median of the cases above the location of the 50th percentile.
 - 3. Subtract the value of the 25th percentile from the value of the 75th percentile. This is the interquartile range.
 - To get a more accurate description of a distribution, people will often provide its *five-number* summary. This consists of the mininum, first quartile, median, third quartile, and the maximum.
- Often times the same theoretical construct (such as temperature) might be measured in different units (such as Farenheit and Celcius). These variables are often *linear transformations* of each other, in that you can calculate one of the variables from another with an equation of the form

$$x^* = a + bx. \tag{4.7}$$

If you have the summary statistics (mean, standard deviation, median, interquartile range) of the original variable x it is easy to calculate the summary statistics of the transformed variable x^* .

- $\circ~$ The mean, median, and quartiles of the transformed variable can be obtained by multiplying the originals by b and then adding a.
- $\circ\,$ The standard deviation and interquartile range of the transformed variable can be obtained by multiplying the originals by b.

Linear transformations do not change the overall shape of the distribution, although they will change the values over which the distribution runs.

Normal Distributions

- A *density curve* is a mathematical function used to represent the distribution of a variable. These curves are designed so that the area under the curve gives the *relative frequencies* for the distribution. This means that the area under the curve within a portion of the range is equal to the probability of getting an case in that portion of the range.
- Normal distributions are a specific type of distribution that have by bell-shaped, symmetric density curves. There are many naturally occuring variables that tend to follow normal distributions. Some examples are biological characteristics, IQ tests, and repeated measurements taken by physical devices.
- A normal distribution can be completely described by its mean and standard deviation. We therefore often use the notation $N(\mu, \sigma)$ to refer to a normal distribution with mean μ and standard deviation σ .

We use μ and σ instead of \bar{x} and s when we describe a normal distribution because we are talking about a *population* instead of a *sample*. For now you can just think about them in the same way. We will talk more about about the difference between populations and samples in Part 3.

- For all normal distributions, 68% of the data will fall witin one standard deviation of the mean, 95% will fall within two standard deviations of the mean, and 99.7% will fall within three standard deviations of the mean. This is called the *68-95-99.7 rule*.
- The *standard normal curve* is a normal distribution with mean 0 and standard deviation 1. Most statistics books will provide a table with the probabilities associated with this curve.
- People often use standardized cases when working with variables that have normal distributions. If your variable X has a mean μ and standard deviation σ then the standardized cases Z are equal to

$$Z = \frac{X - \mu}{\sigma}.$$
(5.1)

These standardized cases have a mean of 0 and a standard deviation of 1. This means that they will follow a standard normal curve.

- Sometimes you will be asked to calculate *normal relative frequencies*. In these problems you are asked to determine the probability of observing certain types of cases in a variable that has a normal distribution. To calculate normal relative frequencies you:
 - 1. Formulate the problem in terms of the original variable X. "I want to know the probability that X > 5."
 - 2. Restate the problem in terms of a standardized normal variable Z. "Because the mean is equal to 2 and the standard deviation is equal to 3, this corresponds to the probability that Z > 1."

- 3. Look up the probabilites associated with the cases you are interested in. Since standardized variables have a standard normal curve you can use the that table. You may also need to use the fact that the total area underneath the curve is equal to 1. "According to the table, the probability that Z < 1 is equal to .8413. This means that the probability that Z > 1 is equal to .1587."
- We can only use the standard normal curve to calculate probabilities when we are pretty sure that our variable has a normal distribution. Statisticians use *normal probability plots* to check normality. Variables with normal distributions have normal probability plots that look approximately like straight lines. It doesn't matter what angle the line has.

Part II

Understanding Relationships

Comparing Groups

- In the first part we looked at how to describe the distribution of a single variable. In this part we examine how the distribution of two variables relate to each other.
- *Boxplots* are used to provide a simple graph of a quantitative variable's distribution. To produce a boxplot you:
 - 1. Identify values of the variable on the y-axis.
 - 2. Draw a line at the median.
 - 3. Draw a box around the middle quartiles.
 - 4. Draw dots for any suspected *outliers*. Outliers are cases that appear to be outside the general distribution. A case is considered an outlier if it is more than 1.5 x IQR outside of the box.
 - 5. Draw lines extending from the box to the greatest and least cases that are not outliers.
- People will often present *side-by-side boxplots* to compare the distribution of a quantitative variable across several different groups.

Scatterplots

- *Response variables* measure an outcome of a study. We try to predict them using *explanatory variables*. Sometimes we want to show that changes in the explanatory variables cause changes in the response variables, but other times we just want to show a relationship between them.
- Response variables are sometimes called *dependent variables*, and explanatory variables are sometimes called *independent variables* or *predictor variables*.
- You can use *scatterplots* to graphically display the relationship between two quantitative variables measured on the same cases. One variable is placed on the horizontal (X) and one is placed on the vertical (Y) axis . Each case is represented by a point on the graph, placed according to its values on each of the variables. Cases with larger values of the X variable are represented by points further to the right, while those with larger values of the Y variable are represented by points that are closer to the top.
- If we think that one of the variables causes or explains the other we put the explanatory variable on the x-axis and the response variable on the y-axis.
- When you want to interpret a scatterplot think about:
 - 1. The overall *form* of the relationship. For example, the relationship could resemble a straight line or a curve.
 - 2. The *direction* of the relationship (positive or negative).
 - 3. The *strength* of the relationship.
- When the points on a scatterplot appear to fall along a straight line, we say that the two variables have a *linear* relationship. If the line rises from left to right we say that the two variables have a *positive* relationship, while if it decreases from left to right we say that they have a *negative* relationship.
- Sometimes people use a modified version of the scatterplot to plot the relationship between categorical variables. In these graphs, different points on the axis correspond to different categories, but their specific placement on the axis is more or less arbitrary.

Correlation

• The correlation coefficient is designed to measure the linear association between two variables without distinguishing between response and explanatory variables. The correlation between the variables x and y is defined as

$$r = \frac{1}{n-1} \sum \left[\left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \right].$$
(8.1)

In this equation, s_x refers to the standard deviation of x and s_y refers to the standard deviation of y.

