Day By Day Notes for MATH 301

Spring 2015

Day 1

Activity: Go over syllabus. Take roll. Overview examples: Randomness - coin example. Gilbert trial.

I have divided this course into three units. Unit 1 (Days 1 through 9) is about summarizing data and basic probability. Unit 2 (Days 10 through 17) is about various common distributions and sampling distributions. Unit 3 (Days 18 through 27) is about statistical inference. One common theme throughout the course is mathematical modeling. We are often trying to explain and predict what we see in the real world with equations and models. Before we can say that a model is appropriate, we must understand the consequences of the equations we have chosen. In Unit 2, we will explore a number of such models and their effects. We will use techniques from Unit 1 to describe our models and ultimately we explore the ideas in Unit 3 to use our models in real life situations. I will try to keep us focused on the "big picture" of statistics as a discipline as we proceed, but sometimes our focus will be on the details of algebra or calculus to get us over some hurdles.

I use the Gilbert trial example on the first day 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 (Unit 1). 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 2. And the trial strategy itself is what statistical inference is all about (Unit 3). Many of you will encounter inference when you read professional journals in your field and researchers use statistics to support their conclusions of improved treatments or to estimate a particular proportion or average.

I believe to be successful in this course you must actually read the text and these notes carefully, working many problems. The most important thing is to **engage** yourself in the material. However, our class activities will often 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, tomorrow 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 am disappointed with the small number of students who come to me for help outside of class. 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; there is no such thing as a stupid question. Often other students have the same questions but are also too shy to ask them in class. If you are still reluctant 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 allotting 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.

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

Review course objectives: collect data, summarize information, model with probability, make inferences.

Reading: (The reading mentioned in these notes refers to what reading you should do for the **next** day's material.)

Section 1.1.

Day 2

Activity: Random Sampling.

Often when we analyze data, our underlying mathematical model will assume the data was collected randomly. Sometimes this means we have sampled some items from a population. Other times we mean that the observations are simply independent of each other. (We will explore the idea of independence in more detail in Chapter 2.)

An example of sampling is when we select some items from our assembly line to check for defects. We will not go into any details of the various methods of sampling but will instead focus only on Simple Random Sampling, which means each item in the population has the same chance to be included in the sample.

An example of the independence idea of random data is repeated measurements on an individual, like blood pressures. We aren't choosing some measurements from a population of measurements; rather we are assuming that each measurement yields a value that is not influenced by previous measurements.

Before we begin summarizing data, I want to have you explore generating some random data.

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

The text briefly discusses collecting random samples. I want us to gain some practical experience collecting real simple random samples, so we will use four methods of sampling today: dice, cards, a table of random digits, 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 trust that hopefully 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 (repeats). 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.

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.

By drawing three cards from a deck of six cards, generate a sample of three people. (Let each card represent a person.) Repeat 20 times, giving 20 samples of size 3.

Using the table of random digits, 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.

Using your calculator, select three people. The TI-83 command MATH 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.) If your calculator has a rand function only, you can achieve the same result as the TI-83 MATH randInt(2, 4) with int(3*rand) + 2. Repeat 20 times, giving 20 more samples of size 3.

Your group should have drawn 80 samples at the end. Keep careful track of which samples you selected; record your results in order, as 125 or 256, for example. (125 would mean items 1, 2, and 5 were selected.) We will pool the results of everyone’s work together on the board.

Goals: Gain practice taking random samples. Understand what a simple random sample is. Become familiar with randInt(. Accept that the calculator is random.

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.

- **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 results in an SRS.

- **Select an SRS from a list of items.** The TI-83 command randInt(will 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.** 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: Section 1.2.

Day 3

Activity: Graphical summaries of data. Pinch Hitting/Defense.

As we begin Unit 1, let's look at the "big picture" first. When we choose a model to describe a real world phenomenon, we have to be able to describe and summarize what we have. Sometimes this will involve numerical calculations; other times we use pictures and graphs. Still other times we will use advanced mathematics like calculus. The first methods we will look at are the graphical measures.

Numerical summaries (like averages) may **over-summarize**. That is, important information may be lost. An alternate approach to just a few summary statistics is a graph of the data, highlighting the key features. We will look at four techniques today, each with their own strengths and weaknesses.

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 with class intervals using the base 10 system. This means the intervals will always be ten "units" wide, such as 10 to 19, or 0 to 9, or 0.00 to 0.09. The "unit" chosen is called the **stem**, and the next digit after the stem is called the **leaf**.

The **histogram** is a picture of the location of data on a number line. It is composed of rectangles whose areas reflect the relative frequency of data. Area is the key idea, not height, although if all rectangles have the same width then area and height are proportional. The vertical scale is **density**, or area per unit width. One drawback of histograms is that knowing the relative frequency in an interval does not indicate **where** in the interval the data may be concentrated. Thus, **clustering** cannot be determined completely with a histogram.

The **box plot** is a graph of the five-number summary. The text discusses the five-number summary in detail in Section 1.4. The box portion is the middle half of the data, from the first quartile to the third quartile. The whiskers are the lines drawn outward from the box, representing the upper and lower quarters of the data. (Boxplots are introduced on page 39.)

QUANTILE is a program I wrote for the TI-83 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 values versus the rank, or percentile, in the data set. Quantile plots always increase from left to right. The syntax 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. The program also changes the settings for STATPLOT 1.

QUANTIL2 is a companion program for comparing two lists of data simultaneously. This program additionally uses C and D as temporary lists, and changes the settings for STATPLOT 2.

Generally, using the TI-83 to view box plots, histograms, scatter plots, and (later) normal probability plots, 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 can have up to three plots on the screen at once. For histograms, we will only use one at a time. For box plots, we often use multiple displays, to compare several lists.

3) ZOOM 9 This command centers the window "around" 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!)

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.

Here are two lists from some baseball data I was looking at recently. The first list is from the National League, and the second list is from the American League. My question is what, if any, are the differences between the leagues. (These represent team totals for pinch hitting appearances where the player then took the field in the next inning to play defense.) I will use this data in class, using the TI-83 and also using a program available on campus computers called MINITAB, to demonstrate the graphical methods of summarization. Tomorrow we will practice on another data set using weather data.

33, 58, 54,34,41,21,44,50,45,55,57,58

95,19,34,39,65,58,38,28,28,49,55,52,158,48

Goals: Be able to use the calculator to make a histogram, box plot, or a quantile plot. Be able to make a stem plot by hand.

Skills:

- **Summarize data into a frequency table.** The easiest way to make a frequency table is to TRACE the boxes in a histogram and record the classes and counts. You can control the size and number of the classes with Xscl and Xmin in the WINDOW menu. The decision as to how many classes to create is arbitrary; there isn't a "right" answer. One popular rule of thumb is use the square root of the number of data values. For example, if there are 25 data points, use 5 intervals. If there are 50 data points, try 7 intervals. This is a rough rule; you should experiment with it. The TI-83 uses its own rule for doing this; I do not know what the rule is. You should experiment by changing the interval width and see what happens to the diagram.

- **Use the TI-83 to create an appropriate histogram, box plot, or quantile plot.** STAT PLOT is our main tool for viewing distributions of data. Histograms are common displays, but have flaws; the choice of class width 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 steep.

