

Day By Day Notes for PBIS 189

Fall 2011

Day 1

Activity: Go over syllabus. Take roll. Overview example: Kristin Gilbert trial. Discussion of variables and graphs.

Goals: Review course objectives: collect data, summarize information, and make inferences.

I have divided this course into four “units”. Unit 1 (Days 1 through 7) is about summarizing one-dimensional data, that is, a single list of data. Unit 2 (Days 8 through 13) is about summarizing two-dimensional data, that is, data that come in two lists. Unit 3 (Days 14 through 21) is about sampling. Unit 4 (Days 22 through 28) is about statistical inference. Throughout the course, we will also focus on the **abuses** people often make of statistics and statistical methods. Sometimes these abuses are unintentional. My hope is that at the end of the course, you will have an appreciation for the “tricks” that people use to lie to you using numbers.

I use the Gilbert trial example because it demonstrates all three of the course’s main ideas in action. The charts we see are examples of how to organize and summarize information that may be complicated by several variables or dimensions (Units 1 and 2). The argument about whether we would see results such as this one if there really were no relationship between the variables is an example of the probability we will study in Unit 3. And the trial strategy itself is what statistical inference is all about (Unit 4). Many of you will encounter inference when you read professional journals in your field and the researchers use statistics to support their conclusions of improved treatments or to estimate a particular proportion or average.

Most if not all of you are taking this course as a general education requirement or a requirement for your major. There is a reason you have been asked to take such a course: often in our modern world we encounter numerical or logical arguments. Without proper skills to deal with these arguments, you will often be fooled by “data pushers”, those with an agenda and who use slick methods to get you to believe their position. You **should** be a skeptic when people use numbers in their persuasive arguments. Too often, and sometimes unknowingly, such arguments are misleading. The University believes that a successful college graduate has the ability to reason quantitatively. Our hope is that your efforts in this course will prepare you appropriately.

I believe to be successful in this course, you must actually **read** the text (and these notes) carefully, and work problems. The most important thing is to **engage** yourself in the material. However, our class activities will sometimes be unrelated to the homework you practice and/or turn in for the homework portion of your grade; instead they will be for **understanding** of the underlying principles. For example, on Day 14 we will simulate Simple Random Sampling by taking 80 samples from each group in the class. This is something you would never do in practice, but which I think will demonstrate several lessons for us. In these notes, I will try to point out to you when we’re doing something to gain **understanding**, and when we’re doing something to gain **skills**.

Each semester, I encourage students to see me for help outside of class. However, I suspect some of you are embarrassed to seek help, or you may feel I will think less of you for not “getting it” on your own. Personally, I think that if you are struggling and cannot make sense of what we are doing, and **don’t** seek help, you are cheating yourself out of your own education. I am here to help you learn statistics. Please ask questions when you have them; I don’t believe there are “stupid questions”. Often other students have the same questions but are also too shy to ask them in class. If you are still too shy to ask questions in class, come to my office hours or make an appointment. Incidentally, when I first took statistics, I didn’t understand it all on my own either, and I too didn’t go to the instructor for help. I also didn’t get as high a grade as I could have!

I believe you get out of something what you put into it. Very rarely will someone fail a class by attending every day, doing all the assignments, and working many practice problems; typically people fail by not applying themselves enough - either through missing classes, or by not allocating enough time for the material. Obviously I cannot tell you how much time to spend each week on this class; you must all find the right balance for you and your life’s priorities. One last piece of advice: don’t procrastinate. I believe statistics is learned best by daily exposure. Cramming for exams may get you a passing grade, but you are only cheating yourself out of understanding and learning.

Units 1 and 2 are about summarization. We will work on several types of summary: graphical summaries of a list of numbers, graphical summaries of a list of characteristics, numerical summaries of a list of numbers, and the summaries for two variables. Sometimes we will summarize situations with one or two numbers, called **summary statistics**. Other times, we need more than that, such as a picture, or many summary statistics. One of your tasks in this course is to gain enough experience to know what is enough summarization for a particular situation.

The first summaries we will work on are the “one-variable graphical techniques”. If you are working with a list of characteristics (attribute data or categorical data), you will need to use a bar graph, or a frequency table, or a pie graph. If you are working with a list of numbers, then you have many more options, including frequency tables, histograms, quantile plots, stem plots, and box plots, just to name a few. Your goal in Units 1 and 2 is to become adept at producing these displays, either by hand or with the calculator, and to interpret the main features from either your own creations or someone else’s work.

In these notes, I will put the daily **task** in gray background.

Your task today is, from a list of numbers, to communicate the important information to the person next to you. (Work in pairs or groups.) Specifically, make and interpret a frequency table and a histogram. In your description to your neighbor, keep in mind these terms: symmetry, skew, center, spread, mode, outlier. (These terms are in the text.) On your calculator, make sure that you try different window settings for your histogram. Remember, you are trying to describe your data to your neighbor without having to tell them every single value; you are trying to point out the main features that you have identified. (For example, deciding whether some list is “skewed” or “symmetric” will often be a matter of opinion; you need to look at many lists before you can make good judgment calls.)

In these notes, I will put sections of computer commands in boxes, like this one. In these notes, I refer to the calculator as the TI-83. The same commands apply to the TI-84.

Frequently we will use the TI-83 to make our work easier. At first, you may think of the machines as burdens, confusing and intimidating. I believe that if you stick with it and become experienced with the tools, you will discover that the calculators are indispensable for calculating statistics. I will show you what I can about the TI-83's, but it is up to you to practice using them. I cannot do that part for you, of course. As always, **ASK** questions as you have them!

Generally, we have three chores to perform before our machine will show us the graphical display we want. We must: 1) Enter the data into the calculator. 2) Choose the right options for the display we want. And 3) Set up the proper window settings. The commands to do these activities on the calculator are:

- 1) **STAT EDIT** Use one of the lists to enter data, **L1** for example; the other **L**'s can be used too. The **L**'s are convenient work lists. At times, you may find that you want more meaningful names. One way to do this is to **store** the list in a new **named** list after entering numbers. The syntax for this is **L1 → NEWL**, assuming the data was entered in **L1** and you want the new name to be **NEWL**. (The **→** key is in the lower left, directly above the **ON** key.) Note: list names are limited to five letters.
- 2) **2nd STATPLOT 1 On** Use this screen to designate the plot settings. You have a choice of six displays: scatter plot, line graph (we won't use this), histogram, modified box plot, box plot, and normal plot. You can have up to three plots on the screen at once. For histograms, we will only use one at a time. Later, when we see box plots and scatter plots, we will make multiple displays.
- 3) **ZOOM 9** This command centers the window "on" your data. It is always a good idea to see what the **WINDOW** settings are. If you then change any of the **WINDOW** settings, you will then press **GRAPH** to see the changes. (If you use **ZOOM 9** again, the changes you just made don't get used!)

Goals: (In these notes, I will summarize each day's activity with a statement of goals for the day.)

Begin graphical summaries (describing data with pictures). Be able to use the calculator to make a histogram.

Skills: (In these notes, each day I will identify **skills** I believe you should have after working the day's activity, reading the appropriate sections of the text, and practicing exercises in the text.

- **Identify types of variables.** To choose the proper graphical displays, it is important to be able to differentiate between Categorical (or Qualitative) and Quantitative (or Numerical) variables. Of course, it is also necessary to know exactly what a variable is. Quite simply, a **variable** is an attribute about a collection of subjects. As an example, for the students in this class, we could measure your height (a quantitative/numerical variable) or your major (a qualitative/categorical variable). The important thing is that each subject in the group has a value for the variable.

- **Be familiar with types of graphs.** To graph categorical variables we use bar graphs or pie graphs. To graph numerical variables, we use histograms, stem plots, or **QUANTILE** plots. In practice, most of our variables will be numerical but it is still important to know the categorical displays.
- **Summarize data into a frequency table and produce a histogram.** A frequency table is a list of class intervals (sometimes called bins or classes) and the number of data values in the class intervals. You can make a frequency table yourself by just counting how many data values are in each interval, but the easiest way is to first use the TI-83 to make a histogram and then to **TRACE** over the boxes and record the classes and counts. In fact, the histogram is just a picture of the frequency table.
- **Know how to modify the TI-83 default histogram.** You can control the size and number of the classes with **Xscl** and **Xmin** in the **WINDOW** menu. The decision of how many classes to create is arbitrary; there isn't a "right" answer, or rather **all** choices of **Xscl** and **Xmin** are "right" answers. You should experiment every time you use the TI-83 to make a histogram by changing the interval width **Xscl** and starting point **Xmin** to see what happens to the display. I will try to do this in class every time I draw a histogram, so that you get used to it.
- **Know how to create and interpret graphs for categorical variables.** The two main graphs for categorical variables are pie graphs and bar charts. Pie graphs are difficult to make by hand, but are popular on computer programs like Excel. Bar charts are also common on spreadsheets. Data represented by pie graphs and bar charts usually are expressed as percents of the whole; thus they add to 100 %. The ordering for categories is arbitrary; therefore concepts such as skew and center make no sense.

Reading: To The Student, pages xxiii-xxix. Chapter 1. (Skip Time Plots.)

Day 2

Activity: Creating and interpreting stem plots, histograms, and quantile plots.

Use the [Arizona Temps](#) data set to practice creating the histograms, stem plots, and quantile plots for several lists. Compare and interpret the graphs. Identify shape, center, and spread. Today's new methods, the stem plot and the quantile plot, are alternate ways to summarize a list of numbers.

Answer these questions about the Arizona Temps:

- 1) Do any of the lists have outliers? (There is not a firm definition of exactly what an outlier is. You will have to develop some personal judgment; the best way is to look at many lists and displays.)
- 2) What information does the stem plot show that the histogram hides?
- 3) What information does the quantile plot show that the stem plot and histogram hide?

4) Which display is most “trustworthy”? That is, which one has the smallest likelihood of misleading you?

The **stem plot** is a hand technique, most useful for small (under 40 values) data sets. It is basically a quick way to make a frequency chart, but always using class intervals based on the base 10 system. This means the intervals will always be ten “units” wide, such as 10 to 19, or 1 to 9, or .01 to .09. The “unit” chosen is called the **stem**, and the next digit after the stem is called the **leaf**.

QUANTILE is a program I wrote that plots the sorted data in a list and “stacks” the values up. This is known as a **quantile plot**. Basically we are graphing the individual data value versus the rank, or percentile, in the data set. Quantile plots always go up from left to right. The command is **PRGM EXEC QUANTILE ENTER**. The program will ask you for the list where you’ve stored the data. **A** and **B** are temporary lists used by the program, so if you have data in these lists already, store them in another list before executing.

Goals: Be able to use make and interpret a quantile plot, using the TI-83 program **QUANTILE**. Be able to make and interpret a stem plot by hand.

Skills:

- **Use the TI-83 to create an appropriate histogram or quantile plot.** **STAT PLOT** and **QUANTILE** are our two main tools for viewing distributions of data on the TI-83. Histograms are common displays, but have flaws; the choice of class intervals is troubling as it is not unique. The quantile plot is more reliable, but less common. For interpretation purposes, remember that in a histogram tall boxes represent places with lots of data, while in a quantile plot those same high-density data places are represented by steepness.
- **Create a stem plot by hand.** The stem plot is a convenient manual display; it is most useful for small data sets, but not all data sets make good stem plots. Choosing the “stem” and “leaf” to make reasonable displays will require some practice. Some ideas for a proper choice of stems: if you have many empty rows, you have too many stems. Move one column to the left and try again. If you have too few rows (all the data is on just one or two stems) you have too few stems. Move to the right one digit and try again. Some data sets will not give good pictures for any choice of stem, while some benefit from splitting or rounding (see the example on page 21).
- **Describe shape, center, and spread.** From each of your graphs, you should be able to make general statements about the shape, center, and spread of the distribution of the variable being explored. One of the main conclusions we want to make about lists of data when we are doing inference (Chapters 14 to 22) is whether the data is close to symmetric; many times “close enough” is, well, close enough! We will discuss this in more detail when we see the Central Limit Theorem in Chapter 11.

Reading: Chapter 2.



Day 3

Activity: *Dance Fever* example. Using the TI-83 to calculate summary statistics and to make the box plots display. **Quiz 1 today.**

Our second topic is numerical measures of data. I hope that you are already familiar with some of these measures, such as the mean and median. Others, like the standard deviation, will perhaps be mysterious. Of course our goal is to unmask the mystery!

When a graphical display shows that a data set is well behaved, perhaps symmetrical with no outliers, we may want to summarize further. We often want two measurements about numerical data: the center and the spread. The reason for these two kinds of measurements will be clearer after we see the Central Limit Theorem in Unit 3. Briefly, “center” is the typical data value, and “spread” refers to how variable the data values are. We will first practice using technology to produce the calculations for us. Along the way, I hope we gain some understanding too.

Use the [Arizona Temps](#) data set to calculate the mean, the standard deviation, the 5-number summary, and the associated box plot for any of the variables.

Compare the box plots and numerical measures with the corresponding histograms and quantile plots you made on Day 2. Note the similarities (where the data values are dense, and where they are sparse) but especially note the differences. The box plots and numerical measures cannot describe shape very well. On the other hand, histograms are messy to use to compare two lists. The stem and leaf is tedious to modify.

Answer these questions about the Arizona Temps:

- 1) Are high and low temperatures distributed the same way, other than the obvious fact that highs are higher than lows? (When we talk of a **distribution** we mean a description of how the data values are centered and how they are spread out. Sometimes this will be a simple statement; other times we need to see a graph.)
- 2) How does a single case influence the calculator’s answers? (What if there was an outlier in the list? If you didn’t have an outlier, change one of the values to a ridiculously large value and recalculate the graphs and measures. Notice the effects of these extreme values.)
- 3) What information does the box plot disguise?

To calculate our summary statistics, we will use **1-Var Stats** (to use List 1) or **1-Var Stats L2** for List 2, for example. There are two screens of output; we will be mostly concerned with the mean \bar{x} (pronounced “x bar”), the standard deviation **Sx**, and the five-number summary from screen two.

To make the **box plot**, we use **STAT PLOT Type** and pick the 4th or 5th icons. The fifth icon is the true box plot, but the fourth one (the **modified box plot**) has a routine built in to flag possible outliers. I recommend using the modified box plot as it shows at least as much

information as the regular box plot, but includes the potential outliers, as defined by this procedure.

Goals: Compare numerical measures of center. Summarize data with numerical measures and box plots. Compare these new measures with the histograms, stem plots, and quantile plots you made on Day 2.