- The correlation between two variables will be positive when they tend to be both high and low at the same time, will be negative when one tends to be high when the other is low, and will be near zero when the value on one variable is unrelated to the value on the second.
- The correlation formula can also be written as

$$r = \frac{\sum (x_i y_i) - \frac{1}{n} \sum x_i \sum y_i}{(n-1)s_x s_y}.$$
(8.2)

This is called the *calculation formula* for r. This formula is easier to use when calculating correlations by hand, but is not as useful when trying to understand the meaning of correlations.

- Correlations have several important characteristics.
 - \circ The value of r always falls between -1 and 1. Positive values indicate a positive association between the variables, while negative values indicate a negative association between the variables.
 - If r = 1 or r = -1 then all of the cases fall on a straight line. This means that the one variable is actually a perfect linear function of the other. In general, when the correlation is closer to either 1 or -1 then the relationship between the variables is closer to a straight line.
 - The value of r will not change if you change the unit of measurement of either x or y.
 - Correlations only measure the degree of *linear* association between two variables. If two variables have a zero correlation, they might still have a strong *nonlinear* relationship.

Least-Squares Regression

- A regression line is an equation that describes the relationship between a response variable y and an explanatory variable x. When talking about it you say that you "regress y on x."
- The general form of a regression line is

$$y = a + bx, \tag{9.1}$$

where y is the value of the response variable, x is the value of the explanatory variable, b is the *slope* of the line, and a is the *intercept* of the line. The slope is the amount that y changes when you increase x by 1, and the intercept is the value of y when x = 0.

- The values on the line are called the *predicted values*, represented by \hat{y} . The difference between the observed and the predicted value $(y_i \hat{y}_i)$ for each case is called the *residual*.
- Least-squares regression is a specific way of determining the slope and intercept of a regression line. The line produced by this method minimizes the sum of the squared residuals. We use the squared residuals because the residuals can be both positive or negative.
- The least-regression line of y on x calculated from n cases is

$$\hat{y} = a + bx,\tag{9.2}$$

where

$$b = \frac{\sum xy - \frac{1}{n} \sum x \sum y}{\sum x^2 - \frac{1}{n} (\sum x)^2}$$

and

$$a = \bar{y} - b\bar{x}.$$

- To see if it is appropriate to use a regression line to represent our data we often examine *residual plots*. A residual plot is a graph with the residual on one axis and either the value of explanatory variable the predicted value on the other, with one point from each case. In general you want your residual plot to look like a random scattering of points. Any sort of pattern would indicate that there might be something wrong with using a linear model.
- A *lurking variable* is a variable that has an important effect on your response variable but which is not examined as an explanatory variable. Often a plot of the residuals against time can be used to detect the presence of a lurking variable.
- In linear regression, an *outlier* is a case whose response lies far from the fitted regression line. An *influential observation* is a case that has a strong effect on the position of the regression line. Points that have unusually high or low values on the explanatory variable are often influential observations.

It is important to carefully examine both outliers and influential observations. Sometimes these points are not actually from the population you want to study, in which case you should drop them from your data set.

- It is important to notice that you get different least-squares regression equations if you regress y on x than if you regress x on y. This is because in the first case the equation is built to minimize errors in the vertical direction, while in the second case the equation is built to minimize errors in the horizontal direction.
- Like correlations, the slope of the regression line of y on x is a measure of the linear assocation between the two variables. These two are related by the following formula:

$$b = r \frac{s_y}{s_x}.\tag{9.3}$$

- The square of the correlation, r^2 , is the proportion of the variation in the variable y that can be explained by the least-squares regression of y on x. It is the same as the proportion of variation in the variable x that can be explained by the least-squares regression of x on y.
- You should notice that the equations for regression lines and correlations are based on means and standard deviations. This means that they will both be strongly affected by outliers or influential observations. Additionally, lurking variables can create false relationships or obscure true relationships.
- Regression lines are often used to predict one variable from the other. However, you must be careful not to extrapolate beyond the range of your data. You should only use the regression line to make predictions within the range of the data you used to make the line. Additionally, you should always make sure that new case you want to predict is drawn from the same underlying population.

Association vs. Causation

- Once you determine that you can use an explanatory variable to predict a response variable you can try to determine whether changes in the explanatory variable *cause* changes in the response variable.
- There are three situations that will produce a relationship between an explanatory and a response variable.
 - 1. Changes in the explanatory variable actually cause changes in the response variable. For example, studying more for a test will *cause* you to get a better grade.
 - 2. Changes in both the explanatory and response variables are both related to some third variable. For example, oversleeping class can cause you to both have uncombed hair and perform poorly on a test. Whether somone has combed hair or not could then be used to predict their performance on the test, but combing your hair will not cause you to perform better.
 - 3. Changes in the explanatory variable are *confounded* with a third variable that causes changes in the response. Two variables are confounded when they tend to occur together, but neither causes the other. For example, a recently published book stated that African-Americans perform worse on intelligence tests than Caucasians. However, we can't conclude that there is a racial difference in intellegence because a number of other important factors, such as income and educational quality, are confounded with race.
- The best way to establish that an explanatory variable causes changes in a response variable is to conduct an *experiment*. In an experiment, you actively change the level of the explanatory variable and observe the effect on the response variable while holding everything else about the situation constant. We will talk more about experiments in Chapter 11.

Part III

Generating Data

Sample Surveys

- The *population* is the group that you want to find out about from your data.
 - If your data represents only a portion of the population you have a *sample*.
 - If your data represents the entire population you have a *census*.
- Most data from surveys are recorded in terms of numbers. However, these numbers do not always mean the same thing. It is important to understand the *scale of measurement* used by a given question so you know what are the appropriate analyses. The four basic scales of measurement are:
 - 1. Nominal scales, where the numbers just serve as labels for categories. Example: coding 1=male, 2=female.
 - 2. Ordinal scales, where the numbers indicate the rank order of a category. Example: coding grade level in high school as 1=freshman, 2=sophomore, 3=junior, 4=senior.
 - 3. *Interval scales*, where equal differences between numbers reflect equal differences on some meaningful quantity. Example: day of the year.
 - 4. *Ratio scales*, which are just like interval scales but a "0" value indicates that something possesses none of the measured characteristic. Example: height.

For our purposes you should treat nominal and ordinal questions as categorical variables, and interval and ratio questions as quantitative variables.