- **Create a stem plot by hand.** The stem plot is a convenient manual display; it is most useful for small datasets, but not all datasets make good stem plots. Choosing the "stem" and "leaves" to make reasonable displays will require some practice. Some notes for proper choice of stems: if you have many empty rows, you have too many stems. Move one digit (place) 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 (place) and try again. Some datasets will not give good pictures for any choice of stem, and some benefit from splitting or rounding (see the examples in class).

- **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 shape very well (apart from rough symmetry); you can only see where the quartiles are. Further, boxplots by their nature are unable to display gaps in the data. Alternatively, you should know that the box plot often gives a reasonable first impression.

- **Compare several lists of numbers, using box plots and quantile plots.** For two lists, the best simple approach is the back-to-back stem plot. For a quick "snapshot" of more than two lists, I suggest trying box plots, side-by-side, or stacked. At a glance, then, you can 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 STAT PLOT. You can superimpose quantile plots to detect differences in the distributions by the different values of the percentiles.

Reading: Section 1.2.

Day 4

Activity: Continued graphical summaries of data. Arizona Temps.

Yesterday we looked at the technical details of making the various displays for summarizing a single list of data. Today we will interpret what these graphs are telling us about the patterns in the data. Specifically we want to be able to say what the center of the data is, how spread out the data are, and what shape the distribution has. Starting tomorrow, we will quantify these measures, but today I want to begin by looking at the graphs.

Use the Arizona Temps dataset to practice creating the histograms, stem plots, box plots, and quantile plots for several lists. Compare and interpret the graphs. Identify shape, center, and spread.

As you explore using technology to summarize data, answer these questions:

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 affect the calculator's routines? (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 or a ridiculously low value and recalculate the graphs and measures. Notice the effects of these extreme values.)

3) What information does the box plot disguise?

Notes: Superimposed histograms on the TI-83 do not do a good job comparing two lists, and the stem and leaf is too difficult to modify. I recommend using box plots or QUANTIL2.

Goals: Be able to qualitatively describe the features of a distribution.

Skills:

- **Describe shape, center, and spread.** From each of our graphs, you should be able to make general statements about the shape, center, and spread of the distribution of the variable being explored. The techniques we've examined display these features differently; your job is to be able to identify the salient features from each of the techniques.

Reading: Sections 1.3 and 1.4.

Day 5

Activity: Numerical summaries of data. *Dance Fever* example.

Our second topic is the 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 2. 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 dataset to calculate means, standard deviations, and the 5-number summaries. To calculate our summary statistics with the TI-83, we will use STAT CALC 1-Var Stats (to use List 1) or STAT CALC 1-Var Stats L2 for List 2, for example. There are two screens of output; we will be mostly concerned with the mean $\bar{x}$, the standard deviation Sx, and the five-number summary on screen two.

Using the summary statistics now instead of the plots, answer these questions again:

1) Are high and low temperatures distributed the same way, other than the obvious fact that highs are higher than lows?

2) How does a single case affect the calculator's routines? (What if we had had an outlier?)

3) What information does the 5-number summary disguise?

Compare these measures with the corresponding graphical measures you calculated on Day 4. Notice that the box plots and numerical measures cannot describe shape very well.

Now, 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, and is one way to interpret roughly what the standard deviation is measuring.

Finally, compare the standard deviation to the Interquartile Range (IQR = Q3 – Q1). Note that for lists without outliers, the IQR will typically be larger than the standard deviation. This is reasonable as the IQR "catches" 50 % of the data while an interval **two** standard deviations wide "catches" roughly 67 % of the data.

Goals: Compare numerical measures of center and spread. Use technology to summarize data with numerical measures. Interpret standard deviation as a measure of spread.

Skills:

- **Understand the effect of outliers on the mean.** The mean (or average) is unduly influenced by outlying (unusual) observations. Therefore, knowing when your distribution is skewed is helpful.

- **Understand the effect of outliers on the median.** 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 may be as simple as entering numbers into your calculator and pressing some buttons: STAT CALC 1-Var Stats. Or, if you are doing some things by hand, you may have to organize information the correct way, such as listing the numbers from low to high. **IMPORTANT NOTE:** Please get used to using the statistical features of your calculator to produce the means, standard deviations, etc. While I know you can calculate the mean by simply adding up all the numbers and dividing by the sample size, you will not be in the habit of using the full features of your machine, and later on you will be wasting time by entering your data twice.

- **Understand standard deviation.** At first, standard deviation will seem foreign to you, but I believe that it will make more sense the more you work 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 dataset is about 4 or 5 standard deviations. Another interpretation is that the standard deviation is roughly ¾ 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 Chapter 3 will further help us interpret standard deviation.

Reading: Sections 2.1 to 2.2

Day 6

Activity: Sample Spaces. Venn Diagrams. Coins, Dice. **Homework 1 due today.**

Today we begin our exploration of probability. This will include the sample space method of doing probability as well as some introductory combinatorics. There is also some benefit to **simulating** experiments to estimate a probability **empirically**; unfortunately, unless the simulation is performed many thousands of times, such estimates are usually not reliable.

In our sample space method, we will list the equally likely ways the experiment **could** have come out, and then simply count which ones are favorable to our particular question. Unfortunately, oftentimes the sample space is too large to list. In such cases, it may still be beneficial to at least **begin** the list or to contemplate the process without actually completely listing the sample space. We will practice these ideas with dice today.

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 randInt(1, 6) on your TI-83 instead of actually rolling dice, that would be fine. One way to compare dice rolls 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. 100 is probably not enough trials, but the TI-83 isn't really a computer, so putting in something like 1000 will probably crash it!

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.

Basic probability rules:

1) Probability is a number between 0 and 1, inclusive.

2) Mutually Exclusive events add when finding the union.

3) Mutually Exclusive and exhaustive events add to one.

Use Venn diagrams to "prove":

$A = (A \cap B) \cup (A \cap B')$

$(A \cup B)' = A' \cap B'$ and $(A \cap B)' = A' \cup B'$.

The Venn diagram is a way of partitioning a sample space into events with (usually) circles. We can use shading to indicate intersections and unions, and we can even prove some results using such diagrams. We can prove the De Morgan's rules and the inclusion/exclusion formula using Venn diagrams. Note to self: Steve Miller band Venn diagram. Image.

Goals: Create sample spaces. Use Venn diagrams to organize sample spaces. Use simulation to estimate probabilities. Know the rules of probability, including addition, complement, and inclusion/exclusion.

Skills:

- **Know the definitions of Sample Space, Event, Outcome, etc.** The basic language of probability will be used throughout the course, so it is important for you to be conversant in it. Basically, the sample space is a collection of outcomes, our basic smallest sets. Events are collections of outcomes, and are usually subsets of the sample space, although the sample space itself is also an event.

- **Be able to use a Venn diagram.** The Venn diagram is a way of partitioning the sample space into mutually exclusive regions. It can be useful for simply organizing sets, or sometimes is quite useful in understanding proofs (the inclusion/exclusion formula is a good example.)

- **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 too 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.