Skills:

- **Understand the effect of outliers on the mean and median.** The mean (or average) is unduly influenced by outlying (unusual) observations, whereas the median is not. Therefore, knowing when your distribution is skewed is helpful. The mean is attracted to the outliers, so for lists with a high outlier, the mean will be higher than the median, and higher than if the outlier is removed. The degree of influence is a matter of judgment. The median is almost completely unaffected by outliers. For technical reasons, though, the median is not as common in scientific applications as the mean.
- **Use the TI-83 to calculate summary statistics.** Calculating summary statistics using the TI-83 is as simple as entering numbers into **STAT EDIT** and then choosing **STAT CALC 1-Var Stats L#**, where **L#** represents the list you are interested in. If you are doing some work by hand, you may have to organize information in a correct way, such as listing the numbers from low to high to find the median. **IMPORTANT NOTE:** I recommend that you get used to using the statistical features of your calculator to produce the mean. While I realize many of you are in the habit of calculating the mean by adding up all the numbers and dividing by the sample size, you will not be using the full features of your machine. Later on you will find yourself entering data twice: once to calculate the mean, then again when you discover you needed the standard deviation and the data should be in a list.
- **Compare several lists of numbers using stem plots and box plots.** For two short lists, the best simple approach is the back-to-back stem plot. For longer lists, or for more than two lists, use box plots, side-by-side, or stacked. At a glance, you can then assess which lists have typically larger values or more spread out values, etc. To graph up to three box plots on the TI-83, enter a different list in each of the 3 plots you can display using **STATPLOT**.
- **Understand box plots.** You should know that the box plots for some lists don't tell the interesting part of those lists. For example, box plots do **not** describe the shape of a distribution very well; you can only see where the quartiles are. Alternatively, you should know that the box plot gives a very good first impression.

Reading: Chapters 2 and 3.

Day 4

Activity: Interpreting standard deviation. Introduce the TI-83's normal calculations. **Homework 1 due.**

Today we will practice the ideas of mean and standard deviation by letting you play with some lists of your own. There are many possible answers to some of the following questions; you should try to see if you **understand** why each of them may have many answers. Also, the fourth list is an attempt to get you thinking about how spread out a data set is. This is a

requirement for understanding the standard deviation. As always, ask questions as you have them; we will be mostly working in groups today and I will listen to you as you work.

Create the following lists:

- 1) A list of 10 numbers that has only one number below the mean. (This seems to violate our interpretation of the average as being “representative”. You should understand how the mean might mislead us!)
- 2) A list of 10 numbers that has the standard deviation greater than the mean. (Note that the standard deviation seems to be more affected by outlying values than the mean is.)
- 3) A list of 10 numbers that has a standard deviation of zero.
- 4) For your fourth list, start with **any** 21 numbers. Find a number N such that 14 of the numbers in your list are within N units of the average. For example, pick an arbitrary number N (say 4), calculate the average plus 4, the average minus 4, and count how many numbers in your list are between those two values. If the count is less than 14, try a larger number for N (bigger than 4). If the count is more than 14, try a smaller number for N (smaller than 4). Continue guessing and counting until you get it close to 14 out of 21. This number you settle on **should** be fairly close to the standard deviation.

Finally, compare the standard deviation to the **Inter Quartile Range** ($IQR = Q3 - Q1$).

Another way to understand standard deviation is to associate the graphical plots with the value of s . Then change a few points, and reexamine the graphs and value of s . Experience like this is not gained by reading about it; you **must** do it yourself! I hope that you use class time in this way.

So far, we have used graphs to summarize distributions and summary statistics to describe certain features such as center and spread. Today we will also look at a specific distribution that is so common it has a name, the **normal distribution**. We will see that data following this curve are extremely easy to describe: we only need to know the mean and the standard deviation. The difference between this new material and yesterday’s work is that if the data really does follow the normal curve, we can draw the distribution’s curve by knowing just those two measures. In yesterday’s work, although we knew the mean and the standard deviation, that still gave us no indication of the **shape** of the distribution.

Generally, finding areas under curves is a calculus problem; ideally you end up with a formula where you simply plug in the endpoints desired. Unfortunately, the normal curve is an example of a function for which the calculus approach doesn’t give a useful answer. Therefore to find areas under the normal curve, we **must** use some sort of approximation. The normal table in the book is one approximation. I don’t recommend using that table though, because our calculator is easier, **and** it happens to be more precise! However, this means that as you read the problems in the textbook, you must skip over the descriptions given about using z -scores to find areas. Also, remember this discrepancy when checking answers in the back of the book for the odd problems: often the book’s answers are different from the calculator’s answers because they used the (inaccurate) table instead of the (accurate) calculator.

My advice to you when calculating areas using the normal curve is to always draw a picture. See your class notes for the actual diagram and the way to label it. I will try to always put two scales on my normal curves: the z -scale and the scale of real units. I do this so that I can make sense of answers I get. For example, if the shaded area on my diagram is obviously less than half of the total area, and if my answer from the calculator is 0.75, then that makes no sense, and I must have entered something into the calculator incorrectly (or drawn the picture wrong!).

Find the following areas:

- 1) The area between -1 and +1 on the standard normal curve.
- 2) The area between -2 and +2 on the standard normal curve.
- 3) The value on the standard normal curve that gives the lower 80 % (the 80th **percentile**).
- 4) The area on the normal curve with a mean of 100 and a standard deviation of 15 above 150.
- 5) The values from a normal curve with a mean of 100 and a standard deviation of 15 that contain the central 95 % of the area. (We will see this idea again when we do confidence intervals in Chapter 13.)

DISTR normalcdf(lower, upper) calculates the area under a normal curve between **lower** and **upper**. If you specify just 2 values, then the mean is 0 and the standard deviation is 1 (this is called the **standard normal curve**). If you want a different mean or standard deviation, add a third and fourth parameter, as in the next paragraph.

DISTR normalcdf(lower, upper, mean, standard deviation) calculates the area under the normal curve between **lower** and **upper** that has the given **mean** and **standard deviation**.

Examples: **DISTR normalcdf(-10, 20, 5, 10)** finds the area between -10 and +20 on a normal curve with mean 5 and standard deviation 10 while **DISTR normalcdf(-2, 2)** finds the area on the standard normal curve between -2 and +2.

DISTR invNorm(value, mean, standard deviation) works backwards, but only gives **upper** as an answer. It is also referred to as a percentile. The 90th **percentile** is that point at which 90 % of the observations are below. The syntax is **DISTR invNorm(.90)** or **DISTR invNorm(.90, 5, 10)**; the first example assumes the standard normal curve and reports the 90th percentile. The second example gives the 90th percentile for a normal curve with a mean of 5 and a standard deviation of 10.

IMPORTANT NOTE: If the desired area is **above** a certain number, you will have to use subtraction and symmetry, as **DISTR invNorm(only** reports values below, or to the left.



Goals: Interpret standard deviation as a measure of spread. Introduce normal curve. Use TI-83 in place of the standard normal table in the text.

Skills:

- **Understand standard deviation.** At first, standard deviation will seem foreign to you, but I believe that it will make more sense the more you become familiar with it. In its simplest terms, the standard deviation is non-negative number that measures how “wide” a dataset is. One common interpretation is that the range of a data set is about 4 or 5 standard deviations. Another interpretation is that the standard deviation is roughly $\frac{3}{4}$ times the IQR; that is the standard deviation is a bit smaller than the IQR. Eventually we will use the standard deviation in our calculations for statistical inference; until then, this measure is just another summary statistic, and getting used to this number is your goal. The normal curve of the next section will further help us understand standard deviation.
- **Know what a z -score is (standardization).** Sometimes, instead of knowing a variable’s actual value, we are only interested in how far above or below average it is. This information is contained in the z -score. Negative values indicate a below average observation, while positive values are above average. If the list follows a normal distribution (the familiar “bell-shaped” curve) then it will be relatively rare to have values below -2 or above $+2$ (only about 5 % of cases). Even if the list is not normal, surprisingly the z -score still tends to have few values beyond ± 2 , although this is not guaranteed.
- **Using the TI-83 to find areas under the normal curve.** When we have a distribution that can be approximated with the bell-shaped normal curve, we can make accurate statements about frequencies and percentages by knowing just the mean and the standard deviation of the data. Our TI-83 has 2 functions, **DISTR normalcdf(** and **DISTR invNorm(** which allow us to calculate these percentages more easily and more accurately than the table in the text. We use **DISTR normalcdf(** when we want the percentage as an answer and we use **DISTR invNorm(** when we already know the percentage but not the value that gives that percentage.

Reading: Chapters 3 and 4.

Day 5

Activity: Practice normal calculations, using real situations (except 2). Creating scatter plots. Calculating the correlation coefficient. **Quiz 2 today.**

Answer the following questions:

- 1) Suppose SAT scores are distributed normally with mean 800 and standard deviation 100. Estimate the chance that a randomly chosen score will be above 720. Estimate the chance that a randomly chosen score will be between 800 and 900. The top 20 % of scores are above what number? (This is the 80th percentile.)
- 2) Find the Inter Quartile Range (IQR) for the standard normal (mean 0, standard deviation 1). Remember that the first quartile is the 25th percentile and the third quartile is the 75th percentile, so $\frac{1}{4}$ of the data are below the first quartile and $\frac{1}{4}$ of the data are above the third

quartile. Observe and note whether this IQR is larger or smaller than 1, the standard deviation.

- 3) Women aged 20 to 29 have normally distributed heights with mean 64 and standard deviation 2.7. Men aged 20 to 29 have mean 69.3 with standard deviation 2.8. What percent of 20 something women are taller than the average 20 something man, and what percent of 20 something men are taller than the average 20 something woman?
- 4) Pretend we are manufacturing fruit snacks, and that the average weight in a package is .92 ounces with standard deviation 0.04. What should we label the net weight on the package so that only 5 % of packages are “underweight”?
- 5) Suppose that your average commute time to work is 20 minutes, with standard deviation of 2 minutes. What time should you leave home to arrive to work on time at 8:00? (You may have to decide a reasonable value for the chance of being late.)

Today we also begin Unit 2, summarizing two variables together. So far we have dealt with just one list of data at a time. The graphical and numerical techniques let us know how the data were spread out, whether there was a skew, where it was centered, etc. With two variables (two lists of data) we will have one graphical display (scatter plots) and one numerical display (correlation). But we will add a new sort of summary procedure to our list of techniques: regression. Regression is a summarized description of the pattern in a scatter plot of data.

The correlation coefficient is a unit-less number between -1 and +1. The two extremes represent data that falls perfectly on a line, with negative or positive slope, respectively. A correlation of 0 represents data that show no trend whatsoever: your prediction of the y -value is unaffected by the x -value you choose. These facts can be explored with the following activity.

Using the [Arizona Temps](#) data, make a scatter plot of “Flagstaff High” versus “Phoenix High”. Then guess what the correlation coefficient might be *without* using your calculator. Use the sample diagrams on page 102 to guide you. Finally, using your calculator, calculate the actual value for the correlation coefficient and compare it to your guess.

Repeat for the variables “Flagstaff High” and “Flagstaff Low”, or for any other pair.

We will have to use the regression function **STAT CALC LinReg(ax+b)** to calculate correlation, r . If you type **ENTER** after the **STAT CALC LinReg(ax+b)** command, the calculator assumes your lists are in columns **L1** and **L2**; otherwise you will type where they are, for example **STAT CALC LinReg(ax+b) L2, L3**. **IMPORTANT NOTE:** You will have to have pressed **DiagnosticOn**; access this command through the **CATALOG (2nd 0)**. You will only need to turn the diagnostics on once, unless you happen to have your RAM cleared on your TI-83.

Goals:

Master normal calculations. Realize that summarizing using the normal curve is the ultimate reduction in complexity, but only applies to data whose distribution is actually bell-shaped. Display two variables and measure (and interpret) linear association using the correlation coefficient.



Skills:

- **Memorize 68-95-99.7 rule.** While we do rely on our technology to calculate areas under normal curves, it is convenient to have some of the values committed to memory. Roughly **68 %** of data from a normal curve are within **one** standard deviation of the mean; roughly **95 %** of data from a normal curve are within **two** standard deviations of the mean; and roughly **99.7 %** of data from a normal curve are within **three** standard deviations of the mean. These values can be used as rough guidelines; if precision is required, you should use the TI-83 instead. (For example, the area between **exactly** -2 and **exactly** +2 is really 95.45 %, not 95 %.) I will assume you know these three numbers by heart when we encounter the normal numbers again in Chapters 13 through 22. An important observation to make here is that rarely is a value from the normal curve further than 3 standard deviations from the center of the curve.
- **Understand that summarizing with just the mean and standard deviation is a special case.** We have progressed from graphs like histograms and quantile plots to summary statistics like medians, means, and standard deviations to finally summarizing an entire list with just two numbers: the mean and the standard deviation. However, this last step in our summarization **only** applies to lists whose distribution resembles the bell-shaped normal curves. If the data's distribution is skewed, or has any other shape, this level of summarization is insufficient. Also, it is important to realize that these calculations are only approximations; usually in the real world data is only **approximately** normally distributed.
- **Plot data with a scatter plot.** This will be as simple as entering two lists of numbers into your TI-83 and pressing a few buttons, just as for histograms or box plots. Or, if you are doing plots by hand you will have to first choose an appropriate axis scale and then plot the points. You should also be able to describe overall patterns in scatter diagrams and suggest tentative models that summarize the main features of the relationship, if any.
- **Use the TI-83 to calculate the correlation coefficient.** Unfortunately, there isn't a quick way to have the TI-83 find the correlation coefficient; we must use the regression function **STAT CALC LinReg(ax+b)**. However, without the scatter plot, it is a mistake to assume the correlation coefficient automatically is a good summary of a relationship. We will learn more about this in Chapter 5. **IMPORTANT NOTE:** You will have to have pressed **DiagnosticOn**; access this command through the **CATALOG (2nd 0)**.
- **Interpret the correlation coefficient.** You should know the range of the correlation coefficient (-1 to +1) and what a "typical" diagram looks like for various values of the correlation coefficient. Again, page 102 is your guide. You should recognize some of the things the correlation coefficient does **not** measure, such as the strength of a **non-linear** pattern. Also, in contrast to the regression results we will learn starting on Day 9, **it does not matter** in what order the variables are entered; correlation is a fact about a **pair** of variables.

Reading: Chapters 1 through 3.

Day 6

Activity: Presentation 1. Graphical and Numerical Summaries. Unit 1 Review. **Homework 2 due today.**



Gather 3 to 5 variables on at least 20 subjects; the source is irrelevant, but knowing the data will help you explain its meaning to us. Summarize **each** of your lists graphically for us, and numerically, where appropriate. Also make sure you have gathered at least one numerical and at least one categorical variable. Your goal today is to demonstrate that you can summarize data, **both** graphically **and** numerically.

Day 7

Activity: Exam 1.

This first exam will cover graphical summaries (pictures), numerical summaries (summary calculations), and normal curve calculations (areas under the bell curve). Some of the questions may be multiple choice or True/False. Others may require you to show your worked out solution. Section reviews are an excellent source for studying for the exams. Don't forget to review your class notes and these notes.