- By selecting a *design* you choose a systematic way of producing data in your study. This includes things like what people or objects you will examine and the situations under which you will examine them. A poorly chosen design will make it difficult to draw any conclusions from the data.
- The basic goal of sampling design is to make sure that the sample you select for your study properly represents the population that you are interested in. Your sample is *biased* if it favors certain parts of the population over others.
- One example of a bad sampling design is a *voluntary response sample*. In this design you ask a question to a large number of people and collect data from those who choose to answer. This sample does not represent the general population because the people who are choose to respond to your question are probably different from the people who do not.
- A good way to get unbiased data is with a *simple random sample*. In a simple random sample every member of the population has an equal chance of being in the sample. Also, every combination of members is as likely as every other combination.
- You should also be careful about introducing bias into your data with the way you write your survey. The wording of each item should be clear and should not influence the way people answer the question.

Designed Experiments

- In observational studies you try to examine and measure things as they naturally occur. In *experimental studies* you impose some manipulation and measure its effects. It is easier to generalize from observational studies, but it is easier to determine causality from experimental studies.
- The objects on which an experiment is performed are called the *experimental units*. Human experimental units used to be called "subjects," but we now call them "research participants."
- The explanatory variables in an experiment are referred to as *factors*, while the values each factor can have are called its *levels*. A specific experimental condition applied to a set of units is called a *treatment*.
- The effectiveness of a treatment is generally measured by comparing the results of units that underwent the treatment to the results of units in a *control condition*. Units in a control condition go through the experiment but never receive the treatment.
- It is generally not a good idea to determine the effect of a treatment by comparing the units before the treatment to the same units after the treatment. Just being in an experiment can sometimes cause changes, even if the treatment itself doesn't really do anything. This is called the *placebo effect*.
- The basic idea of an experiment is that if you control for everything in the environment except for the factor that you manipulate, then any differences you observe between different groups in your experiment *must* be caused by the factor. It is therefore very important that the groups of experimental units you have in each level of your factor are basically the same. If there is some difference between the types of units in your groups, then the results might be caused by those differences instead of your treatment.
- Sometimes people use *matching* to make their groups similar. In this case you assign people to the different levels of your factor in such a way that for every person in one factor level you have a similar person in each of the other levels.
- The preferred way to make your groups similar is by *randomization*. In this case you just randomly pick people to be in each of your groups, relying on the fact that on average the differences between the groups will average out. Not only does this require less effort, but you don't have to know ahead of time what are the important things that you need to match on.
- There are many different ways to randomly assign your experimental units to conditions. You can use a table of random numbers, a mathematical calculator that produces random digits, or you can use statistical software.
- Randomization is one of the basic ideas behind statistics. When we randomly assign units to be in either the treatment or control conditions, there will be some differences between the groups. However, this is not a problem because mathematicians know a lot about how random variables work. When we

see a difference between the groups we can figure out the probability that the difference is just due to the way that we assigned the units to the groups. If the probability is too low for us to believe that the results were just due to chance we say that the difference is *statistically significant*, and we conclude that our treatment had an effect.

- In conclusion, when designing an experiment you want to:
 - 1. Control the effects of irrelevant variables.
 - 2. Randomize the assignment of experimental units to treatment conditions
 - 3. Replicate your experiment across a number of experimental units.

Part IV

Experience with Random Behavior

Randomness

- In Chapter 11 we talked about how and why staticians use randomness in experimental designs. Once we run an experiment we want to determine the probability that any relationships we see are just due to chance. In the next chapters we will learn more about probability and randomness so we can make sense of our results.
- A *random* phenomenon is one where we can't exactly predict what the end result will be, but where we can describe how often we would expect so see each possible result if we repeated the event a number of times. The different possible end results are referred to as *outcomes*.
- The *probability* of an outcome of a random phenomenon is the proportion of times we would expect to see the outcome if we observed the event a large number of times.
- We can determine the probabilities of outcomes by observing a random phenomenon a number of times, but most often we either assign them using common sense or else calculate them using mathematics.

Intuitive Probability

- An *event* is a collection of outcomes of a random phenomenon.
 - An event can be only a single outcome, or it can be many different outcomes.
 - In statistics we often use set notation to describe events. The event itself is a set, and the outcomes included in the event are elements of the set.
 - \circ When we talk about the probability of an event we are referring to the probability that the outcome of a random phenomenon is included in the event.
 - Two events are called *disjoint* if they have no outcomes in common. This means that they can never both occur at the same time.
 - Two events are called *independent* if knowing that one occurs does not change the probability that the other occurs. This is similar to saying that there isn't anything that is simultaneously influencing both events. Note that this is different from being disjoint in fact, two events cannot be both disjoint and independent.
- Probabilities of events follow a number of rules:
 - The probability that event A occurs = P(A).
 - The probability that an event A does not occur = 1 P(A).
 - Probabilities are always between 0 and 1.
 - If the event contains all possible outcomes then its probability = 1.
 - If events A and B are disjoint then P(A or B) = P(A) + P(B).
 - If events A and B are independent then P(A and B) = P(A)P(B).
- When thinking about a random phenomenon one of the first things we typically do is consider the different results we might observe.
 - A sample space is a list of the all of the possible outcomes from a specific random phenomenon.
 - \circ Just like events, we typically use a set to describe the sample space of a random phenomenon. This set is typically named S, and includes each possible outcome as an element.
- A sample space is called *discrete* if we can list out all of the possible outcomes. The probability of an event in a discrete sample space is the sum of the probabilities of the individual outcomes making up the event.
- A sample space is called *continuous* if we cannot list out all of the possible outcomes.
 - In a continuous sample space we use a *density curve* to represent the probabilities of outcomes. The area under any section of the density curve is equal to the probability of the result falling into that section.
 - To calculate the probability of an event you add up the area underneath the density curve for each range of outcomes included in the event.