- **Understand the probability rules.** Being adept at probability begins with knowing definitions and knowing basic formulas. For example, you can't prove things about mutually exclusive sets if you can't recite the definition of mutually exclusive. Memorize at first; later it becomes "learned", not "memorized".

- **Relate the rules to sample spaces.** Remember that the rules we're discussing are all based on counting elements in sample spaces. Sometimes it is helpful to have a few "standard" examples in mind so conjectures or steps in reasoning can be verified. For example, the inclusion exclusion principle is shown well with the two-dice problem "What is the chance of at least one six?" Ignoring the intersection makes the probability found using addition too large.

- **Simulation can be used to estimate probabilities.** If the number of repetitions of an experiment is large, then the resulting observed frequency of success (the empirical probability) 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, 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 precision for our purposes.

Even with 500, 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 3.

Reading: Section 2.3.

Day 7

Activity: Combinations and Permutations. Pascal's Triangle

As we try to count outcomes in an event, to determine the probability of the event, we discover many useful patterns. Pascal's triangle contains what are known as the **binomial coefficients**, or the **combination formula**. The two main relationships we need from this material are $C(n,r)$ and $P(n,r)$. Basically, $C(n,r)$ counts subsets and $P(n,r)$ counts orderings of those subsets. One could easily spend days exploring Pascal's Triangle; we will spend only a few sessions on it.

Combinations are quite useful in our discrete probability distributions of Chapter 3. In particular, the binomial, hypergeometric, and negative binomial distributions all use combinations.

Arrange the letters in FREDA. Arrange the letters in FREED. Arrange the letters in ERRORS. Arrange the letters in SETTER.

There are several approaches to the above tasks. I will cover the two main techniques of counting that we find important. Keep in mind that we will only look at combinations and permutations briefly; one could easily devote an entire course to the topic.

Goals: Explore Pascal's Triangle and how it relates to counting (permutations and combinations).

Skills:

- **Recognize the usefulness and properties of Pascal's Triangle.** Pascal's Triangle is old (known to the Indians, the Persians, and the Chinese as early as the 10th century) yet is still quite useful. There are just two rules to construct Pascal's Triangle: each row begins and ends with a 1, and each entry is the sum of the two entries above it to the left and the right. From such a simple construction, though, we encounter many relationships: the combination formula, the triangular numbers, the Fibonacci numbers, powers of 2, among others. Our chief interest is in the combination formula and its relationship to the binomial distribution.

- **Know the Permutation and Combination formulas.** When counting the number of ways of choosing items or ordering items, our formulas are $C(n,r)$ and $P(n,r)$, respectively. You will need to work enough problems so that you know when to use each of them. One way to keep them straight is to think of a Combination as a Committee of people, and a Permutation as a Photograph of that committee. (There are more permutations than combinations for a particular choice of n and r, because there are many different arrangements of the people on the committee for the committee photograph.) Also don't forget our trick of listing the complete sample space, but only for small problems!

Reading: Section 2.4.

Day 8

Activity: Continued Combinations and Permutations.

Using our algebra skills, we can manipulate the combination formulas and develop some rules. Chief among these rules is Pascal's identity, which is the essential rule that produces Pascal's triangle.

Problems from Tucker.

Goals: Continue exploring Pascal's Triangle and how it relates to counting (permutations and combinations).

Skills:

- **Know the Permutation and Combination formulas.** When counting the number of ways of choosing items or ordering items, our formulas are $C(n,r)$ and $P(n,r)$, respectively. You will need to work enough problems so that you know when to use each of them. One way to keep them straight is to think of a Combination as a Committee of people, and a Permutation as a Photograph of that committee. (There are more permutations than combinations for a particular choice of n and r.) Also don't forget our trick of listing the complete sample space, but only for small problems!

Reading: Section 2.4.

Day 9

Activity: Presentation 1.

Summaries (Chapter 2)

Gather 3 to 5 variables on at least 20 cases; the source is irrelevant, but knowing the data will help you explain its meaning to us. Be sure to have at least one numerical and at least one categorical variable. Demonstrate that you can summarize data graphically and numerically.

Day 10

Activity: Constructing probability trees.

Consider a card trick where two cards are drawn sequentially off the top of a shuffled deck. (There are 52 cards in a deck, 4 suits of 13 ranks.) We want to calculate the chance of getting a heart on the first draw, a heart on the second draw, and hearts on both draws. We will organize our thoughts into a tree diagram, much like water flowing in a pipe. On each branch, the label will be the probability of taking that branch; thus at each node, the exiting probabilities (conditional probabilities) add to one.

On the far right of the tree, we will have the intersection events. Each intersection probability is found by multiplying the branch probabilities leading to that event. At each branch, we have a **conditional probability**, which is a probability calculation based on what has already occurred to get to where we are in the tree. The notation we will use is $P(A|B)$ which reads "the probability of event A given that event B has occurred". The multiplication formula we use in trees is $P(A \cap B) = P(A)P(B|A)$. Note that this can also be written with the roles of A and B reversed: $P(B \cap A) = P(B)P(A|B)$. Combining these two formulas gives us the basis of Bayes' formula: $P(A) = P(B)P(A|B)/P(B|A)$. I will demonstrate how to use these conditional formulas using the rare disease problem on Day 11.

For the 2-cards problem, calculate the chances of:

1) Drawing a heart on the first card.

2) Drawing a heart on the second draw given that a heart was drawn first.

3) Drawing two hearts.

4) Drawing at least one heart.

5) Drawing a heart on the second card.

6) Drawing a heart on the first draw given that a heart was drawn second.

Goals: Be able to express probability calculations as tree diagrams.

Skills:

- **Know how to use the multiplication rule in a probability tree.** Each branch of a probability tree is labeled with the conditional probability **for that branch**. To calculate the joint probability of a series of branches, we multiply the conditional probabilities together. Note that at each branching in a tree, the (conditional) probabilities add to one, and that overall, the joint probabilities add to one.

- **Recognize conditional probability in English statements.** Sometimes the key word is "given". Other times the conditional phrase has "if". But sometimes the fact that a statement is conditional is disguised. For example: "Assuming John buys the insurance, what is the chance he will come out ahead" is equivalent to "If John buys insurance, what is the chance he will come out ahead".

Reading: Section 2.5.

Day 11

Activity: Demonstrating Bayes' with the rare disease problem. **Homework 2 due today.**

Today we will continue the conditional probability discussion by looking at Bayes' Theorem. Essentially, Bayes' Theorem allows us to reverse the conditions in a conditional probability. This commonly occurs when data or information is available in one direction or order of events, but the interesting questions involve the reverse direction or order. The rare disease problem we do today is an example.

Another extremely important notion to our models we will create next chapter is the idea of independence. There are two definitions you will encounter. They are equivalent, as each follows from the other. They are $P(A \cap B) = P(A)P(B)$ and $P(A) = P(A|B)$. The first definition is the one we will use most often, as it says that we can multiply the two probabilities of independent events together to get the joint probability. We saw this already when we looked at the sample space for rolling two dice.

Construct the probability tree for the rare disease problem (Example 2.31, page 79).

Change the 0.001 to .3 and consider the common disease problem.