Day 8

Activity: Outlier effects on correlation. Using the Olympic data, fit a regression line to predict the 2012 race results.

We are not going to focus on the actual formula used to calculate the correlation coefficient. However, it **is** important to note that it involves summation, which is akin to averaging. Therefore, any “problems” we have with averages will also exist with correlation. Specifically, extreme values (outliers) will influence the results. Unfortunately, in two dimensions we encounter a new sort of outlier: it is far from the pattern but is not an outlier in either variable. Today we will find out what effect a single point that is different from the pattern has on the correlation coefficient. **IMPORTANT NOTE:** Today's first activity is one of those we will do for understanding, and is not something you would do in real life to a data set.

The data set we will explore today has 7 data points.

	x	y
	1	4
	2	3
	2	5
	3	4
	4	8
	4	6
	5	9
First new point	3	8
Second new point	7	10
Third new point	7	1

First, plot the original seven points and calculate the correlation coefficient. Now add the eighth point in three different places (making three new data sets, each with 8 points) and for each new data set, calculate the correlation coefficient. Compare the values, noting where the

eighth data point (the outlier) was and how it changed the correlation. Summarize the effect of outliers in a paragraph.

Now that we know how to produce a scatter plot of data, and calculate a correlation coefficient, we will want to further describe the relationship between the two variables, if they lie roughly on a straight line. The chief technique for summarizing such a **linear** relationship is **Least Squares Linear Regression**. This technique is also known as **Least Squares Regression**, **Best Fit Regression**, **Linear Regression**, etc. The important point is that we are going to describe the relationship with a straight line, so if the scatter plot shows some other shape, this technique will be inappropriate. Your tasks are to 1) come up with a line, either by hand or with technology, that “goes through” the data in some appropriate way, 2) to be able to use this model to describe the relationship verbally, and 3) to predict numerically y-values for particular x-values of interest.

Begin by making a scatter plot of the race times. If you want a rough guess for the slope of the best fitting line through the data, you can connect two points spaced far apart (I will show you the details in class.)

Next, use the TI-83's regression features to calculate the **best fit**. The command is **STAT CALC LinReg(ax+b)**, assuming the two lists are in **L1** and **L2**. (**L1** will be the horizontal variable, years in this case.) (We used this command on Day 5 also. However, for regression it is **vital** that you get the order correct; the idea here is that you are predicting the vertical variable from knowing the horizontal variable.)

Interpret what your two regression coefficients mean. Make sure you have units attached to your numbers to help with the meanings.

Have the calculator type this equation into your **Y=** menu (using **VARS Statistics EQ RegEQ**), and **TRACE** on the line to predict the future results. Specifically, see what your model says the 2008 time should have been. Compare to the actual value and check how predictive our model is.

Here are the data: 100-meter dash winning Olympic times:

1896	Thomas Burke, United States	12 sec		
1900	Francis W. Jarvis, United States	11.0 sec		
1904	Archie Hahn, United States	11.0 sec		
1908	Reginald Walker, South Africa	10.8 sec		
1912	Ralph Craig, United States	10.8 sec		
1920	Charles Paddock, United States	10.8 sec		
1924	Harold Abrahams, Great Britain	10.6 sec		
1928	Percy Williams, Canada	10.8 sec	Elizabeth Robinson, United States	12.2 sec
1932	Eddie Tolan, United States	10.3 sec	Stella Walsh, Poland (a)	11.9 sec
1936	Jesse Owens, United States	10.3 sec	Helen Stephens, United States	11.5 sec
1948	Harrison Dillard, United States	10.3 sec	Francina Blankers-Koen, Netherlands	11.9 sec
1952	Lindy Remigino, United States	10.4 sec	Marjorie, Jackson, Australia	11.5 sec
1956	Bobby Morrow, United States	10.5 sec	Betty Cuthbert, Australia	11.5 sec
1960	Armin Hary, Germany	10.2 sec	Wilma Rudolph, United States	11.0 sec
1964	Bob Hayes, United States	10.0 sec	Wyomia Tyus, United States	11.4 sec

1968	Jim Hines, United States	9.95 sec	Wyomia Tyus, United States	11.0 sec
1972	Valery Borzov, USSR	10.14 sec	Renate Stecher, E. Germany	11.07 sec
1976	Hasely Crawford, Trinidad	10.06 sec	Annegret Richter, W. Germany	11.08 sec
1980	Allen Wells, Britain	10.25 sec	Lyudmila Kondratyeva, USSR	11.6 sec
1984	Carl Lewis, United States	9.99 sec	Evelyn Ashford, United States	10.97 sec
1988	Carl Lewis, United States	9.92 sec	Florence Griffith-Joyner, United States	10.54 sec
1992	Linford Christie, Great Britain	9.96 sec	Gail Devers, United States	10.82 sec
1996	Donovan Bailey, Canada	9.84 sec	Gail Devers, United States	10.94 sec
2000	Maurice Greene, United States	9.87 sec	Marion Jones, United States	10.75 sec
2004	Justin Gatlin, United States	9.85 sec	Yuliya Nesterenko, Belarus	10.93 sec
2008	Usain Bolt, Jamaica	9.69 sec	Shelly-ann Fraser, Jamaica	10.78 sec
2012	?		?	

(a) A 1980 autopsy determined that Walsh was a man.

200-meter dash winning Olympic times:

1900	Walter Tewksbury, United States	22.2 sec		
1904	Archie Hahn, United States	21.6 sec		
1908	Robert Kerr, Canada	22.6 sec		
1912	Ralph Craig, United States	21.7 sec		
1920	Allan Woodring, United States	22 sec		
1924	Jackson Sholz, United States	21.6 sec		
1928	Percy Williams, Canada	21.8 sec		
1932	Eddie Tolan, United States	21.2 sec		
1936	Jesse Owens, United States	20.7 sec		
1948	Mel Patton, United States	21.1 sec	Francina Blankers-Koen, Netherlands	24.4 sec
1952	Andrew Stanfield, United States	20.7 sec	Marjorie Jackson, Australia	23.7 sec
1956	Bobby Morrow, United States	20.6 sec	Betty Cuthbert, Australia	23.4 sec
1960	Livio Berruti, Italy	20.5 sec	Wilma Rudolph, United States	24.0 sec
1964	Henry Carr, United States	20.3 sec	Edith McGuire, United States	23.0 sec
1968	Tommy Smith, United States	19.83 sec	Irena Szewinska, Poland	22.5 sec
1972	Valeri Borzov, USSR	20.00 sec	Renate Stecher, E. Germany	22.40 sec
1976	Donald Quarrie, Jamaica	20.23 sec	Barbel Eckert, E. Germany	22.37 sec
1980	Pietro Mennea, Italy	20.19 sec	Barbel Wockel, E. Germany	22.03 sec
1984	Carl Lewis, United States	19.80 sec	Valerie Brisco-Hooks, United States	21.81 sec
1988	Joe DeLoach, United States	19.75 sec	Florence Griffith-Joyner, United States	21.34 sec
1992	Mike Marsh, United States	20.01 sec	Gwen Torrance, United States	21.81 sec
1996	Michael Johnson, United States	19.32 sec	Marie-Jose Perec, France	22.12 sec
2000	Konstantinos Kenteris, Greece	20.09 sec	Marion Jones, United States	21.84 sec
2004	Shawn Crawford, United States	19.79 sec	Veronica Campbell, Jamaica	22.05 sec
2008	Usain Bolt, Jamaica	19.30 sec	Veronica Campbell-Brown, Jamaica	21.74 sec
2012	?		?	

Goals: Understand the impact of outliers on correlation. Gain further practice with the TI-83. Practice using regression with the TI-83. We want the regression equation, the regression line superimposed on the plot, the correlation coefficient, and we want to be able to use the line to predict new values.

Skills:

- **Interpret the correlation coefficient.** You should recognize how outliers influence the magnitude of the correlation coefficient. One simple way to observe the effects of outliers is to calculate the correlation coefficient with and without the outlier in the data set and compare the two values. If the values vary greatly (this is a judgment call) then you would say the outlier is “influential”.
- **Fit a line to data.** This may be as simple as “eyeballing” a straight line to a scatter plot. However, to be more precise, we will use least squares, **STAT CALC LinReg(ax+b)** on the TI-83, to calculate the coefficients, and **VARS Statistics EQ RegEQ** to type the equation in the **Y=** menu. You should also be able to sketch a line onto a scatter plot (by hand) by knowing the regression coefficients.
- **Interpret regression coefficients.** Usually, we want to only interpret slope, and slope is best understood by examining the units involved, such as inches per year or miles per gallon, etc. Because slope can be thought of as “rise” over “run”, we are looking for the ratio of the units involved in our two variables. More precisely, the slope tells us the change in the response variable for a one unit change in the explanatory variable. We don’t typically bother interpreting the intercept, as zero is often outside of the range of experimentation.
- **Estimate/predict new observations using the regression line.** Once we have calculated a regression equation, we can use it to predict new responses. The easiest way to use the TI-83 for this is to **TRACE** on the regression line. You may need to use up and down arrows to toggle back and forth from the plot to the line. You may also just use the equation itself by multiplying the new x -value by the slope and adding the intercept. (This is exactly what **TRACE** is doing.) Note: when using **TRACE**, and the x -value you want is currently outside the window settings (lower than **XMin** or above **XMax**) you must reset the window to include your x -value first.

Reading: Chapter 5.

Day 9

Activity: Revisit outliers data set, adding regression lines. **Quiz 3 today.**

While we have not focused on the actual formulas used to calculate the regression coefficients, it is important to note that they involve summation (just as it was in correlation), which is akin to averaging. Therefore, any “problems” we have with averages will also exist with regression. Specifically, extreme values (outliers) will highly influence the results. Unfortunately, in two dimensions we encounter a new sort of outlier: it is far from the pattern but is not an outlier in either variable. Today we will find out what effect a single point that is different from the pattern has on the regression coefficients. **IMPORTANT NOTE:** Today’s activity is one of those we will do for understanding, and is not something you would do in real life to a data set.

The data set we will explore today is the same as that from Day 8 and has 7 data points. First, plot them and calculate the regression coefficients. Now add the eighth point in three different places (making three new data sets, each with 8 points) and for each new data set, calculate the

regression coefficients. Compare the values, noting where the eighth data point (the outlier) was and how it changed the values. Summarize the effect of outliers in a paragraph.

(You may use any extra time today to discuss Presentation 2 in your groups.)

Goals: Understand the impact of outliers on the regression coefficients. Gain further practice with the TI-83.

Skills:

- **Understand the limitations and strengths of linear regression.** Basically, linear regression should only be used with scatter plots that are roughly linear in nature. That seems obvious by the name. However, there is nothing that prevents us from **calculating** the numbers for any data set we can input into our TI-83's. We have to realize what our data looks like **before** we calculate the regression; therefore a scatter plot is an **essential** first step. In the presence of outliers and non-linear patterns, we should avoid drawing conclusions from the fitted regression line.

Reading: Chapters 5 and 6.

Day 10

Activity: Correlation/Regression summary. U. S. population example. Alternate regression models. Analyzing Two Categorical Variables. **Homework 3 due today.**

Today I want to show you an example similar to the work you will do for your presentation for the state population prediction. Your TI-83 has many other regression models available to you. The models that often are adequate for population models are exponential functions and logistic functions. The polynomial functions (quadratic, cubic, quartic) might fit the data well, yielding small residuals, but they are unlikely to be useful **predictive** models. If you zoom out, you will see unrealistic future effects, for example. One caution with the logistic function on the TI-83: it seems to be finicky. For some data sets it works extremely well, and for others it won't even produce fitted values.

Another modification you may want to try, instead of different models, is different time periods. For example, to predict the population the year 2020, data from before 1850 may not be helpful for developing the model, especially if factors affecting population growth have changed substantially.

Finally, I have listed below some summary statements about correlation and regression that you should be aware of before the second exam.

- 1) For correlation, the variables can be entered in any order; correlation is a fact about a **pair** of variables. For regression, the order the variables are presented matters. If you reverse the order, you get a different regression line.
- 2) We must have **numerical** variables to calculate correlation. For categorical variables, we will use contingency tables, which we will explore next.

- 3) High correlation does not necessarily mean a straight-line scatter plot. The U. S. population growth is an example.
- 4) Correlation is not resistant to the effects of outliers; the data set from Days 8 and 9 showed that the placement of a single point in the scatter plot could greatly influence the value of the correlation and the coefficients in the regression equation.

Is there really a difference between home and away won/loss records in sports? Baseball managers sometimes “platoon” their right- and left-handed batters based on the hand of the opposing pitchers. Can we see evidence of this?

We cannot make a scatter plot with categorical data. Our next best option is to make a **contingency table**. This is simply a cross-classification of the data values. A simple example is the win/loss, home/away record for a team. It is quite easy to make such a table; we just count how many items in the population fit into each cell.

The **real** question is whether the data shows anything meaningful. By that I mean do the categories show any departure from what would be expected if everything were just random? Before we answer this, we will usually want to summarize percentages from the table, including marginal percentages.

We will examine an interesting paradox today called **Simpson’s paradox**. Can averaging over a third variable reverse a trend? The answer is yes, but the reasons are subtle. Hopefully, we will see why this paradox happens, and be able to explain it.

We will look at several examples today, and for each one, we will have a table of counts. For each table, produce new tables of overall percentages, row percentages, and column percentages. You need to practice interpreting these percentages to be able to make quality descriptions of your data.

I will also give you a few examples of Simpson’s paradox. Your task is to recognize why the paradox has happened (it’s related to disparate sample sizes).

Goals: See scatter plots and correlation in practice. Understand correlations limitations and features. Introduce association for categorical variables. Explore Simpson’s paradox.

Skills:

- **Recognize the proper use of correlation, and know how it is abused.** Correlation measures **straight-line** relationships. Any departures from that model make the correlation coefficient less reliable as a summary measure. Just as for the standard deviation and the mean, outliers influence the correlation coefficient. Therefore, it is extremely important to be aware of data that is unusual. Some two-dimensional outliers are hard to detect with summary statistics; scatter plots are a must then.
- **Understand that there are competing regression models.** We have focused our attention on the **linear** regression models, but as we see on our TI-83, there are many other potential

models. If you use an alternate model, be sure to plot the fitted line along with the scatter plot. Be careful though; our usual measure of fit, r^2 , will not accurately tell the story.

- **Create a table summarizing two categorical variables.** Unlike numerical variables, summarization of categorical data is accomplished by making frequency tables. Often along with the tables, one will calculate marginal totals and percentages.
- **Understand that cause and effect is difficult to establish.** The slogan is “Association is not the same as causation.” We will encounter this many times throughout the rest of the course. In the next set of material (Chapters 8 and 9) we will discuss ways to produce data from which we **can** draw conclusions about causation.
- **Recognize Simpson’s paradox.** Sometimes when data is summarized over several sub-categories, an association can be reversed. It seems contrary to good common sense, but it is actually the effects of a lurking variable, and the phenomenon is known as Simpson’s paradox. You should be able to recognize situations where this paradox might occur. Not all tables of categorical variables will exhibit this paradox; the tables must be comparing rates over several groups and the sample sizes must be just so for the paradox to be present.

