Chapter 1:
Why the Social Researcher Uses Statistics

- Why Statistics? Why are you here?
  - The most common response to this question is because you have to be; that is, because the class is required.
  - It is true that in virtually all undergraduate and graduate-level areas of study, statistics is a required field of study.
  - The question that I am asking is broader. In particular, I am asking you to consider *why this university (and others) believe that the study of statistics represents an essential component of your educational experience?*

Chapter 1:
Why the Social Researcher Uses Statistics

- I believe there are several reasons for the requirement.
  - As citizens we are exposed to statistics constantly (i.e., in marketing surveys, voting polls, newspapers, magazines)
  - By learning statistics, you will become a more informed/sharper consumer of statistical material. Often it is said that there are lies, damn lies, and statistics (who said that?).
    - News reports of the 400% rise in heroin use among middle school children.
  - In your professional life, no matter what it may be, you may be expected to interpret and present statistical information. This is particularly true in the field of criminal justice and criminology.

Chapter 1:
Why the Social Researcher Uses Statistics

- The Nature of Criminal Justice Research
  - Unlike many disciplines the use of statistics is particularly important for criminal justice practitioners. There are so many examples of how statistics are used in the field of criminal justice.
    - For example, it is common for both researchers and practitioners to deal with Crime Rates, Incarceration Rates, Recidivism Rates.
    - What is a crime rate? How do you interpret a crime rate? Why are they useful?

Chapter 1:
Why the Social Researcher Uses Statistics

- What are statistics?
  - Broadly speaking, statistics is a set of procedures used by social scientists to organize, summarize and communicate information about the world around us.
  - As I mentioned briefly last class, we will treat statistics as a set of tools; tools which we employ to answer specific questions of interest. We will have to do many calculations, but it is essential that you recognize that we are really just trying to answer questions.
    - For example, crime rates. What are the national crime trends in violent crime over the past ten years? How do you know? What about Boston? Lowell?
  - One way to assess these patterns is by personal experience. Clearly, this is our first means of understanding the world around us. But is it the best/most accurate? Not always, and that is where we can benefit from statistics. The benefit of empirical knowledge over personal experience will be a recurring theme throughout the class. If you ask many people about social patterns, particularly as they relate to crime, their only barometer is personal experience.
  - In this class, we are going to avoid making such rash generalizations, and take a more balanced and evidence-based approach.

Chapter 1:
Why the Social Researcher Uses Statistics

- Statistics can generally be divided into two categories. When we compute statistics, they will be used for either their descriptive or their inferential qualities.

- Descriptive (first module of the class)
  - Generally, descriptive statistics are used to summarize (or describe) data. This family of statistical techniques will allow us to gather basic information from an array of raw data. These procedures will allow us to describe a set of numbers more clearly. This is a process of “boiling down” a large group of data into a more understandable format.
  - Of most interest in this class will be frequency tables, measures of central tendency, and measures of dispersion (or variability).

Chapter 1:
Why the Social Researcher Uses Statistics

- Inferential (second and third modules of the class)
  - This family of statistical techniques are used to make predictions about the entire population. As you will see, one of the strengths of statistics is that it will allow us to make informed decisions about the entire population using data collected only for a sample.
  - In this class, we will learn to calculate confidence intervals, chi-square, ANOVA, correlations and partial correlations, and regression analysis. Each of these measures has unique strenghts (and limitations), which we will discuss.
  - The will help us to improve our understanding of the broader population, using data for a small subset of that population.
  - As such, the manner in which data is collected is directly linked to how we analyze those data.