Notice there is a distinct difference between two events being independent and two sets being disjoint. Two sets are disjoint, or mutually exclusive, if their intersection is empty, which would imply $P(A \cap B) = 0$. But independence implies that $P(A \cap B) = P(A)P(B)$, which is typically a non-zero result. To put it another way, disjoint events displayed in a Venn diagram cannot overlap, but independent events *do* overlap, and in just the right way so that the percentage of the overlap within one of the circles is in the same ratio as that circle is to the whole. I will draw some diagrams in class to demonstrate.

Goals: Be able to reverse the events in a probability tree, which is what Bayes' theorem is about.

Skills:

- **Be able to use the conditional probability formula to reverse the events in a probability tree.** The key here is the symmetry of the events in the conditional probability formula. We exchange the roles of A and B, and tie them together with our formula for $P(A \cap B)$. This reversal is the essence of Bayes' theorem.

- **Know the definition of independence.** Independence is a fact about probability, not about sets. Contrast this to "disjoint" which is a property of **sets**. In particular, independent events are by definition **not** disjoint. Independence is important later as an assumption as it allows us to multiply individual probabilities together without having to worry about conditional probability.

Reading: Sections 3.1 and 3.2.

Day 12

Activity: Continue coins and dice. Introduce Random Variables.

I like to introduce discrete probability distributions by continuing the probability discussions we've had already with dice and coins. Essentially we can further partition the sample space by assigning a numerical value to each outcome (which we call a random variable), and then counting how many outcomes are in the events associated with each possible value of the random variable. A collection of these assigned numbers and their associated probability is what we call a **probability distribution**, or a **probability mass function** (**pmf**). If we cumulate these probabilities, we get the **cumulative distribution function** or **cdf**.

In class, I will produce some distributions for us, with coin tossing and dice rolling. Specifically, we want to find the chance of getting any number of heads on n tosses of a fair coin, and the chance of getting sums when rolling two or more dice.

Answer the following questions, using complete distributions:

1) What is the chance of getting a sum of 8 on two dice?

2) What is the chance of getting a sum of 10 on two dice?

3) What is the chance of getting a sum of x on two dice, where x is between 1 and 13?

4) What is the chance of getting 10 heads on 20 flips of a fair coin?

5) How can you get the TI-83 to graph a probability histogram?

Derive a pmf and its cdf. Use the sum on two dice as an example. Know how to work back and forth from one to the other.

Goals: Understand that variables may have values that are **not** equally likely.

Skills:

- **Understand discrete random 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 Section 3.4. 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.

Reading: Section 3.3.

Day 13

Activity: Continue working with Random Variables. Introduce Expected Value and Variance.

We can use the **Frequency** option of the TI-83 to find μ and σ for discrete distributions. This option allows different weights for the data entered. When we used STAT CALC 1-Var Stats L1 before, the values in L1 were each representing one data value each. In probability distributions, we don't have actual data, but **possible** values, the values of the **Random Variable**, and their associated probabilities. To find the mean and standard deviation of the distribution, then, we let the probabilities be the weights. We will practice with the dice rolling results we created today.

To summarize pmf's, we will typically calculate two values, the **Expected Value** and the **Variance**. You should be familiar with Expected Value already; your GPA is an example. I will go over a few EV calculations for the dice problems, then move on to other functions, like x^2. Finally, we will see that the Variance is a form of Expected Value.

From our distributions for flipping coins and rolling dice that we produced on Day 12, calculate the Expected Value and the Variance. You may either calculate using the formulas or the TI-83 frequency option, but make sure you know what the machine is doing for you.

Goals: Understand how to summarize a pmf with Expected Value and Variance.

Skills:

- **Know the definition of a discrete probability mass function (pmf).** If a non-negative function sums to 1 over some set, then we have a discrete pmf. It is not necessary for the set to be finite; this means we may need to work with infinite sums. Because each item in the sum is a probability, it **is** necessary that each value is less than one. (Contrast this with the continuous distributions on Day 14.)

- **Know the definition of a discrete cumulative distribution function (cdf).** If a non-decreasing function begins at 0 from the left and ends at 1 on the right, and has no place where the derivative is non-zero, then we have a discrete cdf. The key is that discrete cdf's are "steps", a collection of flat sections with discrete jumps in between.

- **Know how to use the TI-83 to calculate the mean and variance for a discrete distribution.** By including a variable of weights or frequencies (probabilities), the TI-83 will calculate μ and σ for a discrete distribution. The syntax is STAT CALC 1-Var Stats L1, L2, where the x-values are entered in L1 and the corresponding probabilities are entered in L2.

Reading: Sections 4.1 to 4.2.

Day 14

Activity: Introduce Continuous Random Variables. Use the cdf technique to randomly generate data. **Homework 3 due today.**

The histogram technique we've seen is a graph that shows where data is concentrated. In a similar way, the probability density function (pdf) shows us where **probability** is concentrated. Because probability is theoretical (as opposed to empirical data) the pdf is a smooth curve instead of a series of rectangles as in the histogram. Calculus is the chief tool we will need to be able to manipulate pdf's for continuous variables.

Just as in the discrete case, the **cumulative distribution function** is a calculation of the probability up to a point, cumulatively. Hopefully, you are reminded of a result from calculus, the definite integral. We will not always be able to find the cdf in a closed form expression, as some indefinite integrals do not have closed form anti-derivatives. Many such examples exist in common statistical distributions, the normal curve being the most common.

Often we will want to generate random data from various distributions. A common technique for doing this is to use the **cdf method**. Caution: the cdf method **only** works for cdf's that have **closed form inverses**. Specifically, the **normal curve** will **not** work with this approach. To use the cdf method, replace $F(x)$ with $rand$, and solve for x. $rand$ is the discrete uniform distribution from 0 to 1, or our most basic notion of a continuous random number.

Goals: Introduce continuous distributions.

Skills:

- **Know the definition of a probability density function (pdf).** If a non-negative function integrates to 1 over some interval, then we have a probability density function. Notice that the function can certainly be over 1 (contrast to pmf's); the key here is that the **area** is one, not the maximum height.

- **Know the definition of a continuous cumulative distribution function (cdf).** If a continuous non-decreasing function begins at 0 from the left and ends at 1 on the right, then we have a discrete cdf. The key is the continuity. If a function has **only** jumps, it is discrete. If a function has **no** jumps, it is continuous. A function with both is **mixed**. An example of a mixed distribution is a question like "If you are employed, what is your income?" People without a job have no income, so a proportion of the data have a zero for that variable, and in the distribution, there is a spike at 0.

- **Realize that these formulas and functions we are exploring are simply models.** What we are trying to do with these functions is model real world data with simple curves. We hope to have both simple equations and close fitting histograms and quantile plots. The best situation is a model that has parameters that can be "tuned". By choosing the parameters judiciously we can find curves that approximate real data.

- **Know how to use the cdf equation to generate random values from the variable.** The cdf ranges from 0 to 1, and each value represents a percentile. If we produce a uniform number between 0 and 1, such as with rand from the TI-83, then by cross-referencing on the graph (or solving in the formula), we can tell which x-value corresponds to that uniform random number. While this can be done visually, it's not very useful unless the equation can be solved explicitly.