Reading: Chapter 6.

Day 11

Activity: Expected Tables. **Quiz 4 today.**

Our first look at inference will be to see what things would look like if everything were random. We will make the **expected table** using our TI-83’s. At the same time, the calculator will give us a number (a P-value) that will help us decide if there is any pattern present. Key characteristics of the expected table are that the row and column totals are identical to the original data table, but the individual cell totals are proportional to the marginal totals. That is, the percentage of cases falling in any column is the same across all rows and vice versa. See the class examples. (Note: we will explore P-value in much more detail in Unit 4. For now, just focus on the rule given below for deciding if the variables are unrelated.)

The expected table represents how things would have worked out if the two variables were unrelated. We will go through examples in class to explain this phenomenon. It is essentially a “what-if” type argument. “What would things look like if the two variables were unrelated?” Then we compare that situation to what has actually occurred, and if the difference is too large, then we conclude that “dumb luck” is not the most likely explanation.

We will use the examples from Day 10 to explore whether there are relationships among the variables.

Goals: Develop intuition for when the observed and expected tables are too different.

Skills:

- **Create the table of expected counts.** The primary method of analyzing categorical tables is comparing the observed data to a table of expected counts. The TI-83 will calculate the expected table for us. Our job is to understand the meaning of the numbers. Basically, the

expected table is the way the table would have come out if the two variables were unrelated. We use it as a baseline in determining association. (This material comes from Chapter 23, but I will not expect you to master Chapter 23.)

- **Recognize when an association is present.** When two categorical variables are associated (much like when two numerical variables are correlated) we detect this with the χ^2 test. We will use a statistical technique to decide if the differences in the tables are too great, **STAT TESTS χ^2 -Test**. You must have the observed table in a matrix. The expected table will be stored in another matrix. If $p < 0.05$, we conclude the two tables are quite different. Our reasoning is this: if the difference between the actual results and the results assuming no association is a small difference, then we have no reason to think that the variables are related. However, if the difference between the two tables is considered large, then we conclude something can be said about the relationship; that is, that one exists.

Reading: Chapters 4 through 7.

Day 12

Activity: Presentation 2. Regression and Correlation. Unit 2 Review. **Homework 4 due today.**

Pick one of the 50 states (the data is at the end of this file). Predict the population in the year 2020 using a regression function (not necessarily linear). Describe how you decided upon your model, and explain how good you think your prediction is. It may be helpful for you to describe what you tried that **didn't** work, so that we can appreciate your choices. Your goal is to convince us to believe your prediction.

Day 13

Activity: Exam 2.

This second exam will cover scatter plots, correlation, regression (two-variable summaries), and two-variable categorical analyses. Some of the questions may be multiple choice or True/False. Others may require you to show your worked out solution. Section reviews are an excellent source for studying for the exams. Don't forget to review your class notes and these notes.

Day 14

Activity: History of polls. Creating random samples.

Unit 3 is about sampling and probability. In addition to sampling, we will also explore the related topics of observational studies and experiments. The underpinnings to Unit 4, statistical inference, rely on knowing precisely how likely a particular outcome is. If the samples we collect are done with no organized method, and without using probability rules, then we will be unable to assess the likelihood of an outcome. Thus, studying sampling and probability is crucial to being able to believe statistical inference in scientific studies.

A very high profile use of sampling is presidential elections. In the early years of the 20th century, many attempts were made to predict elections beforehand. Below are some

consequences of the haphazard “samples of convenience” that were conducted. Because of the lessons learned during these fiascos, we now have very rigorous methods of producing good data.

to 1936: Selection Bias

Literary Digest calls the 1936 election for Alf Landon (remember him), in an Electoral College landslide. The poll was performed by sending postcards to people with telephones, magazine subscribers, car owners, and a few people on lists of registered voters. The first problem was that the sample was biased toward Republican-leaning voters, who could afford magazine subscriptions and telephones during the Great Depression. The second problem is the response rate error: they received only 2.3 million responses, for a 23 % response rate. This was compounded by a volunteer error: the respondents are most likely people who wanted change, and Roosevelt is president. Nevertheless, Franklin Roosevelt, the Democrat, wins in a landslide. Statistician Jessica Utts (Seeing Through Statistics, Jessica M. Utts, Duxbury Press, Wadsworth Publishing Co., 1996 pages 65-66), who examined issues of the Literary Digest from 1936, says “They were very cocky about George Gallup predicting they would get it wrong. [Gallup helped make his reputation in polling by correctly calling the race.] The beauty of something like that is that the winner is eventually known.” [Taken from: <http://whyfiles.org/009poll/fiasco.html>]

to 1948: Quota Sampling

Aided by erroneous polling, newspapers prematurely call the presidential election for challenger Thomas Dewey, leading to the famous photograph of an elatedly re-elected Harry Truman. The problematic polls use then-popular “quota-sampling” techniques. In other words, they sought out a certain number of men, a certain number of women, and similarly for blacks, whites, and various income levels. According to statistician Fritz Scheuren, quota sampling in political polling was abandoned after this debacle in favor of random sampling of the population. [Taken from: <http://whyfiles.org/009poll/fiasco.html>]

to present: Random Sampling

Modern sampling techniques use randomness to select participants. Interviewers are not allowed to choose people; and they are required to read questions from a card. The purpose of sampling is to gather unbiased information, and each of the techniques we employ is an attempt to keep the information untainted.

I will give you the numbers and we can see how the polls have improved since 1936, and if there are still any biases towards either Republicans or Democrats.

The easiest form of random sampling is **simple random sampling**. Under this method, each possible sample could be the one selected. In practice, with samples of 1,000 people out of populations of 100 million, there are an incomprehensibly large number of possible samples. This means that practically, we do not actually take simple random samples, and other methods must be employed. Today, though, we will get at the understanding of the procedure. **IMPORTANT NOTE:** Today’s second activity is one of those we will do for understanding, and

is not something you would do in real life. In practice, you would take one sample and be done. Today each group will take 80 samples.

We will use three methods of sampling today: dice, cards, Table B in our book, and our calculator. To make the problem feasible, we will only use a population of size 6. (I know this is unrealistic in practice, but the point today is to see how randomness works, and hopefully trust that the results extend to larger problems.) Pretend that the items in our population (perhaps they are people) are labeled 1 through 6. For each of our methods, you will have to decide in your group what to do with “ties”. Keep in mind the goal of simple random sampling: at each stage, each remaining item has an equal chance to be the next item selected. Also, remember here we are perhaps interviewing **exactly** three people. Keep careful track of which samples you selected; record your results in order, as 125 or 256, for example. (125 would mean persons 1, 2, and 5 were selected.)

- 1) Dice: By rolling dice, generate a sample of three people. (Let the number on the die correspond to one of the items.) Repeat 20 times, giving 20 samples of size 3.
- 2) Cards: I will give each groups 6 cards, ace through six. Shuffle the cards, select three, and record the sample. Repeat 20 times, giving 20 samples of size 3.
- 3) Table B: Using Table B, starting at any haphazard location, select three people. (Let the random digit correspond to one of the items.) Repeat 20 times, giving 20 more samples of size 3. (See page 198 for an example of how to use Table B.)
- 4) Calculator: Using your TI-83, select three people. select three people. The TI-83 command **MATH PRB randInt(2, 4, 5)** will produce 5 numbers between 2 and 4, inclusive, for example. (If you leave off the third number, only one value will be generated.) Repeat 20 times, giving 20 more samples of size 3.

Your group should have drawn 80 samples at the end. We will pool the results of everyone’s work together on the board.

IMPORTANT NOTE: I want you to be clear what we mean by the word **random**. When we refer to “randomness”, we are generally talking about the results of a mathematical model. Unfortunately in English, random often means something different. For example, people think random means “without a plan, or without conscious decision”. In statistics, we actually mean “with equal chance”. I will try to remind you of this the further we get into our probability discussion.

Goals: Introduce sampling. Identify biases. Explore why **non-random** samples are not trustworthy. Gain practice taking random samples. Understand what a simple random sample is. Become familiar with **MATH randInt(**. Accept that the calculator produces randomness.

Skills:

- **Understand the issues of bias.** We seek representative samples. The “easy” ways of sampling, samples of convenience and voluntary response samples, may or may not produce good samples, and because we don’t know the chances of subjects being in such samples, they

are unreliable sampling methods. Even when probability methods are used, biases can spoil the results. Avoiding bias is our chief concern in designing surveys.

- **Huge samples are not necessary.** One popular misconception about sampling is that if the population is large, then we need a proportionately large sample. This is just not so. My favorite counter-example is our method of tasting soup. To find out if soup tastes acceptable, we mix it up, then sample from it with a spoon. It doesn't matter to us whether it is a small bowl of soup, or a huge vat, we still use only a spoonful. The situation is the same for statistical sampling; we use a small "spoon", or sample. The fundamental requirement though is that the "soup" (our population) is "well mixed" (as in a simple random sample - see Day 14). In practice, most nation-wide samples have only a few thousand participants.
- **Know the definition of a Simple Random Sample (SRS).** Simple Random Samples can be defined in two ways: 1) An SRS is a sample where, at each stage, each item has an equal chance to be the next item selected. 2) A scheme where every possible sample has an equal chance to be the sample is an SRS.
- **Select an SRS from a list of items.** The TI-83 command **MATH randInt(** can be used to select numbered items from a list randomly. If a number selected is already in the list, ignore that number and get a new one. Remember, as long as each **remaining** item is equally likely to be chosen as the next item, you have drawn an SRS.
- **Understand the real world uses of SRS's.** In practice, simple random samples are not that common. It is just too impractical (or impossible) to have a list of the entire population available. However, the idea of simple random sampling is essentially the foundation for all the other types of sampling. In that sense then it is very common.

Reading: Chapter 8.

Day 15

Activity: Alternate Sampling Schemes.

Today we will explore other sampling schemes than the SRS. In practice, as we have discussed, it is impractical or impossible to collect simple random samples. This does **not** mean that we cannot gather useful information, but that we may have to live with some biases. Of course we will try to eliminate as many biases as we can. We will look at **systematic sampling**, **stratified sampling**, and **cluster sampling** in today's exercise. In each method, you will calculate an estimate for the entire U. S. based only on your sample results. Keep track of your results, as we will pool the class results to see the effects of the different methods.

Using a small population, we will explore alternate sampling schemes. Our population today is the 50 United States and the variable of interest is the **Land Area**.

Simple Random Sampling: Randomly select 10 states. Find the average Land Area and the standard deviation of Land Area for your 10 states. To guess the total for the entire U. S., multiply your average by 50.



Systematic Sampling: Using the alphabetically sorted list, choose a random number between 1 and 5. Then select every 5th item after that. In general, you would use a random number between 1 and N/n , and then select every N/n^{th} item. Find the average Land Area and the standard deviation of Land Area for your 10 states. To guess the total for the entire U. S., multiply your average by 50.

Stratified Sampling: From the large states (Alaska, Texas, California, Montana, New Mexico, Arizona, Nevada, Colorado, Wyoming, and Oregon), randomly select 4 states. From the small states (Rhode Island, Delaware, Connecticut, Hawaii, New Jersey, Massachusetts, New Hampshire, Vermont, and Maryland), randomly select 1 state. From the remaining 31 states, randomly select 5 states.

To guess the total Land Area for the entire U. S., guess the totals for the three strata separately. That is, for the small states, multiply your average for them by 9. For the large states, multiply by 10. For the other states, multiply by 31. Then add the 3 sub-totals together. In addition, calculate the standard deviation for your 10 states chosen, by entering them in a list in the TI-83 and performing **1-Var Stats**.

Cluster Sampling: Using the following breakdown of the states (North, Southwest, Central, Southeast, and Northeast), randomly select two regions. From each of the two regions, randomly select 5 states.

North: Alaska, Washington, Oregon, Idaho, Montana, Wyoming, N. Dakota, S. Dakota, Nebraska, and Minnesota.

Southwest: Hawaii, California, Nevada, Utah, Arizona, Colorado, New Mexico, Kansas, Oklahoma, and Texas.

Central: Iowa, Missouri, Arkansas, Wisconsin, Michigan, Illinois, Indiana, Ohio, Kentucky, and Tennessee.

Southeast: Louisiana, Mississippi, Alabama, Georgia, Florida, S. Carolina, N. Carolina, Virginia, W. Virginia, and Maryland.

Northeast: Pennsylvania, Delaware, New Jersey, New York, Vermont, New Hampshire, Massachusetts, Connecticut, Rhode Island, and Maine.

Find the average Land Area and the standard deviation of Land Area for your 10 states. To guess the total for the entire U. S., multiply your average by 50.

For our entire class, which method produced the most reliable results? I will summarize the benefits and drawbacks of each method, despite whether we actually saw these effects in our simulation today.

	State	Land Area		State	Land Area		State	Land Area
1	Alabama	50,744	18	Louisiana	43,562	35	Ohio	40,948
2	Alaska	571,951	19	Maine	30,862	36	Oklahoma	68,667
3	Arizona	113,635	20	Maryland	9,774	37	Oregon	95,997

4	Arkansas	52,068	21	Massachusetts	7,840	38	Pennsylvania	44,817
5	California	155,959	22	Michigan	56,804	39	Rhode Island	1,045
6	Colorado	103,718	23	Minnesota	79,610	40	South Carolina	30,109
7	Connecticut	4,845	24	Mississippi	46,907	41	South Dakota	75,885
8	Delaware	1,954	25	Missouri	68,886	42	Tennessee	41,217
9	Florida	53,927	26	Montana	145,552	43	Texas	261,797
10	Georgia	57,906	27	Nebraska	76,872	44	Utah	82,144
11	Hawaii	6,423	28	Nevada	109,826	45	Vermont	9,250
12	Idaho	82,747	29	New Hampshire	8,968	46	Virginia	39,594
13	Illinois	55,584	30	New Jersey	7,417	47	Washington	66,544
14	Indiana	35,867	31	New Mexico	121,356	48	West Virginia	24,078
15	Iowa	55,869	32	New York	47,214	49	Wisconsin	54,310
16	Kansas	81,815	33	North Carolina	48,711	50	Wyoming	97,100
17	Kentucky	39,728	34	North Dakota	68,976			

Goals: Understand the differences (good and bad) between various sampling schemes.