Conditional Probability

- Many times the probability of an event will depend on external factors. Think about the probability that a high school student will go to college. We know that many factors, such as socio-economic status (SES) can impact this probability. Conditional probability lets us think about the probability of an event given that something else is true.
- We use the notation P(A|B) to refer to the conditional probability that event A occurs given that B is true. For example the probability that someone who is low in SES goes to college would be written as $P(\text{going to college} \mid \text{low SES})$.
- Conditional probabilities can also be thought of as limiting the type of people you want to include in your probability assessment. One good way to see conditional probabilities is in a 2-way table.
- Last chapter we mentioned that two events are *independent* if knowing that one occurs does not change the probability that the other occurs. Using conditional probability we can write this mathematically:

Two events A and B are independent if P(A|B) = P(A). (14.1)

• If we know the probabilities of two events and the probability of their intersection we can calculate either conditional probability:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}, \qquad P(B|A) = \frac{P(A \text{ and } B)}{P(A)}.$$
 (14.2)

• Similarly, if we know the probabilities of two events and the corresponding conditional probabilities we can calculate the probability of the intersection:

$$P(A \text{ and } B) = P(A)P(B|A) = P(B)P(A|B).$$
 (14.3)

This is actually an extension of the intersection formula we learned in chapter 13. While that formula would only work if A and B are independent, this one will always work.

• We can also extend the union formula we learned in chapter 13:

$$P(A \text{or} B) = P(A) + P(B) - P(A \text{and} B).$$
(14.4)

This formula will always be accurate, whether events A and B are disjoint or not.

• People often think that there are patterns in sequences that are actually random. There are many different types of sequences that can be reasonable results of independent processes.

Random Variables

- A random variable is a variable whose value is determined by the outcome of a random phenomenon. The statistical techniques we've learned so far deal with variables, not events, so we need to define a variable in order to analyze a random phenomenon.
- A *discrete random varaible* has a finite number of possible values, while a *continuous random variable* can take all values in a range of numbers.
- The *probability distribution* of a random variable tells us the possible values of the variable and how probabilities are assigned to those values.
 - The probability distribution of a discrete random variable is typically described by a list of the values and their probabilities. Each probability is a number between 0 and 1, and the sum of the probabilities must be equal to 1.
 - The probability distribution of a continuous random variable is typically described by a density curve. The curve is defined so that the probability of any event is equal to the area under the curve for the values that make up the event, and the total area under the curve is equal to 1. One example of a continuous probability distribution is the normal distribution.
- We use the term *parameter* to refer to a number that describes some characteristic of a population. We rarely know the true parameters of a population, and instead estimate them with *statistics*. Statistics are numbers that we can calculate purely from a sample. The value of a statistic is random, and will depend on the specific observations included in the sample.
- The *law of large numbers* states that if we draw a bunch of numbers from a population with mean μ , we can expect the mean of the numbers \bar{y} to be closer to μ as we increase the number we draw. This means that we can estimate the mean of a population by taking the average of a set of observations drawn from that population.

Sampling Distributions

- As discussed in the last chapter, statistics are numbers we can calculate from sample data, usually as estimates of population parameters. However, a large number of different samples can be taken from the same population, and each sample will produce slightly different statistics. To be able to understand how a sample statistic relates to an underlying population parameter we need to understand how the statistic can vary.
- The *sampling distribution* of a statistic is the distribution of values taken by the statistic in all possible samples of the same size.
- The distribution of the sample mean \bar{y} of *n* observations drawn from a population with mean μ and standard deviation σ has mean

$$\mu_{\bar{y}} = \mu \tag{16.1}$$

and standard deviation

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}.\tag{16.2}$$

If the original sample has a normal distribution, then the distribution of the sample mean has a normal distribution.

The distribution of the sample mean will also have a normal distribution (approximately) if the sample size n is fairly large (at least 30), even if the original population does not have a normal distribution. The larger the sample size, the more closer the distribution of the sample mean will be to a normal distribution. This is referred to as the *central limit theorem*.

• It is important not to confuse the distribution of the population and the distribution of the sample mean. This is sometimes easy to do because both distributions are centered on the same point, although the population distribution has a larger spread.

Part V

Statistical Inference

Estimating With Confidence

- We have already learned about using statistics to estimate population parameters. Now that we have an understanding of probability we can also state how much trust we have in our estimates.
- One important thing to keep in mind when thinking about statistical inference is that we always assume that our sample data are a random sample of the population whose parameter we want to estimate. If this is not true then our inferences we make might be invalid.
- Let's think about estimating the mean of a population μ . We already know that one way to calculate this estimate is to take the mean of a sample drawn from the population \bar{y} . We also know that \bar{y} is more likely to be closer to μ if we use a larger sample. We want a way to report how close we think our statistic \bar{y} is likely to be to the parameter μ .
- We know that \bar{y} follows a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. The 68-95-97 rule says that the probability that an observation from a normal distribution is within two standard deviations of the mean is .95. This means that there is a 95% chance that our sample mean \bar{y} is within $\frac{2\sigma}{\sqrt{n}}$ of the population mean μ . We can therefore say that we are 95% confident that the population parameter μ lies within $\frac{2\sigma}{\sqrt{n}}$ of our estimate \bar{y} .
- Reporting the probability that a population parameter lies within a certain distance of a sample statistic is called presenting a *confidence interval*. The basic steps to get a confidence interval are:
 - 1. Calculate your estimate. For a confidence interval around the mean you use \bar{y} .
 - 2. Determine your *confidence level* C. This is basically the probability that the parameter will fall within the interval you calculate.
 - 3. Determine the deviation of your interval. This is the amount above and below your estimate that will be contained in your interval. For a confidence interval around the mean this is equal to

$$z^* \frac{\sigma}{\sqrt{n}},\tag{17.1}$$

where z^* is how far from the center of the normal distribution you would have to go so that the area under the curve between $-z^*$ and z^* would be equal to C. This is the same thing as the number of standard deviations from the center you would need to include a proportion of the observations equal to C.

- 4. Calculate your confidence interval. Your interval will be between (estimate deviation) and (estimate + deviation).
- 5. Present your confidence interval. You should report both the actual interval and your confidence level.
- Notice that there is a tradeoff between your confidence level and the size of your interval. The higher the confidence level you want the larger the confidence interval will be.