- **Know the definition of Expected Value.** Expected Value is calculated as the average product of the random variable and its probabilities. For discrete variables, this is a sum, and for continuous variables it is an integral. You should realize that the definition is really the same for both discrete and continuous distributions. Both definitions involve a summed product. The nature of the summing is the only difference between them. Once you understand that integration is simply repeated addition, the similarity of the two formulas is clear.

- **Know the definition of Variance.** Variance is a particular kind of Expected Value. The Expected Value of a function of x is simply the average product of that **function** of x and its probabilities. For Variance, the function of x that we use is the squared difference from the mean.

- **Understand that Expected Values and Variances are parameters and should have no x-values.** After summing or integrating, the variable of summation or integration disappears. This is most easily seen in the Fundamental Theorem of Calculus where we substitute numbers (the integrands) into the x-variable in the anti-derivative. Explicitly we note that x equals those integrand values. Therefore, Expected Value and Variance are constants.

Reading: None

Day 15

Activity: Presentation 2.

Probability (Chapter 2)

Choose one of the following games and

1) Give us a short history of the game.

2) Describe how randomness is part of the game.

3) Using our probability rules, show us an example using this game.

Games: Risk, Blackjack, Backgammon, Roulette, Battleship, Poker, Minesweeper, Cribbage

Reading: Section 4.3.

Day 16

Activity: Normal Curve. Introduce the TI-83's normal calculations.

Today I will introduce the normal curve, the most used distribution (model) in statistics.

We have encountered this model before, with dice, and we see the shape begin to appear if we graph the values from Pascal's Triangle. Because the pdf for the normal is an equation for which there is no corresponding anti-derivative (in a closed form), we must resort to numerical integration to calculate areas under the curve. Every statistics textbook has a **normal table** for this purpose. You are welcome to use the text's normal table; however, I think you will find it much easier to use the built-in functions in the TI-83, or MINITAB. Incidentally, the first areas under this curve were calculated by Abraham De Moivre, in 1733. Details here.

The TI-83 commands we will use are DISTR normalcdf(and DISTR invNorm(. DISTR normalcdf(is used to find the area under the curve between two x-values, and DISTR invNorm(is used to find a percentile. For example, the command DISTR normalcdf(10, 20, 14, 3) will give the area between 10 and 20 using a mean of 14 and a standard deviation of 3. (Draw a sketch of this.) If you leave off the mean and standard deviation, 0 and 1 will be assumed. The command DISTR invNorm(0.7, 14, 3) will give the x-value that has area 0.7 to the left of it, using a mean of 14 and a standard deviation of 3. Again, if you leave off the mean and standard deviation, 0 and 1 are assumed.

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 7.)

Goals: Introduce the normal curve.

Skills:

- **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 two functions, DISTR normalcdf(and DISTR invNorm(, that 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 cumulative percentage.

Reading: Chapters 1 through 3.3 and Sections 4.1 to 4.2.

Day 17

Activity: Exam 1.

This first exam will cover numerical summaries (summary calculations), graphical summaries (pictures), and probability (including discrete and continuous random variables). The normal calculations will **not** be on Exam 1.

Reading: Section 4.3.

Day 18

Activity: Practice normal calculations.

More

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 Interquartile Range (IQR) for the standard normal (mean 0, standard deviation 1). Compare this to the standard deviation of 1.

3) Women aged 20 to 29 have normally distributed heights with mean 64 and standard deviation 2.7. Men have mean 69.3 with standard deviation 2.8. What percent of women are taller than the average man, and what percent of men are taller than the average 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 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.)

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.

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. These values can be used as rough guidelines; if precision is required, you should use the TI-83 instead. I will assume you know these numbers by heart.

- **Understand that summarizing with just the mean and standard deviation is a special case.** We have progressed from pictures like histograms to summary statistics like medians, means, etc. to finally summarizing an entire list with just 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 incomplete. Also, it is important to realize that these calculations are only approximations; usually in the real world data is only **approximately** normally distributed.

Reading: Section 4.4.

Day 19

Activity: Introduce Gamma.

The next pdf we will explore is the very flexible **Gamma Distribution**. (It seems Euler himself worked on the factorial nature of the gamma function as early as the 1720's.) Unfortunately, calculating areas requires both numerical integration and a hard-to-calculate function called the gamma function, and therefore tables are not readily available. Our book **does** have a table called the **incomplete gamma distribution**, but we can use our TI-83's **chi-square** distribution and some theory from Mathematical Statistics MATH 401 to find these same areas. The procedure is to convert the desired x-value via $y = 2x/\beta$ and to use 2α degrees of freedom, df. Then DISTR χ^2cdf(0, y, df) on the TI-83 will give the probability that the gamma variable is less than x. To find right tail areas (above x), we use complements (subtraction). Also, for certain degrees of freedom we know the pdf, and we can then simply ask our calculators to do a numerical integration to get the desired probability.

One use of the gamma distribution is in describing the shape of the sampling distribution (end of Chapter 5) for the sample variance. As we saw in the preceding paragraph, the chi-squared distribution is a specific case of the gamma distribution, and that table is in our book as well.

Using the pdf, we can derive expressions for the mean and variance. Note: this work will focus on details of calculus; don't forget the "big picture". We are trying to model some real world data, like income data, and we need to know the mean and variance of our model to make sure it matches reality.

Goals: Introduce the Gamma distribution.

Skills:

- **Know the shapes of the gamma distribution.** By adjusting the parameters, the gamma distribution can take on a number of shapes. You should explore the various combinations to see when the curve begins at 0, when it begins at infinity, and when it begins at some value in between.

- **Know the properties of the gamma distribution.** You should know the mean and variance of the gamma distribution. The Expected Value (mean) is $\alpha\beta$ and the Variance is $\alpha\beta^2$.

- **Know how to calculate gamma probabilities when α is an integer.** With the appropriate transformation ($y = 2x/\beta$) we can use our TI-83's χ^2 function to find gamma probabilities. If α is not an integer, you will have to interpolate. Another solution is to use the calculator's definite integral abilities. This of course requires you to know the pdf.

Reading: Section 4.6.

Day 20

Activity: Examine Probability Plots.

If we have memorized all the cdf's of interest, such as the normal distribution, the gamma distribution, or any other distribution for which we have data, we could compare the observed quantile plot with the theoretical cdf to see if they match. Unfortunately this is not a simple exercise. First of all, we are not very good at memorizing **specific** shapes of these cdf's. Secondly, departures from the models are not always of the same sort of magnitude and are then more difficult to detect. However, because of the importance of the normal distribution, much work has been done to "detect normality", as it is called. On the TI-83, the last icon in STATPLOT is a one such quick method for assessing normality. Essentially, the y-axis has been transformed so that the normal cdf, which is usually a squiggly S-shape, is a straight line. We are often pretty good at detecting straight lines, so this technique can be informally done at a glance. A formal treatment using normal probability plots requires fancier statistical mathematics, something you will encounter in Applied Regression MATH 385 or Mathematical Statistics MATH 401.

MINITAB can make empirical probability plots for us, and I will show you the commands in class.