Skills:

- **Systematic Sampling.** In systematic sampling, we decide on a fraction of the population we would like to sample, and then randomly select a number between 1 and N/n , the population size divided by the sample size. Then from the list of items, we take every N/n^{th} item. This scheme works best when we have a list of items, and it is simple to select items periodically from that list. For example, we may have decided to take every 20th name, and there are 20 names per page. It would be thus very convenient to take the 11th item from each page, for example.
- **Stratified Sampling.** In stratified random sampling, we draw a simple random sample from several sub-populations, called strata. The sample size may differ by stratum; in fact, this leads to the most efficient use of stratified sampling. This scheme works best when there is variability between strata. For example, one stratum may be more homogeneous than another stratum. Then, just a few items are needed from the stratum that is homogenous, since all the items in that stratum are basically the same. Alternatively, from a stratum that is quite diverse, you should sample more heavily. In our example in class today, the large states are quite different from one another in land areas, so we sampled 40 % of them. The small states are very similar in area, so we sampled only 11 % of them. The remaining 31 states are in between, so we sampled 16 % of them.
- **Cluster Sampling.** In cluster sampling, we select several groups of items. Within each selected group, we then repeat by selecting from several smaller groups. This process continues until we have our ultimate sampled units. This scheme works best when we do not have a list of all the items in a population, but we have lists of **groups** of items, for example states or counties. One drawback of this method is that the ultimate clusters are generally quite similar (all people living on a block in a town are generally more similar than different), so the effective sample size is lower than it appears. To compensate, cluster samples typically have larger sample sizes.



Reading: Chapter 9 and Data Ethics.

Day 16

Activity: Lurking variables exercises.

Many times, the conclusion drawn about a cause-and-effect relationship is incorrect, because of various alternate explanations. We will explore these today through examples. What you are looking for on these problems are these situations:

- 1) Reversed cause and effect. It may be that instead of X causing Y , Y causes X . Just because two things occur together does not mean that the first causes the second.
- 2) Common response. These are the “lurking” variables. A third unobserved variable might be causing each of the two observed variables. Critics of the anti-smoking movement were (correctly) pointing out that there might be a third set of variables (such as genetics) that cause both smoking propensity **and** cancer susceptibility.
- 3) Coincidence. There may be nothing to the association other than luck. With large samples, though, we tend to discount the luck notion.

The main conclusion from today’s examples is that in observational studies, there are **always** potential confounding factors. We may try to control for the other factors, such as by looking at different age categories separately, but critics can always argue that a critical variable that explains the observed association has not been measured.

I have a set of problems from *Statistics*, by Freedman, Pisani, and Purves, which I think will help us think about looking for alternative explanations than the proffered “attractive” conclusion. Work on each problem for a few minutes; we will discuss each of them in detail before the class is over. Be prepared to defend your explanations.

Goals: Explore experimentation ideas. Discover potential lurking variables and ways to control for them.

Skills:

- **Examine situations and detect lurking variables.** When searching for lurking variables, it is not enough to suggest variables that might also explain the response variable. Your potential lurking variable must also be associated with the explanatory variable. So, for example, suppose you are trying to predict a child’s height using their weight. A possible lurking variable might be age, because a child’s age tends to be associated with both weight **and** height. On the other hand, a variable associated with height that is unlikely to be related to weight (and therefore would **not** be a lurking variable) is arm span.
- **Know appropriate ways to attempt to control for lurking variables.** Once a potential lurking variable has been identified, we can make more appropriate comparisons by dividing our subjects into smaller, more homogeneous groups and **then** making the comparison. This is why you will see comparisons made by age groups, for example.

- **Understand that experimentation, done properly, will allow us to establish cause-and-effect relationships.** Observational studies have lurking variables; we can try to control for them by various methods, but we cannot eliminate them. If the data is collected appropriately through good experimentation, however, the effects of lurking variables can be eliminated. This is done through randomization, the thinking being that if a sample is large enough, it can't realistically be the case that all of one group contains **all** the large values of a lurking variable, for example.

Reading: Chapter 10.

Day 17

Activity: What is Randomness? **Quiz 5 today.**

Our formal probability theory is based on the “long run”, but our everyday lives are dominated by the “short run”. This contradiction will follow us throughout our discussions for the rest of the course. I believe it is an important issue, and I will try to convince you of that too. To begin our probability material, we will explore the most basic idea we have of randomness: coin flipping. I believe that if we understand coin flipping, we can understand the other models of randomness we will encounter.

Today we will look at some everyday sequences to see if they exhibit “random” behavior.

Coin experiment 1: Write down a sequence of H's and T's representing head and tails, pretending you are flipping a coin. Then flip a real coin 50 times and record these 50 H's and T's. Without knowing which list is which, in most cases I will be able to identify your real coin.

Pro sports teams: In sports you often hear about the “hot hand” or “momentum”. We will pick a team, look at their last few games, and see if flipping a coin will produce a simulation that resembles their real performance. Then we will examine whether we could pick out the simulation without knowing which was which.

Coin experiment 2: Spin a penny on a flat surface, instead of tossing it into the air. Record the percentage of heads.

Coin experiment 3: Balance a nickel on its edge on a flat surface. Jolt the surface enough so that the nickel falls over, and record the percentage of heads.

If it is possible to list the entire sample space for an experiment, you can answer any probability question asked by just counting. For example, with two dice, to find the chance of rolling doubles, look at the diagram from class and count how many of the 36 elements are doubles. Then express the answer as a ratio (out of 36 in this case). We will do examples like this in class before you do the three exercises below. Sometimes, however, listing the entire sample space is a daunting task, especially if the sample space is quite large. In those cases, it may be more appropriate to approximate the answer by simulating the experiment; that is actually perform the experiment a large number of times. The difficulty with simulation is that we often will need an extremely large number of trials before the results are trustworthy.

We will try both methods.

Using both complete sampling spaces (theory) and simulation, find (or estimate) these chances:

- 1) Roll two dice, one colored, one white. Find the chance of the colored die being less than the white die. Do this part both with a theoretical sampling space diagram **and** with a simulation of 100 dice rolls. If you would rather use the calculator instead of actually rolling two dice 100 times, that would be fine. One way to roll dice with the calculator is to use `(randInt(1, 6) > randInt(1, 6))`. (The `>` key is found in the **TEST** menu.) This command will record a “1” if the result is true and a “0” if it is false. Using this clever trick of the calculator, we can essentially convert dice roll numbers into yes/no responses. Note that by continuing to hit **ENTER** the command will be repeated; this way you do not have to retype for every trial. However, a more efficient method is to use `seq()`, found in the **LIST OPS** menu. The command is `seq((randInt(1, 6) > randInt(1, 6)), X, 1, 100) -> L1`. The simulated results have been stored in **L1**, and you can now find out how many times the first die was larger than the second die by using **1-Var Stats** and looking at the average, which will be a percentage now. However, see the skill note below for simulations; you may need to do this too many times to make it worthwhile.
- 2) Roll three dice and find the chance that the largest of the three dice is a 6. (If two or more dice are tied for the largest, use that value; for example, the largest value when 6, 6, 4 is rolled is 6.) To simulate this one, you can keep track of what the largest number is on three rolls using `max(randInt(1, 6, 3))`. **max** is found in the **LIST MATH** menu. You will still use `seq`, but substitute the expression with **max** for the expression with the `>` from 1. (I underlined them for you). After the simulation stores the results in a list, we can make a histogram to see how many 6's (or any other number) were produced. Again, see the skill note below for simulations; you may need to do this too many times to make it worthwhile.
- 3) Roll three dice and find the chance of getting a sum of less than 8. I suggest drawing sample spaces to do this one. You do not need to draw the entire sample space for three dice, but maybe by drawing a portion of it you will be able to see what you need to pay attention to. I will give you hints in class when you get to this one if you ask me. The simulation expression for this one is `sum(randInt(1, 6, 3))`. **sum** is in the same menu as **max**. Once again, see the skill note below for simulations; you may need to do this too many times to make it worthwhile.

Goals: Observe some real sequences of random experiments. Develop an intuition about variability. Create sample spaces. Use simulation to estimate probabilities.

Skills:

- **Recognize the feature of randomness.** In statistics, random does not mean haphazard, or without pattern. It means “equal chances”, from trial to trial. While we are unable to predict what will happen on a single toss of a coin, we **are** able to predict what will happen in 1,000 tosses of a coin. This is the hallmark of a random process: uncertainty in a small number of trials, but a predictable pattern in a large number of trials.

- **Resist the urge to jump to conclusions with small samples.** Typically our daily activities do **not** involve large samples of observations. Therefore our ideas of “long run” probability theory are not applicable. You need to develop some intuition about when to believe an observed experiment, and when to doubt the results. We will hone this intuition as we develop our upcoming inference methods. For now, understand that you may be jumping to conclusions by just believing the simulation’s observed value.
- **List simple sample spaces.** Flipping coins and rolling dice are common events to us, and listing the possible outcomes lets us explore probability distributions. We will not delve deeply into probability rules; rather, we are more interested in the ideas of probability and I think the best way to accomplish this is by example.
- **Know the probability rules and how to use them.** We have three rules we use primarily: the complement rule, the addition rule for disjoint events, and the multiplication rule for independent events. The complement rule is used when the calculation of the complement turns out to be simpler than the even itself. For example, the complement of “at least one” is “none”, which is a simpler event to describe. The addition rule for disjoint events is used when we are asking about the chances of one event **or** another occurring. If the two events are disjoint (they have no elements in common) we can find the chance of their union (one event or the other) by adding their individual probabilities. The multiplication rule for independent events is used when we have a question about the intersection of two sets, which can be phrased as a question about two events occurring simultaneously. **Both** sets occurring at once can be phrased as one event **and** the other event occurring, so the multiplication rule is used for “and” statements.
- **Simulation can be used to estimate probabilities, but only for a very large number of trials.** If the number of repetitions of an experiment is large, then the resulting observed frequency of success can be used as an estimate of the true unknown probability of success. However, a “large” enough number of repetitions may be more than we can reasonably perform. For example, for problem 1 today (and using some calculations from Chapter 20), a sample of 100 will give results between 32/100 and 51/100 (0.32 to 0.51) 95 % of the time. That may not be good enough for us. But even with 500 rolls, the range is 187/500 to 230/500 (0.374 to 0.460). Eventually the answers will converge to a useful percentage; the question is how soon that will occur. We will have answers to that question after Chapter 20.

Reading: Chapters 10 and 11.

Day 18

Activity: Continue coins and dice. Introduce Random Variables and Probability Histograms. Central Limit Theorem exploration. **Homework 5 due today.**

To be successful at probability models, we need to understand Random Variables and Probability Distributions. Our goal with a probability model is to predict real-world behavior. If we have chosen a good model and know how it works, we can then make calculations that match observed results. This will allow us to make conclusions theoretically instead of having to rely on simulations, which as we have seen may require huge sample sizes and therefore may be prohibitively costly.

Today we will explore the sample spaces for the sum when rolling two dice and when rolling three dice. (We actually have already seen the two-dice situation.) In addition to the sample space, we will include **random variables**, which are numbers assigned to each element. For example, the sum of the two values on the dice is a random variable. Another example is the larger of the two values on the dice. If we keep track of the possible values (for example 2 to 12 for the sum on two dice) then we can tabulate the **probabilities** in a chart, which we call the **probability distribution**. We will see this idea again and again as we continue towards statistical inference (Unit 4).

We will finish up the problems from Day 17, and also examine Pascal's triangle, which is a way of figuring binomial probabilities (chances on fair coins). Also in our tables, we will include random variables.

Our last topic in Unit 3 is the important **Central Limit Theorem**. In your everyday lives, you have already seen the evidence of this famous theorem, because most of the normal distributions you've encountered are results of this theorem. There are two main conclusions to the Central Limit Theorem that will be important to us. One is that averages are more reliable than individual measurements. This says that the average for a bunch of items in a sample is likely to be closer to the true population average than a single measurement. The other main conclusion is that while the sample average is getting closer to the population average, it is also getting closer in an organized way; in fact, the normal curve describes the sampling distribution of the average quite well for large samples. Many processes in nature are the result of sums, so the normal distribution is quite common in nature.

One of the problems students have with the Central Limit Theorem is recognizing that the problem at hand is about sample averages instead of individual measurements. It is critical for you to have this ability. One way to think of the situation is that **all** problems are really Central Limit Theorem problems, but some of them are just samples of size 1 ($n = 1$). The calculation for the spread of an average ($s_{\bar{x}} = s/\sqrt{n}$) still works when $n = 1$. Using this line of thinking, all the problems we did in Chapter 3 were CLT problems with $n = 1$.

Another problem I have noticed is dealing with sums rather than averages. There are two approaches, and we will look at both in 3) below. The first method is to label the normal curve with both the scale for the average and the scale for the sum. You first find the "average" scale using the standard deviation for the average, as we have been doing. Then the "sum" scale is the "average" scale multiplied by n . The second method is to calculate the standard deviation for the sum directly, using $s_{sum} = s\sqrt{n}$. Notice, for standard deviations, that for the average, you **divide** by the square root of n whereas for the sum you **multiply** by it. That is how I remember which is which: sums are larger than averages. These new standard deviations (for the sum or for the average) are called "standard errors", a new term to keep our various standard deviations straight.

IMPORTANT NOTE: We have many standard deviations we have been referring to. It is important for you to keep them all straight. We have the standard deviation in a **population** of items, σ . We have the standard deviation in a **sample** from a population, s . We have the standard deviation of a **series** of trials, like sums, or averages, and we call these "**standard errors**". We have the standard deviation of a probability model, which is much like the population standard deviation σ . I know all these terms are confusing, using the same word to

describe them, but that is why your practicing problems and having discussions with me are important.

In addition to coins and dice, **MATH PRB rand** on your calculator is another good random mechanism for exploring “sampling distributions”. These examples will give you some different views of sampling distributions. The important idea is that each time an experiment is performed, a potentially different result occurs. How these results vary from sample to sample is what we seek. You are going to produce many samples, and will therefore see how these values vary.

- 1) Sums of two items: Each of you in your group will roll two dice. Record the sum on the dice. Repeat this 100 times, generating 100 sums. Make a histogram and a **QUANTILE** plot of your 100 sums. Compare to the graphs of the other members in your group, particularly noting the shape. Sketch the graphs you made and compare to the theoretical results. A simulation on the calculator uses **seq (sum(randInt(1, 6, 2)), X, 1, 100) -> L1**. The 100 two-dice sums are stored in **L1**. (**seq** is found in the **LIST OPS** menu.)
- 2) Averages of 4 items: Each of you in your group will generate 4 random numbers on your calculator, add them together, average them, and record the result; repeat 100 times. The full command is **seq (rand + rand + rand + rand, X, 1, 100) / 4 -> L2**, which will generate 100 four-number averages and store them in **L2**. Again, make a graph of the distribution. (**seq** is found in the **LIST OPS** menu.)
- 3) Sums of 12 items: Each of you in your group will generate 12 random **normal** numbers on your calculator using the command **MATH PRB randNorm(65, 5, 12)**. Add them together and record the result; repeat 100 times. The full command is: **seq (sum (randNorm(65, 5, 12)), X, 1, 100) -> L3**. Again, make a graph of the distribution. (This example could be describing how the weight of a dozen eggs is distributed.) (**sum** is found in the **LIST MATH** menu.)