Chapter 1:
Brief Review of Research Methods

- It is important that we all begin with a review of a number of key terms that we will use throughout the semester. Many of these concepts will surely be familiar, even if their meanings are a little fuzzy.
  - Variable
  - Independent Variable (often symbolized as X)
  - Dependent Variable (often symbolized as Y)
  - Distribution
  - Frequency Distribution

Chapter 1:
Brief Review of Research Methods

- Variable Types
  - Discrete-these are the variables that can take only specific values, that is they cannot be subdivided (i.e., number of cars, number of siblings, etc.).
    - Values cannot be compared numerically
    - Tell us “what kind” or “what type”
    - Refers to categories
  - Continuous-these are variables that refer to quantities and can be divided into smaller categories. That is, they have a nearly infinite set of possible values.
    - Values can be compared numerically
    - Tell us “how much” or “how many”

Chapter 1:
Brief Review of Research Methods

- Levels of Measurement
  - Nominal
    - Variables measured at this level indicate only that there is a difference between categories.
    - Categories must be both mutually exclusive and exhaustive.
  - Ordinal
    - Variables at this level have the property of magnitude, that is, there is a clear difference between categories and these categories can be ranked.
  - Interval/Ratio
    - Variables at this level have everything from the previous two levels, that is they represent distinct categories, they can be ranked, and the precise distances between categories can be determined

Chapter 1: Brief Review of Research Methods

- Units of Analysis
  - This concept refers to the unit that constitutes an observation in our data
  - This is the item for which we have data. The units of analysis can be individuals or higher levels of organization:
    - Persons
    - Cities
    - Prisons
  - Often macro units are simply aggregations of smaller units

Chapter 1: Introductory Descriptive Measures

- For the purposes of descriptive analyses, variables can be presented in a number of different ways. The primary objective is to provide information about a variable efficiently, that is in a way that is simple to interpret.
- The presentation of counts, or the number of times a given event occurred is a common way for data to be presented. Typically count-data are listed as a frequency distribution.
  - A limitation of reporting counts is that they do not take into consideration the underlying population for a given area. That is, a simple comparison of the counts does not offer information differences in the risk (or hazard) of the event taking place across areas.
- Proportions represent the share of the total sample size (or population) that is found in a given category.
  P=f/N
  where:
  f=frequency of cases in a particular category
  N=total number of cases

Chapter 1:
Introductory Descriptive Measures

- Percentages are calculated by multiplying proportions by 100. In other words, the numeric values for proportions and percentages are the same, however, the latter are easier to interpret:
  %=(f/N)*100

- Ratios are measures that provide information of the size of one category *relative* to one another. Ratios are computed by dividing the number of observations in category “x” by the number of observations is category “y.” When interpreting ratios, keep in mind that the value is a comparison of the relative sizes of the given categories. That is, the resulting ratio indicates how much larger (or smaller) one category is than the other.

- Rates represent one of the most common ways in which crime data are reported. Rates provide information about 3 key pieces of information:
  1. a specific event
  2. a particular geographic area
  3. a particular period of time

- Further, many times in criminal justice/criminology research rates are discussed as the potential, or hazard associated with a particular event taking place.

Chapter 1:
Introductory Descriptive Measures

- Rates are computed by dividing the number of times a given event has taken place by the at-risk population and then multiplying this value by a constant. Often crime rates use a value of 100,000 as the constant.

  Rate=(# of actual events/# at-risk units)*K

- Note that the constant value has the effect of "standardizing" rates, meaning that it allows the rates to be compared across areas that have different at-risk population sizes.
- Although they are very useful measures, in part because they permit comparison across jurisdictions, rates are limited in some respects. How so?
- Percent change, or rate of change, is also a measure that is commonly used by researchers because it provides information about how much a given phenomenon has changed over a period of time:
  % Change=$((F_2-F_1)/F_1)*100$
  where:
  F_1=# of times the event happened at Time 1 (i.e., the earlier period)
  F_2=# of times the event happened at Time 2 (i.e., the later period)
- When interpreting percent change, positive values indicate that an event increased over time, while negative values suggest that the event decreased over time.

Chapter 1: Frequency Distributions

- There are two type of frequency distributions that we will work with in this class: Simple Frequency Distributions and Grouped Frequency Distributions.
  - Simple Frequency Distributions report information for every observation in a given distribution. Simple frequency distributions are very common, as they provide counts of how many times a given value occurs in a distribution.
  - Grouped Frequency Distributions are useful for many situations researchers experience in the social sciences. In particular, this type of reporting method is useful when the data are spread over a wide area, making it difficult to arrange them neatly into a simple frequency distribution. Furthermore, in such a case, it will likely also be difficult to make sense of the general patterns in the data.
  - A grouped frequency distribution is one where we divide the distribution into equally sized intervals, which will make it easier to interpret the distribution. We can also use grouped frequency distributions for a variety of different analyses.