Confidence Intervals for a Mean

- In the last section we introduced the idea of a confidence interval around the estimated mean. However, to calculate our interval we needed to know σ , the standard deviation of the population. Like most parameters, however, we generally don't know what this is. In this section we learn a way to calculate confidence intervals without having to know *any* population parameters.
- Remember that the validity of our confidence interval relies on the assumption that our sample mean has a normal distribution centered on the population mean. This is true as long as either the original population had a normal distribution or you have a sample size of at least 30.
- When we don't have access to the population standard deviation we can estimate it with the standard deviation of our sample, s.
- Estimating the standard deviation, however, adds some additional error into our measurements. We have to change our confidence intervals to take this into account. Specifically, our confidence intervals will be slightly wider when we estimate the standard deviation than when we know the parameter value.
- To calculate a confidence interval for the mean from a sample of size n when you do not know the population standard deviation you:
 - 1. Calculate the sample mean \bar{y} and standard deviation s.
 - 2. Determine your confidence level C.
 - 3. Determine the deviation of your interval. This is equal to

$$t^* \frac{s}{\sqrt{n}},\tag{18.1}$$

where t^* is how far from the center of the t distribution with n-1 degrees of freedom you would have to go so that the area under the curve between $-t^*$ and t^* would be equal to C.

- 4. Calculate your confidence interval. As before, your interval will be between (estimate deviation) and (estimate + deviation).
- 5. Present your confidence interval. You should report both the actual interval and your confidence level.
- Notice that this is basically the same as when you do know the population standard deviation except that you are drawing values from a *t* distribution instead of the normal distribution. Statisticians have proven that the values of the *t* distribution make the correction we need when we estimate the standard deviation from our sample. A *t* distribution looks very much like the normal distribution except that it is lower in the center and higher in the tails.
- There are actually a number of different t distributions, and the one you should use will depend on how many people you have in your sample. All t distributions have the same shape, but those with more degrees of freedom are look more like the normal distribution.

Testing Hypotheses

- When we calculate a confidence interval we are interested in estimating a population parameter. Sometimes, however, instead of estimating a parameter we want to test a specific theory about the population. For example, you might want to test whether the mean of the population is equal to a specific value. In this section we will talk about using sample statistics to test specific hypotheses about the population.
- The first thing you do is to write a mathematical statement of the theory you want to test. Since this is a theory about the population, your statement will be written in terms of population parameters.
- You write your statement in terms of a null hypothesis H_0 and an alternative hypothesis H_a . Usually H_0 is a statement that a parameter is equal to a constant (typically zero) while H_a states that the parameter is different from the constant. You typically write these so that your theory says that H_a is true.
- The test we will use attempts to find evidence for H_a over H_0 , so the two hypotheses must be disjoint. Tests of this sort (pitting a null hypothesis against an alternative hypothesis) are very common in statistics.
- Many times we are interested in determining if the sample mean is equal to zero. One set of hypotheses that would test this is

$$\begin{aligned}
 H_0: & \mu = 0 \\
 H_a: & \mu > 0.
 \end{aligned}$$
 (19.1)

Notice that the hypotheses can't both be true at the same time. The alternative hypothesis $\mu > 0$ is called *one-sided* because it only considers possibilities on one side of the null hypothesis. It is also possible to have *two-sided* alternative hypotheses, such as $\mu <> 0$.

- Once we have decided on our hypotheses we examine our sample data to determine how unlikely the results would be if H_0 is true, and if the observed findings are in the direction suggested by H_a .
- We quantify the result of our test by calculating a *p*-value. This is the probability, assuming that H_0 is true, that the observed outcome would be at least as extreme as that found in the data. The lower the p-value, the more confident we are in rejecting the null hypothesis.
- Exactly how you calculate the p-value for your test depends on what your null and alternative hypotheses are. For example, think about hypotheses about the value of the population mean. We know that the sample mean follows a normal distribution centered on the population mean. If we determine how many standard deviations there are between the sample mean and value specified under H_0 , we can look up the probability that a population with a mean equal to the H_0 value would produce a sample with a mean at least as extreme as that observed in the sample. This method is called a *one-sample z test*.
- The specific steps to perform a one-sample z test are:

- 1. Calculate the sample mean \bar{y} .
- 2. Calculate the one-sample z statistic

$$z = \frac{\bar{y} - \mu_0}{\frac{\sigma}{\sqrt{n}}},\tag{19.2}$$

where μ_0 is the mean assumed in the null hypothesis, σ is the known population standard deviation, and n is the sample size.

- 3. Look up the p-value associated with z in a normal distribution table. If you are performing a two-sided test you need to multiply this value by 2 (because you have to consider the probability of getting extreme results of the opposite sign).
- Once we calculate a p-value for a test we have to decide whether the probability is low enough to conclude that the null hypothesis is unlikely. Before analyzing the data you should choose a value that you would accept as indicating that the null is probably false. This value is called your *significance level*, and you say that your results are *statistically significant* if the p-value is less than your significance level.
- The exact significance level you should use will depend on the field that you are in. In psychology we typically use either .05 or .01.
- If the results are statistically significant we say that we reject H_0 in favor of H_a . Otherwise we say that we fail to reject H_0 .

Tests for a Mean

- In the last chapter we learned how to perform one-sample z tests to test hypotheses about the sample mean. However, this test assumes that we know the population standard deviation σ . In this chapter we will learn how to test hypotheses about the value of a mean when we do not know σ . We will also learn how to test for a difference between measurements taken on matched observations.
- Just as when calculating confidence intervals, when we don't have access to the population standard deviation we can estimate it with the standard deviation of our sample, s.
- As previously discussed, estimating the standard deviation will add some additional error into our measurements. To take into account we will use a t distribution instead of a normal distribution.
- To examine hypotheses about the mean when we don't know the population standard deviation we use a *one-sample t test*. To perform this test you
 - 1. Calculate the sample mean \bar{y} and standard deviation s.
 - 2. Calculate the one-sample t statistic

$$t = \frac{\bar{y} - \mu_0}{\frac{s}{\sqrt{n}}},\tag{20.1}$$

where μ_0 is the mean assumed in the null hypothesis and n is the sample size.

- 3. Look up the p-value associated with t in a t distribution table using n-1 degrees of freedom. If you are performing a two-sided test you need to multiply this value by 2.
- We will often find ourselves calculating t statistics. It has basically the same interpretation as a z statistic: in this case it is the number of standard devations between the sample mean \bar{y} and the value we are testing under the null hypothesis μ_0 . Whenever we calculate a t statistic it will have the form

$$t = \frac{\text{estimate} - \text{value assumed in } H_0}{\text{SE}_{\text{estimate}}},$$
(20.2)

where "SE" stands for "standard error."