For all the lists you generated, calculate the standard deviation and the mean. We will find these two statistics to be immensely important in our upcoming discussions about inference. It turns out that these means and standard deviations can be found through formulas instead of having to actually generate repeated samples. These means depend only on the mean and standard deviation of the original population (the dice or **rand** or **randNorm** in this case) and the number of times the dice were rolled or **rand** was pressed (called the *sample size*, denoted n).

Goals: Understand that variables may have values that are **not** equally likely. Examine histograms or quantile plots to see that averages are less variable than individual measurements. Also, the shape of these curves should get closer to the shape of the normal curve as n increases.



Skills:

- **Understand sampling distributions and how to create simple ones.** We have listed sample spaces of equally likely events, like dice and coins. Events can further be grouped together and assigned values. These new groups of events may not be equally likely, but as long as the rules of probability still hold, we have valid probability distributions. Pascal's Triangle is one such example, though you should realize that it applies only to fair coins. We will work with "unfair coins" (proportions) later, in Chapter 20. Historical note: examining these sampling distributions led to the discovery of the normal curve in the early 1700's. We will copy their work and "discover" the normal curve for ourselves too using dice.
- **Understand the concept of sampling variability.** Results vary from sample to sample. This idea is **sampling variability**. We are very much interested in knowing what the likely values of a statistic are, so we focus our energies on describing the sampling distributions. In today's exercise, you simulated samples, and calculated the variability of your results. In practice, we only do one sample, but calculate the variability with a formula. In practice, we also have the Central Limit Theorem, which lets us use the normal curve in many situations to calculate probabilities.

Reading: Chapter 11. (Skip "Statistical Process Control".)


Day 19

Activity: Practice Central Limit Theorem (CLT) problems. **Quiz 6 today.**

The Central Limit Theorem is special in that it allows us to use the normal curve to find probabilities for distributions that are not normal, but only if we are talking about a large enough sample. Unfortunately, we sometimes forget that we **cannot** use the normal curve on **small** samples. We **should** check to see how close the actual data is to the normal curve, though, because if the original data is reasonably symmetric, with no outliers or extreme skewness, then samples of 20 or 30 may be large enough to use the normal curve approximation. I admit that this determination is a judgment call; there are fancier techniques to determine this that we are skipping over. I don't expect you to always know if the sample size is large enough; I **do** expect that you know about the issue, and that you have looked at the data to at least check for outliers.

We will have examples of non-normal data and normal data to contrast the diverse cases where the CLT applies.

- 1) People staying at a certain convention hotel have a mean weight of 180 pounds with standard deviation 35. The elevator in the hotel can hold 20 people. How much weight will it have to handle in most cases? Do we need to assume weights of people are normally distributed?
- 2) Customers at a large grocery store require on average 3 minutes to check out at the cashier, with standard deviation 2. Because checkout times cannot be negative, they are obviously not normally distributed. Can we calculate the chance that 85 customers will be handled in a four-hour shift? If so, calculate the chance; if not, what else do you need to know?

- 
- 3) Suppose the students at UWO take an average of 15 credits each semester with a standard deviation of 3 credits. What is the chance that the average in a sample of 20 randomly chosen UWO students is between 14.5 and 15.5?
- 4) Suppose the number of hurricanes in a season has mean 6 and standard deviation $\sqrt{6}$. What is the chance that in 30 years the average has been less than 5.5 hurricanes per year? What assumptions must you make?
- 5) The number of boys in a 4-child family can be modeled reasonably well with the binomial distribution, and therefore the normal approximation works well. The mean turns out to be 2 and the standard deviation is 1. If ten such families live on the same street, what is the chance that the total number of boys is 24 or more?

Goals: Use normal curve with the CLT.

Skills:

- **Recognize how to use the CLT to answer probability questions concerning sums and averages.** The CLT says that for large sample sizes, the distribution of the sample average is approximately normal, even though the original data in a problem may not be normal. Sums are simply averages multiplied by the sample size, so any question we can answer about averages, we can also answer about the corresponding sums.
- **For small samples, we can only use the normal curve if the actual distribution of the original data is normally distributed.** It is important to realize when original data is not normal, because there is a tendency to use the CLT even for small sample sizes, and this is inappropriate. When the CLT **does** apply, though, we are armed with a valuable tool that allows us to estimate probabilities concerning averages. A particular example is when the data is a count that **must** be an integer, and there are only a few possible values, such as the number of kids in a family. Here the normal curve wouldn't help you calculate chances of a **single** family having 3 boys. However, we could calculate quite accurately the number of boys in **100** such families.

Reading: Chapters 8 through 11.

Day 20

Activity: Presentation 2. Sampling. Unit 3 Review. **Homework 6 due today.**

Sample 20 students from UWO. For each student, record the number of credits they are taking this semester, what year they are in school, and whether or not they are graduating this semester. Summarize each variable for us, using both graphical and numerical summaries, as appropriate. At the very least, report to us the average credit load and the standard deviation of credit loads. Make your sample as representative as you can, but you **must** have some sort of a **probability sample** to get **full credit**. Discuss the biases your sample has and what you did to try to avoid bias.



Day 21

Activity: Exam 3.

This third exam is on sampling, experiments, and probability, including sampling distributions. Some of the questions may be multiple choice or True/False. Others may require you to show your worked out solution. Section reviews are an excellent source for studying for the exams. Don't forget to review your class notes and these on-line notes.

Day 22

Activity: Guess m&m's percentage. Changing confidence levels and sample sizes.

Unit 4 is about statistical inference, where we collect a sample of data and try to make statements about the entire population the sample came from. The classic situation is where we have an actual population of individuals from which we have chosen a sample, for example pre-election polls. A more prevalent use of statistical inference though is in the situation where we extend our results from the current individuals to the entire population of individuals, even though we don't have a random sample. An example of this kind of inference is when we use subjects in an experiment and then make claims about the whole population. If it turns out our subjects were not representative in some important way, then our calculations will be ridiculous. For example, does it matter to us that saccharin causes cancer in laboratory rats? Are we claiming that rat physiology is like human physiology? I do not know the answer to that question, but it is at the heart of the inference.

There are two main tools of statistical inference, confidence intervals and hypothesis testing. The first one we will examine is **confidence intervals**, where we "guess" the population value using the sample information. We **know** our guess isn't right though; what is the probability that our example is **perfectly** representative! So, to convey our belief that we are **close** but not **exact**, we report an interval of guesses. Ideally this interval would be narrow; then we would be saying our guess is very close to the truth. In practice though it is hard to get narrow intervals, because as we will see the narrow intervals are associated with large samples and large samples are expensive.

What fraction of m&m's are blue or green? Is it 25 %? 33 %? 50 %? We take samples today to find out. Technically the samples we take today will not be SRS's; we are not labeling the m&m's individually and then randomly selecting them by number. However, if the population is sufficiently mixed, I believe we can use our results and have reasonable faith that they make sense. (Note that this situation differs from samples of convenience such as asking students in Reeve about campus life.)

Each of you will sample from my jar of m&m's, and you will all calculate your own confidence interval. Of course, not everyone will be correct, and in fact, some of us will have "lousy" samples. But that is the point of the confidence coefficient, as we will see when we jointly interpret our results.

It has been my experience that confidence intervals are easier to understand if we talk about sample proportions instead of sample averages. Thus I will use techniques from Chapter 8. Each of you will have a different sample size and a different number of successes. In this case the sample size, n , is the total number of m&m's you have selected, and the number of



successes, x , is the total number of blue or green m&m's in your sample. Your guess is simply the ratio x/n , or the **sample proportion**. We call this estimate p -hat or \hat{p} . Use **STAT TEST 1-PropZInt** with 70 % confidence for your interval here today. However, in practice you will rarely see anyone using less than 90 % confidence.

When you have calculated your confidence interval, record your result on the board for all to see. We will jointly inspect these confidence intervals and observe just how many are "correct" and how many are "incorrect". The percentage of correct intervals *should* match our chosen level of confidence. This is in fact what is meant by confidence.

We want narrow confidence intervals. What effect do the confidence level and the sample size have on interval width? Remember, what we are using this technique for is to guess the true parameter, in these cases population percentage (the first chart) or population average (the second chart). We have already used and interpreted the command for the first chart. For the second chart, we will use a new command today, the **ZInterval**. We interpret this command the same as the other: it gives a range of guesses, and our confidence is expressed as a percentage telling us how often the technique produces correct answers.

Today we will also explore how changing confidence levels and sample sizes influence CI's. Fill in enough of the cells in the following table so that you can see the trends, recording the **confidence interval width** in the body of the table. Use **STAT TEST 1-PropZInt** but in each case make x equal to 50 % of n . **IMPORTANT NOTE:** Today's exercise is not one you would do in practice. Our only goal is to see the effects of sample size and confidence on the **width** of the intervals.

Confidence Level =====> Sample Size	70 %	90 %	95 %	99 %	99.9 %
10					
20					
50					
100					
1000					

We will try to make sense of this chart, keeping in mind the meaning of confidence level, and the desire to have narrow intervals.

Now repeat the above table using **STAT TEST ZInterval**, with $\sigma = 15$ and $\bar{x} = 100$.

Confidence Level =====> Sample Size	70 %	90 %	95 %	99 %	99.9 %



10					
20					
50					
100					
1000					

Goals: Introduce statistical inference - Guessing the parameter. Construct and interpret a confidence interval. See how the TI-83 calculates our CI's. Interpret the effect of differing confidence coefficients and sample sizes.

Skills:

- **Understand how to interpret confidence intervals.** The calculation of a confidence interval is quite mechanical. In fact, as we have seen, our calculators do all the work for us. Our job is then not so much to **calculate** confidence intervals as it is to be able to understand **when** one should be used and how best to **interpret** one.
- **Know what the confidence level measures.** We would like to always make correct statements, but in light of sampling variability we know this is impossible. As a compromise, we use methods that work **most** of the time. The proportion of times our methods work is expressed as a **confidence coefficient**. Thus, a 95 % confidence interval method produces correct statements 95 % of the time. (By “correct statement” we mean one where the true unknown value is contained in our interval, that is, it is between our two numbers.)
- **Understand the factors that make confidence intervals believable guesses for the parameter.** The two chief factors that make our confidence intervals believable are the sample size and the confidence coefficient. The key result is larger confidence makes wider intervals, and larger sample size makes narrower intervals.
- **Know the details of the Z Interval.** When we know the population standard deviation, σ , our method for guessing the true value of the mean, μ , is to use a z confidence interval. This technique is unrealistic in that you must know the true population standard deviation. In practice, we will estimate this value with the sample standard deviation, s , but a different technique is appropriate (See Day 24).

Reading: Chapter 15 and Part of Chapter 16 (“Type I and Type II Errors”).

Day 23

Activity: Argument by contradiction and the Scientific method. Type I and Type II error diagram. Courtroom terminology. Practice problems on hypothesis testing. Introduce z -test as an example.

We have been introduced to the first of our two statistical inference techniques. Today we will begin our introduction to the other technique: **statistical hypothesis testing**. In this technique, we again collect a sample of data, but this time instead of guessing the value of the parameter, we check to see if the sample is close to some pre-specified value for the parameter. This pre-specified value is the **statistical hypothesis**, a statement of equality. The result of our test will be either “I believe the hypothesis is reasonable” or “the hypothesis looks wrong”. Just as in confidence intervals, we will give a percentage that can be interpreted as the success rate of our method.

Some terminology:

Null hypothesis. The null hypothesis is a statement about a parameter (a population fact). The null hypothesis is **always** an equality or a single claim (like two variables are independent). We assume the null hypothesis is true in our following calculations, so it is important that the null be a specific value or fact that can be assumed.

Alternative hypothesis. The alternative hypothesis is a statement that we will believe if the null hypothesis is rejected. The alternative does not have to be the complement of the null hypothesis. It just has to be some other statement. It **can** be an inequality, and usually is.

One- and Two-Tailed Tests. A one-tailed test is one where the alternative hypothesis is in only one direction, like “the mean is **less** than 10”. A two-tailed test is one where the alternative hypothesis is indiscriminate about direction, like “the mean is **not equal** to 10”. When a researcher has an agenda in mind, he will usually choose a one-tailed test. When a researcher is unsure of the situation, a two-tailed test is appropriate.

Rejection rule. To decide between two competing hypotheses, we create a rejection rule. It’s usually as simple as “Reject the null hypothesis if the sample mean is greater than 10. Otherwise fail to reject.” We always want to phrase our answer as “reject the null hypothesis” or “fail to reject the null hypothesis”. We **never** want to say “accept the null hypothesis”. The reasoning is this: rejecting the null hypothesis means the data have contradicted the assumptions we’ve made (assuming the null hypothesis was correct); failing to reject the null hypothesis doesn’t mean we’ve proven the null hypothesis is true, but rather that we haven’t seen anything to doubt the claim yet. It **could** be the case that we just haven’t taken a large enough sample yet.

Type I Error. When we reject the null hypothesis when it is in fact true, we have made a Type I error. We have made a conscious decision to treat this error as a more important error, so we construct our rejection rule to make this error rare.

Type II Error. When we fail to reject the null hypothesis, and in fact the alternative hypothesis is the true one, we have made a Type II error. Because we construct our rejection rule to control the Type I error rate, the Type II error rate is not really under our control; it is more a function of the particular test we have chosen. The one aspect we **can** control is the sample size. Generally, larger samples make the chance of making a Type II error smaller.

Significance level, or size of the test. The probability of making a Type I error is the significance level. We also call it the size of the test, and we use the symbol α to represent it. Because we want the Type I error to be rare, we usually will set α to be a small number, like

0.05 or 0.01 or even smaller. Clearly smaller is better, but the drawback is that the smaller α is, the larger the Type II error becomes. Remember, we often express this error rate as a percentage, so 0.05 is also 5 %.

P-value. There are two definitions for the P-value. Definition 1: The P-value is the alpha level that will cause us to **just** reject our observed data. Definition 2: The P-value is the chance of seeing data as extreme or more extreme than the data actually observed. Using either definition, we calculate the P-value as an area under a tail in a distribution. Caution: the P-value calculation will depend on whether we have a one- or a two-tailed test.

Power. The power of a test is the probability of rejecting the null hypothesis when the alternative hypothesis is true. We are calculating the chance of making a correct decision. Because the alternative hypothesis is usually not an equality statement, it is more appropriate to say that power is a **function** rather than just a single value.