To demonstrate the normal probability plots, and the effects of random variation on the "straightness" of the resulting lines, we will simulate some data today on the calculator. Using the command MATH PRB randNorm(0, 1, 100) -> L1, store 100 observations in L1. Produce a normal probability plot with STAT PLOT.

Now, generate random numbers using the uniform distribution (randInt). The uniform distribution is nothing like the normal curve, so we expect the normal probability plot to be quite different from a straight line. Note that the uniform distribution is "tail heavy", unlike the normal curve which is "tail light".

Typically in real world problems we are more concerned with outliers, or skewed asymmetric distributions. For our third example, generate standard normal numbers and then square them. This particular transformation creates a heavily right skewed distribution. Make a normal probability plot to see the effect of a heavy right skew.

For all three of these examples, you should repeat the exercise several times, to get some experience with how variable randomness is. In particular, our first list of pure normal numbers should theoretically produce a straight line every time, but it doesn't. We need to be very aware of this seeming paradox when we move on to inference procedures, as that is the basis for most of our results. How much variability is within the normal bounds of randomness is the fundamental context for our scientific conclusions.

Goals: Know about the normal probability plot.

Skills:

- **Have some expertise using the normal probability plot.** Using the TI-83 we can produce normal probability plots. With MINITAB, we can even make other probability plots. In either case, we are comparing our observed data to the theoretical expectation and noting whether any observed differences are beyond what would be expected under randomness. The normal probability plot on the TI-83 should look like a straight line if the normal model fits the data well.

Reading: Section 3.4.

Day 21

Activity: Introduce Binomial. **Homework 4 due today.**

Our topic today is the most useful discrete distribution, the **binomial distribution**. Our most common example is coin flipping. We need four conditions to apply the binomial model to our data. 1) We must have only two possible outcomes, successes and failures. 2) We must have a fixed number of trials. In other words, the number of trials performed is always the same number. 3) We must have independence between trials. And 4) We must have the same chance of success on each trial.

Coin flipping is a good way to understand randomness, but because most coins have a probability of heads very close to 50%, we don't get the true flavor of the binomial distribution. Today we will simulate the flipping of an unfair coin; that is, a binomial process with probability not equal to 50%.

Experiment 1: Our unfair "coin" will be a die, and we will call getting a 6 a success. Roll your die 10 times and record how many sixes you got. Repeat this process 10 times each. Your group should have 40 to 50 trials of 10 die rolls. Pool your results and enter the data into a list on your calculator. We want to see the histogram (be sure to make the box width reasonable) and calculate the summary statistics, in particular the mean and variance. Also produce a quantile plot. Compare the simulated results with theory. (Of course, you may use the TI-83 to "roll" your die if you like. See Day 6.)

Experiment 2: Your calculator will generate binomial random variables for you, but it is not as illuminative as actually producing the raw data yourself. Still, we can see the way the probability histogram looks (if we generate enough cases; this is an application of the law of large numbers). I suggest 100 at a minimum. Again be sure to make your histogram have an appropriate width. The command is MATH PRB randBin(*n*, *p*, *r*), where *n* is the sample size, *p* is the probability of success, and *r* is the number of times to repeat the experiment.

Goals: Know the definition of a binomial distribution.

Skills:

- **Know the four assumptions.** The binomial distribution requires four assumptions: 1) There must be only two outcomes. 2) Trials must be independent of one another. 3) There must be a fixed sample size, chosen ahead of time and not determined while the experiment is running. 4) The probability of success is the same from trial to trial. Number 2 is the most difficult to check, so it is usually simply assumed. Number 4 rules out finite populations, where the success probability depends on what is left in the pool. Number 3 rules out situations where the experiment continues until some event happens.

- **Random values can be generated using MATH PRB randBin(.** The usual cdf method of generating random values is useless here because we don't have an explicit invertible formula for the binomial cdf. Thus we rely on other mechanisms. One could generate a series of Bernoulli trials and add the 0's and 1's together. The easiest way, however, is

MATH PRB randBin(*n*, *p*, *r*), where *n* is the sample size, *p* is the probability of success, and *r* is the number of values required.

Reading: Section 3.4.

Day 22

Activity: Binomial.

Today we will explore calculating binomial probabilities. The difficulties come from the discrete nature of the graph, and the issue of the height of the steps. You must be careful with "less than" and "less than or equal to" in your calculations.

Using DISTR binompdf(and/or DISTR binomcdf(, calculate these probabilities:

1) What is the chance of 4 or fewer successes with $n = 10$ and $p = 0.4$?

2) What is the chance of 3 to 10 successes, inclusive, with $n = 20$ and $p = 0.7$?

3) What is the chance of more than 6 successes with $n = 15$ and $p = 0.2$?

Our last item of business with the binomial is the mean and variance calculation. I will go through the "trick" in class, which is just a few algebra steps, but it might throw some of you. I suggest paying attention carefully, and trying to focus on the overall ideas instead of the details. Then later, knowing what the purpose of each step is you can pay attention to the details. This trick will appear in other distributions, so it is worth knowing about.

Goals: Become familiar with the important binomial distribution.

Skills:

- **Become familiar with the TI-83 commands to calculate binomial probablities.** Our TI-83 has two functions, DISTR binompdf(and DISTR binomcdf(which allow us to calculate binomial probabilities. As their names imply, one is cumulative and the other is marginal. With sufficient memory, one could resort to the actual formula and use a sequence command to accomplish the same thing. In particular, you can **verify** that you are using the commands correctly by memorizing a small check example. The command syntax is DISTR binompdf(*n*, *p*), which will give the whole pmf as a list, or DISTR binompdf(*n*, *p*, *x*), which will give just the probability of exactly *x* successes. For DISTR binomcdf(*n*, *p*, *x*), we have the cumulative probability below (and including) *x* successes.

- **Know the mean and variance for the binomial.** For each distribution we encounter, we will want to know the mean and variance. These become useful later when we examine the Central Limit Theorem. For the binomial, the mean is np and the variance is $np(1 - p)$.

Reading: Section 3.5.

Day 23

Activity: Hypergeometric Distribution.

All hypergeometric problems can be thought of as "lottery" problems. (Leonhard Euler was a Swiss mathematician and physicist who lived during the eighteenth century. In 1749, the King of Prussia asked Euler to study a proposed lottery system, and the results of Euler's analysis were published in 1769. Euler's paper included a special series from which the hypergeometric distribution was later derived.) (Reference) The winning balls are our "successes"; the losing balls are the "failures". In this respect, the hypergeometric is like the binomial. However, balls are chosen **without** replacement, and this is where the two distributions differ. I will do three types of examples: quality control, poker, and a large sample binomial.

Is my program generating hypergeometric data? You will check to see if it is close to theory. Warning: this will be difficult to answer unequivocally, due to the "Law of Small Numbers". In your groups, generate a number of observations with the program, using some parameters you choose, and then decide on some devices (graphs, calculations, etc.) to compare the program output with theory. At the end of the period, we will share notes, so be prepared in your group to summarize your findings.

Goals: Understand the Hypergeometric distribution.

Skills:

- **Recognize situations where the hypergeometric distribution is an appropriate model.** The example I think of when imagining the hypergeometric distribution is the decision to accept or reject a shipment of goods. Another good example is the full house calculation for poker. In any case, what we are modeling is a binomial-like situation, with successes and failures. But see the next item.