• You can also use the one-sample t test to test for differences between two groups of observations if you have a *matched pairs design*. In a this design, for every observation in one group you have a matching observation in the other. They could either be different observations from the same participant, or they could be from different participants that are similar on important characteristics.

To test whether the two groups are the same you first create a new variable that is the difference between the matched participants. You will have a total of $\frac{n}{2}$ of these. You then perform a means test on this new variable just as explained above. When calculating your t statistic and degrees of freedom you substitute $\frac{n}{2}$ for the sample size.

• A statistical inference is called *robust* if the confidence level or p-value does not change very much when the assumptions of the procedure are violated.

• The one-sample t test is fairly robust to violations of normality, but it is not very robust to violations of the assumption that the observed data are a random sample of the population.

Part VI

Topics in Inference

Comparing Two Means

- At the end of the last chapter we learned how to test for a difference between two groups on a variable when we have a matched pairs design. In this chapter we will learn how to test for a difference between groups when the two groups have independent observations. This is one of the most-often used procedures in statistics.
- To compare the means of two populations on a single variable you can use a *two-sample t test*. To use this test
 - 1. You must take independent samples from both populations.
 - 2. The sample means from both populations must be normally distributed. This will automatically be satisfied if you take large samples.
 - 3. The size of your two samples can be different.
 - 4. Your experimental hypotheses should be of the form

$$H_0: \quad \mu_1 - \mu_2 = 0
 H_a: \quad \mu_1 - \mu_2 > 0.$$
(21.1)

The alternative hypothesis can be either one-sided (as in the example) or two sided (such as when H_a is $\mu_1 - \mu_2 \neq 0$).

- The steps to perform a two-sample t test are:
 - 1. Calculate the sample means \bar{y}_1 and \bar{y}_2 , and the sample standard deviations s_1 and s_2 . The statistics with the same subscript are calculated from the same sample, although it doesn't matter which subscript you give to each sample.
 - 2. Calculate the standard error of the difference:

$$SE_{\bar{y}_1-\bar{y}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$
(21.2)

where n_1 and n_2 are your sample sizes.

3. Calculate the two-sample t statistic

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\mathrm{SE}_{\bar{y}_1 - \bar{y}_2}}.$$
(21.3)

4. Unfortunately this statistic does not have a t distribution. The t distribution corrects for the error introduced by estimating one standard deviation, but to calculate this t we have to estimate two. We therefore typically rely on computers to calculate the p-value for this statistic.

5. If you do need to calculate a p-value for a two-sample t statistic by hand you can look it up in a t distribution table with k degrees of freedom, where k is equal to the smaller of $n_1 - 1$ and $n_2 - 1$. The actual p-value is guaranteed to be no larger than that produced by this method, so this is a *conservative* approximation.

Remember to multiply your obtained p-value by 2 if you are performing a two-sided test.

• The boundaries for a level C confidence interval for $\mu_1 - \mu_2$ are

$$(\bar{y}_1 - \bar{y}_2) \pm t^* (\mathrm{SE}_{\bar{y}_1 - \bar{y}_2}),$$
 (21.4)

where t^* is the critical value from a t distribution with k degrees of freedom such that the area between $-t^*$ and t^* is equal to C. k can either be obtained from the method above or (more accurately) from a statistics program.

- There is a special version of the two-sample t test that assumes that to two populations have the same standard deviation. The advantage of this is that the resulting test statistic only estimates a single standard deviation so p-values can be directly calculated from the t distribution. The procedure is called the *pooled* t test.
- The steps to perform a pooled t test are:
 - 1. Calculate the sample means \bar{y}_1 and \bar{y}_2 , and the sample standard deviations s_1 and s_2 .
 - 2. Calculate the pooled standard deviation

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}.$$
(21.5)

This is an estimate of the common standard deviation using a weighted average of the sample standard deviations.

3. Calculate the pooled t statistic

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$
(21.6)

4. Look up the p-value associated with t in a t distribution table using $(n_1 + n_2 - 2)$ degrees of freedom. If you are performing a two-sided test you need to multiply this value by 2.

Inference for Proportions

- Up until now we've only thought about problems where the response variable was continuous. Sometimes, however, we might want to consider a variable that only has the values "true" or "false" such as whether a woman becomes pregnant or not. In this case we can test hypotheses about the *proportion* of a specific outcome in the population.
- We can estimate the proportion of people who have a specific outcome in the population with the proportion of people who have that outcome in a random sample taken from the population. The population proportion is the parameter p, while the estimate taken from the sample is the statistic \hat{p} .
- Assuming that the population is much larger than our sample, the estimate \hat{p} has an approximately normal distribution with mean p and standard deviation

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}},\tag{22.1}$$

where n is the sample size.

• To test whether a population proportion is equal to a specific value you can perform a *one-sample z* test for a proportion. Your experimental hypotheses should be of the form

$$H_0: \quad p = p_0 \\
 H_a: \quad p < p_0.$$
(22.2)

The alternative hypothesis can be either one-sided (as in the example) or two sided (such as $H_a: p \neq p_0$).

- The steps to perform a one-sample z test for a proportion are:
 - 1. Calculate your estimate of the proportion \hat{p} from your sample.
 - 2. Calculate the one-sample z statistic for a proportion

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}},\tag{22.3}$$

where p_0 is the proportion assumed in your null hypothesis, and n is your sample size.

- 3. Look up the p-value associated with z in a normal distribution table. If you are performing a two-sided test you need to multiply this value by 2.
- The boundaries for a level C confidence interval for p_0 are

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$
(22.4)

where z^* is the critical value from a normal distribution that will include an area C between $-z^*$ and z^* .

• We can also test whether the proportions of two different populations are equal. For this we would use a *two-sample z test for proportions*. Your experimental hypotheses should be of the form

$$\begin{aligned} H_0: \quad p_1 - p_2 &= 0 \\ H_a: \quad p_1 - p_2 &> 0. \end{aligned}$$
 (22.5)

The null hypothesis could also be written as $H_0: p_1 = p_2$. The alternative hypothesis can be either one-sided (as in the example) or two sided (such as $H_a: p_1 - p_2 \neq 0$).