We will examine these ideas using the z -test. The TI-83 command is **STAT TEST ZTest**. The command gives you a menu of items to input. It assumes your null hypothesis is a statement about a mean μ . You must tell the assumed null value, μ_0 , and the alternative claim, either the two-sided choice, or one of the one-sided choices. You also need to tell the calculator how your information has been stored, either as a list of raw **DATA** or as summary **STATS**. If you choose **CALCULATE** the machine will simply display the test statistic and the P-value. If you choose **DRAW**, the calculator will graph the P-value calculation for you. You should experiment to see which way you prefer.

One difficulty students have with hypothesis testing is setting up the hypotheses. Today we will work some examples so you become familiar with the process and the interpretations. Before we begin the examples, though, I want to mention several ideas.

Researchers generally have agendas in their research. They are after all trying to show their ideas are true. But we require a statement of equality in our statistical inference, so the strategy generally goes this way: assume the opposite of what you are trying to prove. When the data don't look right under that assumption, you have "proven" the equality is untrue, or at least it is very unlikely. You can then believe your original idea was the right one. For example, a researcher thinks adding fertilizer to plants will increase growth. She will assume for the sake of argument that the fertilizer doesn't work. Her null hypothesis is that the mean growth will be exactly zero. After she collects some data using fertilizer on plants, she will be able to say the data agree with the "no-growth" claim or that they disagree with the claim, in which case she will happily assume that fertilizer instead helps plants grow.

The notation we use for hypotheses is a shorthand. Instead of saying "The null hypothesis is that the population mean is 20" we say " $H_0: \mu = 20$ ". Because it is a sentence, you need a subject and a verb. The subject is H_0 . The verb is the colon. $\mu = 20$ is the clause. It would be incorrect to write just $H_0: = 0$. That would be like saying "The null hypothesis is that equals zero."

We have two interpretations of P-values. To perform a statistical hypothesis test using P-values, we compare the calculated P-value to the pre-chosen significance level, often 5 %. If the P-value is small (at or below 5 %) we reject the claim, stating that the data contradict the

null hypothesis and we then conclude the alternate hypothesis is more believable. If the P-value is large (above 5 %) we do not have any reason to think the null hypothesis is incorrect. The data agree with the claim. We “fail to reject” the null hypothesis and conclude that it could indeed be true. Notice the weak language: “could” instead of “is definitely”. When we reject the null hypothesis, we are saying with high confidence that the null isn’t right. When we **fail** to reject the null hypothesis we are saying it is plausible, but we are not saying it is for sure true. We will explore this particular phenomenon more on Day 24.

The second interpretation for a P-value is that it is the chance of seeing more extreme results than those observed. It is our measure of how unusual our current results are. If our P-value is 0.25, or 25 %, we say “there is a 25 % chance of seeing data like this or worse when the null hypothesis really is true”. That doesn’t seem so unusual then. Conversely, if the P-value is 0.03, or 3 %, we say there is only a 3 % chance of seeing data this extreme or worse. That seems too unusual to us, and we therefore conclude that the claim on which this calculation was based is wrong. Specifically, we assume the alternate hypothesis to be the correct sentence.

Work on the following problems. Compare answers with your neighbor or those in your group. We will summarize before the end of class.

- 1) An educational researcher wants to show that the students in his school district score better than the national average (which is 65) on a standardized test. He collects a sample of 20 test scores, and finds the sample average is 66. If σ is 3, what is his conclusion? Set up the hypothesis and perform the test. Use $\alpha = 0.05$. What assumptions is the researcher making?
- 2) A downhill skier makes 5 timed runs on a course with these times in seconds: 52, 56, 55, 58, and 51. In the past, this skier’s times have had a standard deviation of 2, so you can assume σ is 2. To qualify for a local competition, a skier must beat 55 seconds. Is there evidence that our skier’s true ability is enough to qualify for the local competition? Set up the hypothesis and perform the test. Use $\alpha = 0.05$. What assumptions must we make about the skier’s times?
- 3) You work for a company that makes nuts for bolts. You have to calibrate a machine to put 100 nuts in a bag. (The label on the bag says “contains 100 nuts”.) Suppose you run the machine 10 times, and find the average is 98. Assume the standard deviation is 1. Is the machine working or do you have to adjust it?

Goals: Introduce statistical inference - Hypothesis testing. Practice contradiction reasoning, the basis of the scientific method.

Skills:

- **Recognize the two types of errors we make.** If we decide to reject a null hypothesis, we might be making a Type I error. If we fail to reject the null hypothesis, we might be making a Type II error. If it turns out that the null hypothesis is true, and we reject it because our data looked weird, then we have made a Type I error. Statisticians have agreed to control this type of error at a specific percentage, usually 5 %. On the other hand, if the alternative hypothesis

is true, and we **fail** to reject the null hypothesis, we have also made a mistake. We generally do not control this second type of error; the sample size is the determining factor.

- **Understand why one error is considered a more serious error.** Because we control the frequency of a Type I error, we feel confident that when we reject the null hypothesis, we have made the right decision. This is how the scientific method works; researchers usually set up an experiment so that the conclusion they would like to make is the alternative hypothesis. Then if the null hypothesis (usually the opposite of what they are trying to show) is rejected, there is some confidence in the conclusion. On the other hand, if we **fail** to reject the null hypothesis, the most useful conclusion is that we didn't have a large enough sample size to detect a real difference. We aren't really saying we are confident the null hypothesis is a true statement; rather we are saying it **could** be true. Because we cannot control the frequency of this error, it is a less confident statement.
- **Become familiar with “argument by contradiction”.** When researchers are trying to “prove” a treatment is better or that their hypothesized mean is the right one, they will usually choose to assume the opposite as the null hypothesis. For election polls, they assume the candidate has 50 % of the vote, and hope to show that is an incorrect statement. For showing that a local population differs from, say, a national population, they will typically assume the national average applies to the local population, again with the hope of rejecting that assumption. In all cases, we formulate the hypotheses **before** collecting data; therefore, you will never see a sample average in either a null or alternative hypothesis.
- **Understand why we reject the null hypothesis for small P-values.** The P-value is the probability of seeing a sample result “worse” than the one we actually saw. In this sense, “worse” means even more evidence against the null hypothesis; or even more evidence favoring the alternative hypothesis. If this probability is small, it means either we have observed a rare event, or that we have made an incorrect assumption, namely the value in the null hypothesis. Statisticians and practitioners have agreed that 5 % is a reasonable cutoff between a result that contradicts the null hypothesis and a result that could be argued to be in agreement with the null hypothesis. Thus, we reject our claim only when the P-value is a small enough number.

Reading: Chapter 15. Chapter 16. (Skip “Power”.) Part of Chapter 17. Chapter 18.

Day 24

Activity: Testing Simulation. Gosset simulation. **Quiz 7 today.**

Today I want to give you practice performing hypothesis tests, and interpreting results. This is a day for **understanding**, not something you would do in practice. Work in pairs; one of you will generate some data for the other person. The other person will conduct a test and decide whether the first person chose 20 or not for the mean. Then you will repeat the procedure a number of times. Each time, the second person will not know what value the first person chose, but the second person will always answer with “You could have used 20” (Failing to reject the null hypothesis) or “I believe you did **not** use 20” (Rejecting the null hypothesis).

In this experiment, you will work in pairs and generate data for your partner to analyze. Your partner will come up with a conclusion (either “reject the null hypothesis” or “fail to reject the null hypothesis”) and you will let them know if they made the right decision or not. Keep

careful track of the success rates. Record the sample averages and the P-values so we can check our work.

For each of these simulations, let the null hypothesis mean be 20, $n = 10$, and $\sigma = 5$. You will let μ change for each replication.

- 1) **Without your partner knowing**, choose a value for μ from this list: 16, 18, 20, 22, or 24. Then use your calculator and generate 10 observations. Use **MATH PRB randNorm(M, 5, 10) → L1** where **M** is the value of μ you chose for this replication. Clear the screen (so your partner can't see what you did) and give them the calculator. They will perform a hypothesis test using the 0.05 significance level and tell you their decision. Remember, **each** test uses the same null hypothesis: "The mean is 20." vs. the alternative "The mean is different from 20". The decision will be either "Reject the null hypothesis" or "Fail to reject the null hypothesis". Finally, compare the decision your partner made to the true value of μ that you chose.
- 2) Repeat step 1 until you have each done at least 10 hypothesis tests; it is not necessary to have each value of μ exactly twice, but try to do each one at least once. Do $\mu = 20$ at least twice each. (We need more cases for 20 because we're using a small significance level.)
- 3) Keep track of the results you got (number of successful decisions and number of unsuccessful decisions) and report them to me so we can all see the combined results.

One last point today is that these techniques all require random samples, and typically either normally distributed data or large sample sizes (to invoke the CLT). In particular be cautious of situations where the entire population has been measured. In these cases, asking the statistical question "Is the difference due to chance?" makes no sense. Because the entire population was measured, there is no chance error involved; there is no sampling error due to items not being chosen for the sample. This situation occurs often in science. For example, a psychologist may conduct an experiment on a litter of rats. This litter is of course not a random sample of all rats; rather it is the rats available to that researcher at that time. What is typically assumed then is that this sample of rats is **like** a random sample. Whether this assumption is true or not cannot be checked. Most researchers simply proceed. I think it is important that you at least acknowledge this "fudging". In practice, though, there will be little you will do about it.

Assuming we know the population standard deviation is unrealistic. We will examine a little history into the problem discovered 110 years ago with small samples. Over a century ago, William Sealy Gosset noticed that for small samples, his confidence intervals were not as correct as predicted. For example, using samples of two measurements he was supposed to capture the parameter in his intervals 95 % of the time and found he was only accomplishing this 70 % of the time. The problem was he did not know the true population standard deviation. At the time, the standard practice was to just use the sample standard deviation as if it were the unknown population standard deviation. This turns out to be the wrong way to do things, and Gosset guessed what the true distribution should be. A few years later Sir R. A. Fisher proved that Gosset's guess was correct, and the statistical community accepted the t

distribution, as it came to be known. Gosset was unable to publish his results under his own name (to protect trade secrets), so he used the pseudonym “A. Student”. You will therefore sometimes see the t distribution referred to as “Student’s t distribution”.

In practice, we seldom know the true population standard deviation, so we need to select the t procedures instead of the z procedures in our TI-83. Our interpretations of confidence intervals and hypothesis tests are exactly the same.

Take samples of size 5 from a normal distribution. Use s instead of σ in the standard 95 % confidence z -interval. Repeat 100 times to see if the true coverage is 95 %. (My program **GOSSET** accomplishes this.) We will pool our results to see how close we are to 95 %.

While we will use the TI-83 to calculate confidence intervals, it will be helpful to know the formulas in addition. All of the popular confidence intervals are based on adding and subtracting a **margin of error** to a point estimate. This estimate is almost always an average, although in the case of proportions it is not immediately clear that it is an average.

Goals: Interpret significance level. Observe the effects of different values of the population mean. Recognize limitations to inference. Realize the potential abuses of hypothesis tests. Introduce t -test. Understand how the z -test is inappropriate in most small sample situations.

Skills:

- **Interpret significance level.** Our value for rejecting, usually 0.05, is the percentage of the time that we falsely reject a true null hypothesis. It does not measure whether we had a random sample; it does not measure whether we have bias in our sample. It **only** measures whether random data could look like the observed data.
- **Understand how the chance of rejecting the null hypothesis changes when the population mean is different than the hypothesized value.** When the population mean is **not** the hypothesized value, we expect to reject the null hypothesis more often. This is reasonable, because rejecting a false null hypothesis is a correct decision. Likewise, when the null hypothesis is in fact true, we hope to seldom decide to reject. If we have generated enough replications in class, we should see a power curve emerge that tells us how effective our test is for various values of the population mean.
- **Know the limitations to confidence intervals and hypothesis tests.** Chapter 16 has examples of when our inference techniques are inappropriate. The main points to watch for are non-random samples, misinterpreting what “rejecting the null hypothesis” means, and misunderstanding what error the margin of error is measuring. Be sure to read the examples in Chapter 16 carefully as I will not go over them in detail in class.
- **Know why using the t -test or the t -interval when σ is unknown is appropriate.** When we use s instead of σ and do **not** use the correct t distribution, we find that our confidence intervals are too narrow, and our hypothesis tests reject H_0 too often. In effect, we must pay a “penalty” for not knowing the true population standard deviation. That penalty from the t distribution is that P-values are larger and confidence intervals are wider.
- **Realize that the larger the sample size is, the less serious the problem is.** When we have larger sample sizes, say 15 to 20, we notice that the simulated success rates are much closer to

the theoretical. Thus the issue of t vs. z is a moot point for large samples. I still recommend using the t -procedures any time you do not know σ .

Reading: Chapters 18 and 19.

Day 25

Activity: Matched Pairs vs. Two-Sample. Two-Sample Problems. **Homework 7 due today.**

Matched Pairs problems are really one sample data sets disguised as two sample data sets because two measurements on the same subject are taken. Sometimes “subject” is a person; other times it is less recognizable, such as a year. The key issue is that two measurements have been taken that are related to one another. One quick way to tell if you have a two-sample problem is whether the lists are of different lengths. Obviously if the lists are of different lengths, they are not paired together. Naturally the tricky situation is when the lists are of the same length, which occurs often when researchers assign the same number of subjects to each of treatment and control groups. Another way to confirm that you have matched pairs is to imagine rearranging one of the lists. Would we still be able to analyze the data or would the rearranging make the results meaningless? Does the first item in each list “belong” to each other?

Once you realize that a sample is a matched pairs data set and that the **difference** in the two measurements is the important fact, the analysis proceeds just like one-sample problems, but you use the list of differences. In this respect, there is nothing new about the matched pairs situation.

I will generate some data for you in class where there is a small difference in means between the two lists, but a real difference, in that for each case the difference is about the same. However, by including variability within each list, we will have the situation where analyzing the data incorrectly as two-sample data will yield insignificant results. This is the danger; failing to reject a significant difference, and thereby perhaps wasting your research resources.

Two sample problems involve comparing whether the means from two populations are the same. Our null hypothesis is that the means are equal, and the alternate is that they differ. In the case of the one-tailed test, we claim that one population’s mean is smaller than the other’s. The test statistic is a t -statistic. We will skip over the details of the formula in favor of the interpretation of computer output, in our case the TI-83. However, the basic strategy is the same: formulate a null hypothesis (in this case, we **usually** assume the means are equal, or that their difference equals zero), calculate a test statistic, and make a decision about which hypothesis to believe. The interpretation of the P-value is the same as before. The interpretation of failing to reject is the same. The **only** difference is the calculation of the test statistic.

We will look at several of the two-sample problems from the text today.

Goals: Recognize when matched-pairs applies. Complete two-sample t -test.