- **Know how the binomial and the hypergeometric are related.** The key difference is the hypergeometric comes from a finite population. If one imagines balls in an urn, then the binomial is a hypergeometric using a **huge** urn. (Of course, we're talking about limits.)

- **Know the mean and variance for the hypergeometric.** Because of the close relationship with the binomial, we can use the corresponding formulas, with a slight modification. If we let $p = M/N$, and if we use the **finite population correction** factor $(N-n)/(N-1)$, we see the mean and variance are easy to remember. For the hypergeometric distribution, the mean is nM/N and the variance is $(N-n)/(N-1)nM/N(1-M/N)$.

- **Use HYPER to generate random values from a hypergeometric distribution.** The program HYPER will generate random data which follows the hypergeometric distribution. The program is simple; it creates a cdf and compares it to a random number MATH PRB rand until the cdf exceeds MATH PRB rand. That value is the one reported. In the program, M is the population size (N), K is the number of successes in the population (M), N is the sample size (n), and R is the number of random numbers desired.

Reading: Section 3.5.

Day 24

Activity: Negative Binomial Distribution. **Homework 5 due today.**

I will derive the pmf for the Negative Binomial by a counting argument. This will be one of those few times where I think the **derivation** of a distribution is one you should memorize. If you know how the formula is developed, you will know the pmf without having to memorize a formula. The essence of this proof is filling in the slots of a sequence of successes and failures, but requiring the last one to be a success. It is then just a matter of using the right binomial coefficient and multiplying by the right number of p's and $(1-p)$'s.

How can we generate random data from this distribution? Will a program similar to HYPER work? Or does the infinite support worry us? Can we simply flip coins (with our calculator)? After we consider these ideas, I will share the two programs I wrote.

Goals: Understand the Negative Binomial distribution.

Skills:

- **Relate the Negative Binomial to the Binomial and the Hypergeometric.** The negative binomial differs from the binomial because the number of trials isn't known beforehand; we continue the experiment **until** r successes occurs. The negative binomial differs from the hypergeometric because the probability of a success on any trial, p, is fixed, as in the binomial.

Reading: Section 4.3.

Day 25

Activity: Normal Approximation to the Binomial.

As we have seen, the pmf for the binomial looks very similar to the pdf for the normal curve. In fact, with appropriate adjustments for discrete versus continuous we can use the normal curve as an approximation for the binomial pmf. In the past, this was a more important issue than it is now with the computing power we have available. As you know, in the binomial pmf we calculate $C(n,r)$. If n is large, $C(n,r)$ may be too large for our computers to calculate. The normal approximation helps us in such cases. Using the mean and variance for the binomial (that you have memorized) we can just use the normal cdf for our probability calculations. However, there can be an error according to the diagram of the two functions, which we will explore in class. The modification we will make to overcome this error is the **continuity correction**.

Calculate the chance of getting between 40 and 50 heads on 100 tosses of a fair coin. Use the binomial. Then use the normal approximation.

Now try the chance of exactly 50 heads on 100 tosses.

Now redo the first one using the continuity correction.

Goals: Understand the details of using the normal curve to approximate binomial probabilities.

Skills:

- **Know why we add or subtract ½ before doing the normal approximation to the binomial.** We can use the normal curve to approximate binomial probabilities. But because the binomial is a discrete distribution and the normal is continuous, we need to adjust our endpoints by ½ unit. I recommend drawing a diagram with rectangles to see which way the ½ unit goes. For example, calculating the chance of getting 40 to 50 heads inclusive on 100 coin flips requires using 39.5 to 50.5.

Reading: None

Day 26

Activity: Cauchy Distribution.

The last pdf we will explore is a counter-example to some theorems. The Cauchy random variable (which has a cdf shaped like arctangent) is the ratio of two standard normal numbers. It has very unusual behavior though. (The unusual behavior seems to not have been noticed until 1824, by Poisson.) While its pdf is symmetric, it has no mean. This is a curiosity of calculus results; while it surely has a median, the “center of mass” integral is actually undefined. This behavior is apparent after a long string of outcomes. Now and then, extreme outliers occur. Even worse, if you keep a cumulative average, it will not seem to approach any fixed value; it wanders around aimlessly. This is in direct contrast to the results of the Central Limit Theorem (Days 27 and 28) because the Cauchy doesn’t satisfy one of the hypotheses. One lucky group will further examine this distribution in the next presentation (Day 33).

Calculate the ratio of two standard normal random numbers. Generate 100 values and compare to the normal distribution. Repeat, calculating the mean and variance each time for the 100 values.

Goals: Examine Cauchy distribution.

Skills:

- **Know the curious features of the Cauchy distribution.** The most unusual characteristic of the Cauchy distribution is the lack of a mean. This paradoxical result is due to improper integrals and limits. The cdf of the Cauchy distribution is the arctangent function, appropriately shifted.

Reading: Sections 5.3 to 5.5.

Day 27

Activity: Linear Combinations. Central Limit Theorem exploration.

An important aspect of mathematical statistics is what happens to random variables when they are added together in linear combinations. We see linear combinations most often when we do averages. In fact, averages are very common, and the resulting theorem about their sums' distribution is **the** theorem in statistics, the **Central Limit Theorem**.

Our first exploration today will be about the formulas for means and variances of linear combinations. Basically, we are talking about combining information by adding and subtracting, and multiplying by constants. An example of a linear transformation is the Fahrenheit/Centigrade conversion $F = (9/5)C + 32$. An example of a linear combination is total profit for a company that makes \$10 each for X chairs and \$30 each for Y tables; their profit would be $10X + 30Y$. The main example I will use is the average of a sample of measurements. This derivation will lead to the means and variances that are critical to the Central Limit Theorem. The main assumption we need is independence.

Our second exploration today will be about the shape of sampling distributions for sums (or averages), which is the essence of the **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. (The CLT was first discovered by Laplace, in the early 1800's.)

One of the problems I see students have when trying to apply 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** probability problems are really Central Limit Theorem problems, but some of them are just samples of size 1 ($n = 1$). Then the calculation for the spread of an average ($s_{\bar{x}} = \frac{s}{\sqrt{n}}$) always works.

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 graph 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 that for the average, you **divide** by the square root of n whereas for the sum you **multiply** by it. This is how I remember which is which: sums are larger than averages.

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 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 people around you, 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 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 generate 12 random **normal** numbers on your calculator using MATH PRB randNorm(65, 5, 12). Add them together and record the result; repeat 100 times. Use 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 these examples) and the number of times the dice were rolled or rand was pressed (called the *sample size*, denoted n).

Goals: Examine histograms 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:

- **Know how to combine Expected Value and Variance for Linear Combinations of Random Variables.** We often are interested in linear functions of two random variables. For instance, we may know the mean and variance of times for workers traveling to a job site. We may also know the mean and variance of the length of the job. What we **really** want to know is the total length of time the workers are gone from the main plant.

- **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: Section 5.3 to 5.5.

Day 28

Activity: Practice Central Limit Theorem (CLT) problems. We will have examples of non-normal data and normal data to contrast the diverse cases where the CLT applies. **Homework 6 due 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.