- The steps for this test are:
 - 1. Calculate your two proportion estimates \hat{p}_1 and \hat{p}_2 your samples.
 - 2. Calculate the *pooled sample proportion*. This is the overall proportion combining both of your samples:

$$\hat{p} = \frac{\text{count of successes in both samples combined}}{\text{total number of observations in both samples combined}}.$$
(22.6)

3. Calculate the two-sample z statistic for proportions

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$
(22.7)

where \hat{p} is the pooled sample proportion from above, and n_1 and n_2 are your sample sizes.

- 4. Look up the p-value associated with z in a normal distribution table. If you are performing a two-sided test you need to multiply this value by 2.
- To calculate a confidence interval around $p_1 p_2$ you
 - 1. Determine your confidence level C.
 - 2. Calculate the standard error of the difference in proportions. This is

$$SE_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$
(22.8)

3. The boundaries for a level C confidence interval for $p_1 - p_2$ are

$$(\hat{p}_1 - \hat{p}_2) \pm z^* (SE_{\hat{p}_1 - \hat{p}_2}),$$
 (22.9)

where z^* is the critical value from a normal distribution that will include an area C between $-z^*$ and z^* .

Two-Way Tables

- In the last chapter we learned how to perform a two-sample z test for proportions, which allows us to determine if the proportions of an outcome are the same in two different populations. In this chapter we extend this idea and learn how to test whether a proportion is the same across three or more groups.
- If you have a large number of categories you might think about just using the two-sample z test several times. For example, if you have three categories you could calculate three different z's: one comparing the first and second group, one comparing the first and third group, and one comparing the second and third group. If you discover any significant differences then you could claim that the means are not equal.

Unfortunately do this will give you invalid results. The problem is that the more groups you have the bigger the probability that any two of them would happen to be different just due to chance. We not only need a statistic that can test for differences between groups, we also need one that won't be more likely to find differences when we have more groups.

- A good way to look at your data when both your predictor and response variables are categorical is with a *two-way table*. A two-way table is organized so that you have a different predictor group in each column and a different outcome in each column. The inside of the table contains counts of how many people in each group had each outcome.
- We can use the information in a two-way table to calculate a chi-square statistic that will test whether the proportion of outcomes is the same for each of our groups. In this case our hypotheses are

$$\begin{aligned} H_0: \quad p_1 &= p_2 = p_3 \\ H_a: \quad \text{at least one pair is not equal.} \end{aligned}$$
 (23.1)

- The steps to perform a chi-square test are:
 - 1. Arrange your data in a two-way table.
 - 2. Find the counts that would be expected in each cell of the table if H_0 were true. These may be computed using the formula

expected count =
$$\frac{(\text{row total}) \times (\text{column total})}{n}$$
. (23.2)

3. Calculate the chi-square statistic

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}.$$
 (23.3)

- 4. Look up the p-value associated with χ^2 in a χ^2 distribution table using (r-1)(c-1) degrees of freedom, where r is the number of rows and c is the number of columns in your two-way table.
- You should not use the χ^2 test if more than 20% of your *expected* cell counts are less than 5. As long as the expected counts are high it is ok if the observed cell counts are less than 5.

Inference for Regression

- In earlier chapters we learned how we can use correlation and least-squares regression to describe the strength of a linear relationship between two quantitative variables. In this chapter we will learn how to test hypotheses about linear relationships, and we will learn how to use a regression line to predict new values.
- You should always check a scatterplot of your data before trying to make inferences about the regression line. You should make sure
 - 1. That there are no outliers in your data. As we've seen before outliers can have a strong influence on the least-squares regression line.
 - 2. That the relationship between the two variables is approximately linear. Your inferences will be invalid if the relationship between the variables is seriously curved.
- Remember that a least-squares regression line has the form

$$\hat{y} = a + bx,\tag{24.1}$$

and that the values of a and b are given by the following equations:

$$b = \frac{\sum (x_i y_i) - \frac{1}{n} \sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n} (\sum x_i)^2}$$
(24.2)

$$a = \bar{y} - b\bar{x}.$$

Notice that a and b are statistics since they are computed strictly from sample data.

• Imagine that you can measure x and y for every person in the population. You could then calculate a population regression line which would have the form

$$\hat{y}_i = \alpha + \beta x_i. \tag{24.3}$$

- In order for us to draw inferences from our sample regression line (such as testing hypotheses or building confidence intervals) we have to make several assumptions. If we violate these assumptions then our inferences might not be correct. We assume
 - 1. If we examined a large number of responses at a single value of x the values we observe in y would have a normal distribution.
 - 2. All the observations are independent of each other except for the influence of x.
 - 3. The standard deviation of y (which we will call σ) around the regression line is the same for all values of x.

- If we took a number of samples and calculated the least-squares regression line for each you would find that we would get different values for a and b each time. In fact, you would find that these would have a normal distribution. Staticians have proven that the mean of a's distribution is α and the mean of b's distribution is β . This means that we can use a to estimate α and b to estimate β .
- To test hypotheses about the parameters α and β we need estimates of these distributions. The first step to this is to calculate an estimate of the standard deviation of y around the regression line:

$$s = \sqrt{\frac{1}{n-2}\sum(y_i - \hat{y}_i)^2}.$$
(24.4)

This is something like an average deviation from the regression line. We divide by n-2 instead of n because we had to estimate two parameters (a and b, which are used to calculate \hat{y}_i) to make this calculation. Notice that this *not* the same as the standard deviation of y. For this standard deviation we don't consider the variability in y that is explained by the regression line.

• Once we have this we can estimate the standard deviation of these distributions with the following equations:

$$SE_a = s \sqrt{\frac{1}{n} + \frac{\bar{x}_i^2}{\sum (x_i - \bar{x})^2}}$$
 (24.5)

$$SE_b = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}.$$
 (24.6)

- Now that we have the estimates a and b, and the standard errors of these estimates SE_a and SE_b we can test theories about the parameters α and β .
- Most often we want to test theories about the relationship between x and y. We can do this by proposing hypotheses about the parameter β , which is the slope of the regression line. Our hypotheses will be of the form

$$\begin{array}{ll}
H_0: & \beta = b_0 \\
H_a: & \beta \neq b_0.
\end{array}$$
(24.7)

We can also examine one-sided hypotheses, such as $H_a: \beta > b_0$.