Skills:

- **Detect situations where the matched pairs t -test is appropriate.** The nature of the matched pairs is that each value of one of the variables is associated with a value of the other variable. The most common example is a repeated measurement on a single individual, like a pre-test and a post-test. Other situations are natural pairs, like a married couple, or twins. In all cases, the variable we are **really** interested in is the difference in the two scores or measurements. This single difference then makes the matched pairs test a one-variable t -test.
- **Know the typical null hypothesis for two-sample hypothesis tests.** The typical null hypothesis for two-sample problems, both matched and independent samples, is that of “no difference”. For the matched pairs, we say $H_0: \mu = 0$, and for the two independent samples we say $H_0: \mu_1 = \mu_2$. As usual, the null hypothesis is an equality statement, and the alternative is the statement the researcher typically wants to end up concluding. In both two-sample procedures, we interpret confidence intervals as ranges for the **difference** in means, and hypothesis tests as whether the observed difference in means is far from zero.

Reading: Chapters 20 and 21.

Day 26

Activity: Proportions. **Quiz 8 today.**

We revisit the m&m type problems from Day 22.

Proportions: What are the true batting averages of baseball players? Do we believe results from a few games? A season? A career? We can use the binomial distribution as a model for getting hits in baseball, and examine some data to estimate the true hitting ability of some players.

For a typical baseball player, we can look at confidence intervals for the true percentage of hits he gets. On the calculator, the command for the one-sample problem is **STAT TEST 1-PropZInt**.

Technical note: the Plus 4 Method will give more appropriate confidence intervals. As this method is extraordinarily easy to use (add 2 to the numerator, and 4 to the denominator), I recommend you always use it when constructing confidence intervals for proportions. For two-sample problems, divide the 2 and 4 evenly between the two samples; that is, add 1 to **each** numerator and 2 to **each** denominator. Furthermore, the Plus 4 Method seems to work even for very small sample sizes, which is not the advice generally given by textbooks for the large sample approximation. The Plus 4 Method advises that samples as small as 10 will have fairly reliable results; the large sample theory requires 5 to 10 cases in **each** of the failure and success group. Thus, at **least** 20 cases are required, and that is only when p is close to 50 %.

Goals: Introduce proportions.

Skills:

- **Detect situations where proportions z-test is correct.** We have several conditions that are necessary for using proportions. We must have situations where only **two** outcomes are

possible, such as yes/no, success/failure, live/die, Rep/Dem, etc. We must have independence between trials, which is typically simple to justify; each successive measurement has nothing to do with the previous one. We must have a constant probability of success from trial to trial. We call this value p . And finally we must have a fixed number of trials in mind beforehand; in contrast, some experiments continue **until** a certain number of successes have occurred.

- **Know the conditions when the normal approximation is appropriate.** In order to use the normal approximation for proportions, we must have a large enough sample size. A typical rule of thumb is to make sure there are at least 10 successes and at least 10 failures in the sample. For example, in a sample of voters, there must be at least 10 Republicans and at least 10 Democrats, if we are estimating the proportion or percentage of Democrats in our population. (Recall the m&m's example: when you each had fewer than 10 blue or green m&m's, I made you take more until you had at least 10.)
- **Know the Plus 4 Method.** A recent study (1998) from statistical research suggested that the typical normal theory failed in certain unpredictable situations. Those researchers recommend a convenient fix: pretend there are 4 additional observations, 2 successes and 2 failures. By adding these pretend cases to our real cases, the resulting confidence intervals almost magically capture the true parameter the stated percentage of the time. Because this fix is so simple and is more accurate, it is the recommended approach in all **proportions confidence interval** problems. Hypothesis testing procedures remain unchanged.

Reading: Chapters 14 to 22.

Day 27

Activity: Presentation 4. Statistical Inference. Unit 4 Review. **Homework 8 due today.**

Make a claim (a statistical hypothesis) and perform a test of your claim. Gather appropriate data and choose an appropriate technique to test your claim. (I do not want you to just do a problem from the book; be creative, but see me if you are stuck.) Discuss and justify any assumptions you made. Explain why your test is the appropriate technique.

Day 28

Activity: Exam 4.

This last exam covers the z - and t - tests and intervals in Chapters 14 through 22. Some of the questions may be multiple choice or True/False. Others may require you to show your worked out solution. Section reviews are an excellent source for studying for the exams. Don't forget to review your class notes and these on-line notes.

Arizona Temps Dataset:

Date	Flagstaff					Phoenix				
	Temperature (°F)		Humidity (%)		Wind (mph)	Temperature (°F)		Humidity (%)		Wind (mph)
	high	low	high	low	avg	high	low	high	low	avg
5/1/06	66	35	89	29	3	99	69	31	8	7
5/2/06	72	35	85	11	7	96	69	25	7	8
5/3/06	66	36	49	16	10	95	69	25	7	10

5/4/06	65	31	41	15	13	93	65	24	8	10
5/5/06	59	28	82	26	6	87	65	26	11	7
5/6/06	65	27	92	16	6	91	67	34	12	6
5/7/06	69	34	64	14	4	94	67	37	9	7
5/8/06	71	36	59	18	11	95	69	29	7	7
5/9/06	70	34	62	20	6	94	69	31	10	6
5/10/06	72	32	82	12	6	99	70	31	8	5
5/11/06	74	31	41	10	6	100	73	31	7	6
5/12/06	79	38	40	9	7	100	71	27	6	6
5/13/06	78	44	40	11	6	101	73	20	7	8
5/14/06	80	40	43	14	3	105	74	31	5	7
5/15/06	79	46	50	20	8	101	74	24	7	7
5/16/06	75	44	83	27	6	104	76	31	12	10
5/17/06	76	37	92	17	4	104	78	38	11	7
5/18/06	78	45	60	16	6	104	77	36	8	9
5/19/06	79	44	65	15	8	105	78	28	8	7
5/20/06	78	45	58	14	10	105	78	28	7	6
5/21/06	77	39	44	18	8	105	78	23	7	10
5/22/06	63	32	70	24	14	84	73	36	9	12
5/23/06	72	27	85	17	5	92	66	40	11	5
5/24/06	81	32	61	9	6	102	71	31	7	5
5/25/06	79	43	40	9	8	103	74	25	6	8
5/26/06	74	39	43	11	12	103	76	22	5	9
5/27/06	68	50	46	14	23	96	72	19	10	12
5/28/06	63	33	65	6	13	88	70	41	4	10
5/29/06	67	27	33	7	7	92	65	26	5	6
5/30/06	74	29	29	7	4	97	68	26	4	5
5/31/06	79	30	36	6	4	102	70	23	4	5
6/1/06	81	33	41	5	7	108	73	19	3	5
6/2/06	84	39	28	8	6	110	80	17	6	7
6/3/06	84	46	37	10	6	112	83	27	6	6
6/4/06	86	53	30	9	7	112	83	23	6	9
6/5/06	87	53	35	8	7	108	84	27	9	11
6/6/06	86	53	34	9	7	105	86	26	12	10
6/7/06	75	19	62	34	9	106	82	32	14	10
6/8/06	75	50	87	28	5	106	81	49	12	9
6/9/06	77	50	80	12	12	105	81	28	8	9
6/10/06	78	42	58	9	10	104	78	25	6	7
6/11/06	80	40	43	8	7	107	76	25	2	7
6/12/06	80	35	33	5	8	107	74	21	3	9
6/13/06	83	36	32	14	10	111	79	17	6	7
6/14/06	82	57	36	5	17	106	84	20	4	9
6/15/06	75	43	30	11	10	101	76	25	7	10
6/16/06	77	35	64	13	6	102	77	21	7	7
6/17/06	84	37	48	8	5	106	77	23	6	6
6/18/06	87	42	37	7	7	109	80	26	6	6



Populations for the 50 states, DC, and the USA, by decade. (in thousands)

	AL	AK	AZ	AR	CA	CO	CT	DE	DC	FL	GA	HI	ID	IL	IN	IA	KS	KY
1790							238	59			83							74
1800	1						251	64	8		163				6			221
1810	9			1			262	73	16		252			12	25			407
1820	128			14			275	73	23		341			55	147			564
1830	310			30			298	77	30	35	517			157	343			688
1840	591			98			310	78	34	54	691			476	686	43		780
1850	772			210	93		371	92	52	87	906			851	988	192		982
1860	964			435	380	34	460	112	75	140	1057			1712	1350	675	107	1156
1870	997		10	484	560	40	537	125	132	188	1184		15	2540	1680	1194	364	1321
1880	1263	33	40	803	865	194	623	147	178	269	1542		33	3078	1978	1625	996	1649
1890	1513	32	88	1128	1213	413	746	168	230	391	1837		89	3826	2192	1912	1428	1859
1900	1829	64	123	1312	1485	540	908	185	279	529	2216	154	162	4822	2516	2232	1470	2147
1910	2138	64	204	1574	2378	799	1115	202	331	753	2609	192	326	5639	2701	2225	1691	2290
1920	2348	55	334	1752	3427	940	1381	223	438	968	2896	256	432	6485	2930	2404	1769	2417
1930	2646	59	436	1854	5677	1036	1607	238	487	1468	2909	368	445	7631	3239	2471	1881	2615
1940	2833	73	499	1949	6907	1123	1709	267	663	1897	3124	423	525	7897	3428	2538	1801	2846
1950	3062	129	750	1910	10586	1325	2007	318	802	2771	3445	500	589	8712	3934	2621	1905	2945
1960	3267	226	1302	1786	15717	1754	2535	446	764	4952	3943	633	667	10081	4662	2758	2179	3038
1970	3444	303	1775	1923	19971	2210	3032	548	757	6791	4588	770	713	11110	5195	2825	2249	3221
1980	3894	402	2717	2286	23668	2890	3108	594	638	9747	5463	965	944	11427	5490	2914	2364	3660
1990	4040	550	3665	2351	29760	3294	3287	666	607	12938	6478	1108	1007	11430	5544	2777	2478	3685
2000	4447	627	5131	2673	33872	4301	3406	784	572	15982	8186	1212	1294	12419	6080	2926	2688	4042
2010	4780	710	6392	2916	37254	5029	3574	898	602	18801	9688	1360	1568	12831	6484	3046	2853	4339

	LA	ME	MD	MA	MI	MN	MS	MO	MT	NE	NV	NH	NJ	NM	NY	NC	ND	OH
1790		97	320	379								142	184		340	394		
1800		152	342	423			8					184	211		589	478		45
1810	77	229	381	472	5		31	20				214	246		959	556		231
1820	153	298	407	523	9		75	67				244	278		1373	639		581
1830	216	399	447	610	32		137	140				269	321		1919	736		938
1840	352	502	470	738	212		376	384				285	373		2429	753		1519
1850	518	583	583	995	398	6	607	682				318	490	62	3097	869		1980
1860	708	628	687	1231	749	172	791	1182		29	7	326	672	94	3881	993		2340
1870	727	627	781	1457	1184	440	828	1721	21	123	42	318	906	92	4383	1071	2	2665
1880	940	649	935	1783	1637	781	1132	2168	39	452	62	347	1131	120	5083	1400	37	3198
1890	1119	661	1042	2239	2094	1310	1290	2679	143	1063	47	377	1445	160	6003	1618	191	3672
1900	1382	694	1188	2805	2421	1751	1551	3107	243	1066	42	412	1884	195	7269	1894	319	4158
1910	1656	742	1295	3366	2810	2076	1797	3293	376	1192	82	431	2537	327	9114	2206	577	4767
1920	1799	768	1450	3852	3668	2387	1791	3404	549	1296	77	443	3156	360	10385	2559	647	5759
1930	2102	797	1632	4250	4842	2564	2010	3629	538	1378	91	465	4041	423	12588	3170	681	6647
1940	2364	847	1821	4317	5256	2792	2184	3785	559	1316	110	492	4160	532	13479	3572	642	6908
1950	2684	914	2343	4691	6372	2982	2179	3955	591	1326	160	533	4835	681	14830	4062	620	7947
1960	3257	969	3101	5149	7823	3414	2178	4320	675	1411	285	607	6067	951	16782	4556	632	9706
1970	3645	994	3924	5689	8882	3806	2217	4678	694	1485	489	738	7171	1017	18241	5084	618	10657
1980	4206	1125	4217	5737	9262	4076	2521	4917	787	1570	801	921	7365	1303	17558	5880	653	10798
1990	4220	1228	4781	6016	9295	4375	2573	5117	799	1578	1202	1109	7730	1515	17990	6629	639	10847
2000	4469	1275	5296	6349	9938	4919	2845	5595	902	1711	1998	1236	8414	1819	18976	8049	642	11353
2010	4533	1328	5774	6548	9884	5304	2967	5989	989	1826	2701	1316	8792	2059	19378	9535	673	11537

	OK	OR	PA	RI	SC	SD	TN	TX	UT	VT	VA	WA	WV	WI	WY	USA
1790			434	69	249		36			85	692		56			3929
1800			602	69	346		106			154	808		79			5308
1810			810	77	415		262			218	878		105			7240
1820			1049	83	503		423			236	938		137			9638
1830			1348	97	581		682			281	1044		177			12866
1840			1724	109	594		829			292	1025		225	31		17069
1850		12	2312	148	669		1003	213	11	314	1119	1	302	305		23192
1860		52	2906	175	704	5	1110	604	40	315	1220	12	377	776		31443
1870		91	3522	217	706	12	1259	819	87	331	1225	24	442	1055	9	38558
1880		175	4283	277	996	98	1542	1592	144	332	1513	75	618	1315	21	50189
1890	259	318	5258	346	1151	349	1768	2236	211	332	1656	357	763	1693	63	62980
1900	790	414	6302	429	1340	402	2021	3049	277	344	1854	518	959	2069	93	76212
1910	1657	673	7665	543	1515	584	2185	3897	373	356	2062	1142	1221	2334	146	92228
1920	2028	783	8720	604	1684	637	2338	4663	449	352	2309	1357	1464	2632	194	106022
1930	2396	954	9631	687	1739	693	2617	5825	508	360	2422	1563	1729	2939	226	123203



1940	2336	1090	9900	713	1900	643	2916	6415	550	359	2678	1736	1902	3138	251	132165
1950	2233	1521	10498	792	2117	653	3292	7711	689	378	3319	2379	2006	3435	291	151326
1960	2328	1769	11319	859	2383	681	3567	9580	891	390	3967	2853	1860	3952	330	179323
1970	2559	2092	11801	950	2591	666	3926	11199	1059	445	4651	3413	1744	4418	332	203302
1980	3025	2633	11865	947	3121	691	4591	14226	1461	511	5347	4132	1950	4706	470	226542
1990	3146	2842	11882	1003	3487	696	4877	16987	1723	563	6187	4867	1793	4892	454	248710
2000	3451	3421	12281	1048	4012	755	5689	20852	2233	609	7079	5894	1808	5364	494	281422
2010	3751	3831	12702	1053	4625	814	6346	25146	2764	626	8001	6725	1853	5687	564	308746

[Return to Chris' Homepage](#)



[Return to UW Oshkosh Homepage](#)

Managed by chris edwards:
click to email [chris edwards](#)
Last updated August 1, 2011