Chapter 1:
Frequency Distributions

- There are a few pieces of information to keep in mind with respect to grouped frequency distributions:

  Substantively

  - They do not provide information about each specific observation, but rather only how many scores fall in a given range. As such, when they are used, researchers are forced to sacrifice precision in understanding the true underlying distribution.
  - However, this sacrifice is made for practical purposes. The precision is lost in order to gain clarity, that is, to achieve a better overall idea about patterns in the distribution. It would be difficult, or impossible, to describe or analyze the data using a simple frequency distribution.

  Practically

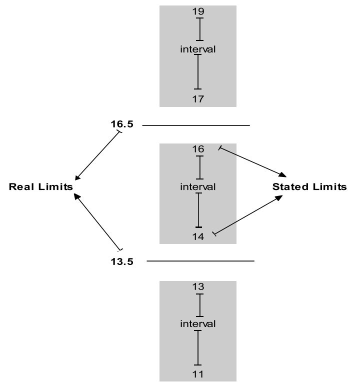
  - When constructing a grouped frequency interval, the intervals must satisfy the following conditions:
    1. Intervals must be of equal size
    2. Intervals must be mutually exclusive and exhaustive
    3. Intervals must begin with whole numbers

Chapter 1:
Frequency Distributions

- There are a number of calculations that we are able to perform using grouped frequency distributions. In a sense, these calculations are used as a means of obtaining some of the more precise information about the distribution that is lost when we use grouped frequencies rather than simple frequency distributions.
- Before introducing some of the calculations that are possible using frequencies, it is important to first introduce a number of concepts.
- The first has to do with the idea of interval limits. There are two different types of limits that we will have to be concerned with: Stated Limits, and Real Limits.
  - Stated limits are the interval boundaries determined by the researcher when constructing the frequency distribution. The stated limits are the high and low values for each interval.
  - Real limits are conceptual values, which are found at the half-way point between successive intervals. The real limits are useful because they allow researchers to consider values that may fall just beyond the stated limits of a give interval.

Chapter 1:
Frequency Distributions

- What does that mean? Funny you should ask. Below is a diagram that illustrates the difference between Stated and Real interval limits. For this example, imagine that the class interval we are interested in contains all values in a distribution between 14 and16.



Chapter 1:
Frequency Distributions

- Note that for our purposes, each interval will have an Upper Stated Limit (USL) and a Lower Stated Limit (LSL). From our example, these values are 16 and 14, respectively.
- Further, each interval will also have an Upper Real Limit (URL) and a Lower Real Limit (LRL). Determining these values is straightforward. In this class, we will simply use the following equations:

  URL=USL+.5
  again following this example URL=16+.5, or 16.5

  LRL=LSL-.5
  or LRL=14-.5, or 13.5

Chapter 1:
Frequency Distributions

- The second concept that needs to be discussed is the idea of cumulative frequency (cf).
- Cumulative frequency is a manner of presenting information, that provides a fuller picture about the overall distribution. Computing cumulative frequency is a straightforward process.
- The value of including information about cumulative frequency is that it tells us how many cases fall ***at or below*** a given interval. This is useful because it provides information about the relevant sizes of the categories, or how observations are distributed across the various interval categories.
- In addition to cumulative frequency, it is also useful to include information about the cumulative percent (c%). These values are computed using the same technique, and provide information about the percent of all observations that fall ***at or below*** a given interval.
- The method used to compute both cf and c% is called the diagonal method.

Chapter 1:
Cumulative frequency and Cumulative Percent

- Note that before computing cf or c%, the categories need to be organized in descending order.