• To test these hypotheses we calculate a one-sample t statistic

$$t = \frac{b - b_0}{\mathrm{SE}_b},\tag{24.8}$$

and determine the p-value from a t distribution with n-2 degrees of freedom. We must remember to multiply the p-value by 2 if we perform a two-sided test.

- We are most often interested in testing whether or not there is a significant relationship between x and y. In this case $b_0 = 0$.
- The boundaries for a level C confidence interval for β are

$$b \pm t^*(SE_b), \tag{24.9}$$

where t^* is the critical value from a t distribution with n-2 degrees of freedom such that the area between $-t^*$ and t^* is equal to C.

- We can test hypotheses and build confidence intervals around α in the same way, but this is not typically done. All that α is is the mean value of y when x = 0, which usually is not very important.
- We've just learned how to use a regression model to test the strength of the linear relationship between two variables. Sometimes, however, we are more interested in using the regression line to predict new observations.

- As you might suspect, if we want to predict the value of y at a specific value of x (which we will refer to as x*), we just put that value of x into the regression equation. We talked about this back in Lesson 9. What we will learn in this section is how to build a confidence interval around our predicted value.
- There are actually two different confidence intervals we might want. First, we might want to build a *confidence interval around the mean* of \hat{y} at x^* . For example, we might want to make an interval such that we are 95% confident that it contains the mean value of \hat{y} at x^* . The variability in our interval is caused by our imprecision in estimating the population mean.

The limits for a level C confidence interval around the mean are

$$\hat{y} \pm t^* (\mathrm{SE}_{\hat{\mu}}), \tag{24.10}$$

where t^* is the critical value from a t distribution with n-2 degrees of freedom such that the area between $-t^*$ and t^* is equal to C, and $SE_{\hat{\mu}}$ is the standard error for predicting the mean response, calculated from the formula

$$SE_{\hat{\mu}} = s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}.$$
(24.11)

• We also might want to build a *prediction interval for a single observation* of \hat{y} at x^* . For example, we might want to make an interval that will contain 95

The limits for a level C prediction interval for a single observation are

$$\hat{y} \pm t^*(\mathrm{SE}_{\hat{y}}),\tag{24.12}$$

where t^* is the critical value from a t distribution with n-2 degrees of freedom such that the area between $-t^*$ and t^* is equal to C, and $SE_{\hat{y}}$ is the standard error for predicting the mean response, calculated from the formula

$$SE_{\hat{y}} = s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}.$$
(24.13)

We can examine the distribution of the residuals from fitting a regression line to determine how well the problem satisfies the assumptions required for valid regression inference.

One-Way Analysis of Variance

- In Lesson 21 we learned how to compare the means of two populations. In this chapter we will learn how to perform an Analysis of Variance (usually just referred to as ANOVA) which will allow us to simultaneously compare the means of three or more groups.
- We already know how to use a two-sample t test to compare the means of two groups. To compare more than two groups we might just calculate several t statistics, one comparing each of our groups. However, this procedure would be invalid. Think back to when we talked about using several z tests for proportions to compare the likelihood of an outcome in three or more groups. We would have the same problem: the more groups you have the bigger the probability that any two of them would happen to be different just due to chance.
- Just like we use the chi-square statistic to simultaneously compare the proportions of an outcome in three or more populations, we can use *Analysis of Variance* (ANOVA) to simultaneously compare the means of three or more populations.
- What the ANOVA does is compare the between-group variance (which is how different your groups are from each other) to the within-group variance (which is how different members of the same group are from each other). If the between-group variance is much larger than the within-group variance then you conclude that the means are significantly different.
- In order to draw inferences from an ANOVA we have to make several assumptions. Violating these assumptions makes our inferences less valid. These assumptions are:
 - 1. We have independent random samples from each population. The number we take from each population can be different.
 - 2. Each individual population has a normal distribution.
 - 3. All of the individual distributions have the same standard deviation. This is similar to the assumption we make when performing linear regression. The distributions can (and we expect them to) have different means.
- An ANOVA will test the hypotheses

$$\begin{aligned} H_0: \quad \mu_1 &= \mu_2 = \ldots = \mu_I \\ H_a: \quad \text{at least two means are different,} \end{aligned}$$
 (25.1)

where I is the number of different groups we want to compare.

• To measure the within- and between-group variability we calculate *sums of squares*. The between-groups sum of squares is calculated as

$$SSG = \sum_{\text{groups}} n_i (\bar{x}_i - \bar{x})^2, \qquad (25.2)$$

while the within-groups sum of squares is calculated as

$$SSE = \sum_{\text{groups}} (n_i - 1) {s_i}^2, \qquad (25.3)$$

where n_i is the number of observations in group i, \bar{x}_i is the mean of group i, and s_i is the standard deviation of group i.

The sum of SSG and SSE gives us the *total sum of squares* SST. This is a measure of the total variability found in x.

- Each one of our sums of squares has degrees of freedom associated with it. This is related to the number of different sources that contribute to the sum. We typically divide the sums of squares by the degrees of freedom to get a measure of the average variation. Working with these *mean squares* lets us compare many groups simultaneously without inflating the probability of finding significant results.
- The relationships between these variables is summarized in the ANOVA table below:

Source of variation	SS	df	MS
Between Group Within Group		I-1 N-I	$\frac{\text{SSG}}{I-1}$ $\frac{\text{SSE}}{N-I}$
Total	SST	N-1	$\frac{\text{SST}}{N-1}$

In the table, I is the number of groups and N is the total number of subjects. Notice that the total df is equal to the sum of the between-group and within-group df.

- We can perform an Ftest to determine if we have a significant amount of between group variance. This is the same thing as testing whether any of our means are significantly different. To perform an F test you
 - 1. Collect the information needed to build an ANOVA table.
 - 2. Calculate the F statistic

$$F = \frac{\text{MSG}}{\text{MSE}}.$$
(25.4)

3. Look up the corresponding p-value in an F distribution table with I-1 numerator degrees of freedom and N-I denominator degrees of freedom.