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 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?

4) 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?

5) The number of boys in a 4-child family can be modeled reasonably well with the binomial distribution. 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 **family** having 3 kids. However, we could calculate quite accurately the number of kids in **100** such families.

Reading: Chapters 3 and 4 and Sections 5.3 to 5.5.

Day 29

Activity: Catchup

We will practice more CLT problems, and you can work on Presentation 3, which explores the CLT.

Reading: Chapters 3, 4, and 5.

Day 30

Activity: Exam 2.

This second exam is on particular distributions, both discrete and continuous, and the idea of sampling distributions. You should know facts about each distribution we have encountered. You should be able to generate random data from each distribution. You should know the strengths and limitations of each of them. The Central Limit Theorem results WILL be on the exam.

Reading: Section 7.1.

Day 31

Activity: Guess m&m's percentage. What fraction of m&m's are blue or green? Is it 25%? 33%? 50%? We take samples to find out.

Unit 3 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** a single-number guess isn't right though; what is the probability that our sample 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.

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 Section 4.7. 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.

Goals: Introduce statistical inference - Guessing the parameter. Construct and interpret a confidence interval.

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.

Reading: Section 7.2.

Day 32

Activity: Explore Confidence Interval interpretation.

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, 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.

Now we will explore how changing confidence levels and sample sizes influence CI's. Complete the following table, filling in the **confidence interval width** in the body of the table. Use STAT TEST 1-PropZInt but in each case make x close to 50% of n. (The calculator will not let you use non-integers for x; round off if needed.)

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: 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 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 (undesirable), and larger sample size makes narrower intervals (desirable).

- **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 37).

Reading: None

Day 33

Activity: Presentation 3.

> Central Limit Theorem (Section 5.4)
>
> Choose from among the following distributions: Exponential, Normal, Cauchy (**CAUTION: this distribution violates the CLT hypotheses**), Uniform, Bernoulli, Problem 3.95's distribution.
>
> Generate a sample of 5 items. Calculate the mean and record the result. Repeat several hundred times. Finally, graph the quantile plot of the list of means from your several hundred samples.
>
> Repeat for a sample of 50 items. Repeat for a sample of 200 items. Compare all of your results to the theory from the CLT.

Reading: Section 8.1.

Day 34

Activity: Argument by contradiction. Scientific method. Type I and Type II error diagram. Courtroom terminology.

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.

***P*-value.** There are two definitions for the P-value. Definition 1: The P-value is the smallest alpha level that will cause us to 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 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.

Goals: Introduce statistical inference - Hypothesis testing.

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 do generally not control this second type of error; the sample size is the determining factor here.

- **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 actually 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.

Reading: Sections 7.3 and 8.2.

Day 35

Activity: Practice problems on hypothesis testing. Use z-test as an example. **Homework 7 due today.**

One difficulty I see students having 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 two 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 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, as explained on Day 34. 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 36.

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: Practice contradiction reasoning, the basis of the scientific method.

Skills:

- **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, 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: Section 8.4.

Day 36

Activity: Testing Simulation.

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. 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” or “I believe you did **not** use 20”.

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 μ =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.

In addition to the simulation above, which uses the TI-83’s built in routines, you should also be able to calculate the observed probabilities we saw in our chart exactly. For example, what **is** the probability that the sample average will cause you to reject the null hypothesis that the mean is 20 when the true mean is really 24? These calculations are known as power calculations, and are a vital part of the choice of a test or the design of an experiment. For if you only had a very small chance of detecting an important difference, then the procedure was really a waste of resources; you weren’t going to reject anyway. Typically, these explorations will become sample size problems.

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 randomness in the sampling method 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.

Goals: Interpret significance level. Observe the effects of different values of the population mean. Recognize limitations to inference.

Skills:

- **Interpret significance level.** Our value for rejecting, usually 0.05, is the fraction 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 does not even measure whether our null hypothesis is true! It **only** measures whether random data could look like the observed data, under the assumption that the null hypothesis is in fact true.

- **Understand how the chance of rejecting the null hypothesis changes when the population mean is different from 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.** Often users of statistical inference techniques will blindly use them without checking the implicit hypotheses. 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.

Reading: Section 8.2.

Day 37

Activity: Gosset simulation. **Homework 8 due today.**

Today, a little history into the problem discovered over a century ago with small samples. William Sealy Gosset noticed that for small samples, his confidence intervals were not as correct as predicted. For example, he was supposed to capture the parameter in his intervals 95% of the time with samples of size two 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: Know about the t-test.

Skills:

- **Know why using the *t*-test or the *t*-interval when σ is unknown is appropriate.** When we use s instead of σ and use the z distribution, we find that our confidence intervals are too narrow, and our hypothesis tests reject H_0 too often. We must pay a "penalty" for not knowing the true population standard deviation, in the form of the appropriate t distribution.

- **Realize that the larger the sample size, the less serious the problem.** 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: Sections 8.2.

Day 38

Activity: Continue hypothesis testing problems.

Note the correspondence between confidence intervals and hypothesis testing. For example, a 95% confidence interval captures the parameter 95% of the time. If the null hypothesis value is in the confidence interval, then we would fail to reject that null hypothesis, as it is considered plausible.

Practice framing and solving hypothesis testing problems.

Goals: Explore the correspondence between intervals and tests.

Skills:

- **Know the one-to-one correspondence between confidence intervals and hypothesis tests.** We can show that for a fixed confidence level, the corresponding hypothesis test would be rejected if the null value is **not** in the interval. Thus, given a confidence interval, one can perform the corresponding hypothesis test. Note however that a P-value cannot be calculated; the fixed alpha level test is as much as we can perform.

Reading: Sections 7.2 and 8.3.

Day 39

Activity: Proportions.

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. Keep in mind as we do this the four assumptions of the binomial model, and whether they are truly justifiable.

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. 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. Using the "Plus 4 Method", even 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 the 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. The typical rule of thumb is to make sure there are at least 5 successes and at least 5 failures in the sample. For example, in a sample of voters, there must be at least 5 Republicans and at least 5 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 5 blue or green m&m's, I made you take more until you had at least 5.)

- **Know the Plus 4 Method.** A recent (1998) result from statistical research suggested that the typical normal theory failed in certain unpredictable situations. Those researchers found 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, it is the recommended approach in all **confidence interval** problems. Hypothesis testing procedures remain unchanged.

Reading: Chapters 7 and 8.

Day 40

Activity: Review. **Homework 9 due today.**

Review Inference.

Review.

Goals: Conclude course topics. Know everything.

Reading: Chapters 7 and 8.

Day 41

Activity: Presentation 4.

Statistical Inference (Chapters 7 and 8)

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.

Goals: Conclude course topics. Know everything.

Reading: None

Day 42

Activity: Exam 3.

This last exam covers statistical inference, including the *t*-tests and intervals and the *z*-tests and intervals for proportions in Chapters 7 and 8.

Arizona Temps Dataset:

	Flagstaff					Phoenix				
Date	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

Return to Chris' Homepage

Return to UW Oshkosh Homepage

Managed by: chris edwards

Last updated January 10, 2015