# prior convictions	f	cf	%	c%
14	4	**25**	16	**100**
13	3	21	12	84
10	5	18	20	72
8	3	13	12	52
7	6	10	24	40
6	2	4	8	16
5	1	2	4	8
3	1	1	4	4
Totals	**25**		**100**	

- Also, as a check, the values at the top of the cf and c% columns, should be equal to the total sample size and 100%, respectively.

Raw Score to Percentile Ranks

- Percentile Ranks tell us the percent of observations that fall at or below a given score. Common uses of percentile ranks are national standardized tests such as the SAT, GRE, LSAT, etc.
- In essence, percentile ranks indicate how one score ranks with respect to the entire distribution of values, or in relation to other values. Typically, percentile ranks are given in deciles or quartiles, but it is possible to compute a percentile rank for any value in a distribution.
- Keeping this in mind, statisticians developed a procedure that will allow a researcher to derive a percentile rank from a raw score. To do this, we use the following equation:

$$PR = L + [(\frac{score - LRL}{h}) * I]$$

where
L=C% in the interval **below** the target interval
Score=raw score you are converting to PR (This value is given to you in the problem)
LRL=Lower real limit of the target interval
h=interval size (USL-LSL+1)
I=% of cases within the target interval

Percentile Ranks to Raw Scores

- It is also possible to derive a raw score from a percentile rank. To do this we will use the following equation:

$$Score_p = LRL + [(\frac{pN - SFB}{f}) * h]$$

where
LRL=Lower real limit of the target interval
p=percentile rank (again this value will be provided in the problem)
N=sample size
SFB=the cf value in the interval **below** the target interval
f=the frequency (or the number of observations within the target interval)
h=interval size (USL-LSL+1)

Chapter 5:
Probability and the Normal Curve

- Understanding probability is the key to inferential statistics.
  - Probability allows us to assign a numerical value to the likelihood that a given outcome will be observed.
  - It is important to keep in mind that probability is a *theoretical* concept; that is, probabilities tell us what we would expect to find in the long run, but does not guarantee any particular outcome.
    - Probability is calculated as follows:
      - $P(A) = \frac{\text{\# times event A can occur}}{\text{total number of outcomes}}$

Rules of probability

- Bounding Rule:
  - This rule specifies that all probabilities are bound between 0 and 1.
    - A zero probability indicates that there is no chance that a given outcome will happen;
    - A probability of 1 indicates that we can be certain that a given outcome will occur.
  - Probabilities are often expressed as percentages, but technically, they are best expressed as decimals.

Rules of probability

- Rule of the Complement/ Converse Rule
  - This rule states that if we know the likelihood of success of a given outcome, we can calculate the likelihood of failure (or non-success):
    - P(F)=1-P(A)
    - Note that the sum of the probability of success and the probability of failure always equal 1
  - Applying the same logic, we can also calculate the likelihood of success if we know the probability of failure for a given outcome.

Rules of probability

- Addition Rule
  - This rule is applied when we want to calculate the likelihood of an outcome when there is more than one result that will be considered a success.
    - In such situations the probability of success is equal to the *sum* of the respective probabilities.
      Example: What is the probability of drawing the Jack of Spades *or* the Jack of Diamonds?

      P(S)=1/52 + P(D) = 1/52 or

      .0192 + .0192 = .0385

    - This tells us that we have approximately a 4% chance that we will draw the Jack of Spades or Diamonds from a deck of cards.

Rules of probability

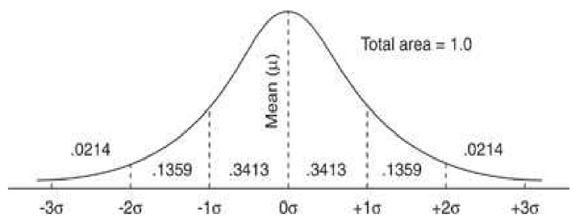
- Multiplication Rule
  - This rule is applied when we want to calculate the likelihood of outcomes in succession.
    - In such situations the probability of success is equal to the *product* of the respective probabilities.
      Example: What is the probability of drawing the Jack of Spades *followed by* the Jack of Diamonds?
      P(S)=1/52 * P(D) = 1/52 or
      .0192 * .0192 = .0004
    - This tells us that we have approximately a .04% chance (1 in 2,500) that we will draw the Jack of Spades and the Jack of Diamonds in two successive draws from a deck of cards.

The Normal Curve

- The normal curve, as you will see, provides a foundation for many of the inferential statistical calculations we will cover in this class.
- Before introducing some of the ways that we make use of the normal curve for statistical purposes, it is important to note its properties.
- There are three characteristics of the normal curve that make it unique:
  - The distribution is symmetrical, that is, the mean, median, and mode are identical
  - Because the normal curve is a theoretical construct, its tails extend infinitely in both directions. Put another way, the tails of the distribution never cross the x (horizontal) axis.
  - Also note that when we discuss a theoretical (or probability) distribution we use different symbols. We symbolize the mean with the Greek symbol mu (μ) and for the standard deviation we use the symbol baby sigma (σ).

Area under the normal curve

- 100% of all possible values fall beneath the normal curve
- A constant proportion of observations fall within a given distance from the mean.
- The image below illustrates the proportion of the distribution that falls with one, two, and three standard deviations from the mean:
  - +/- 1σ=68.26%
  - +/- 2σ=95.44%
  - +/- 3σ=99.74%



Area under the normal curve

- Because of the characteristics of the normal curve, we can also determine the area under the curve between any two points.
- However, in order to determine this percentage, we will need to convert a raw score into a standard scores, or a z-score.
- A z-score indicates how far a given observed value falls from its respective mean.
- The formula for calculating a z-score is as follows:

$$z = \frac{(x - \mu)}{\sigma}$$

Area under the normal curve

- In most cases, observed values do not fall at even integers away from the mean. In such cases, we will have to compute a z-score and then look up the percentage that corresponds to this value using Table A in the back of the book.

- As an example, we will use data regarding the IQ distribution in this country. The mean of the distribution (μ) is 100 and the standard deviation (σ) is 15.

- Imagine we wanted to know the percent of the population that has an IQ somewhere between 100 and 119.

Area under the normal curve

- For this calculation first we will have to convert the observed IQ value (119) into a z-score:

$$Z = \frac{119 - 100}{15} \qquad Z = \frac{19}{15} \qquad Z = 1.27$$

- Next, we have to refer to Table A in the back of the text, and find the percent associated with this value. The table indicates that 39.8% of the distribution falls within 1.27 standard deviations beyond the mean.
- Substantively, this means that 39.8% of the population has an IQ between 100 and 119.

Area under the normal curve

- There are a number of different permutations on this type of problem. Below I list a number of things to keep in mind when seeking to determine the area under the curve between *any* two points.
- First, 100% of all values fall under the curve, 50% above the mean and 50% below the mean.
- If the question asks for us to compute the area between two points and those two points are on opposite sides of the mean, we get the total area by summing the respective percentages.
- If the question asks for us to compute the area between two points and those two points are on the same side of the mean, we get the total area by subtracting the smaller percentage from the larger percentage.
- Finally, when we treat the normal curve as a probability distribution, we can determine the *probability* of finding a value that falls within a given range of the distribution by dividing the percentage by 100.

Chapter 6:
Samples and Populations

- As the authors describe, social scientists rely predominantly on sample-based data in their research.
- There are a number of reasons for this, most notably:
  - It is difficult, or effectively impossible, to obtain a full and complete list of all members of a population.
  - Collecting data for an entire population would be expensive.
  - Even if it were possible, collecting data for an entire population is time consuming, which is problematic because timeliness is an important consideration for much research.

Random and Non-random sampling

- Sampling, is the process whereby a researcher collects and analyzes data for a subset of a population. How those data are collected has direct implications for whether the information can be used for inferential purposes.
- Non-random sampling techniques are those in which all members of a population do not have an equal probability of being selected for inclusion. Examples of non-random procedures typically used are as follows:
  - Convenience Samples
  - Quota Samples
  - Snowball Samples

Random and Non-random sampling

- By contrast, in order for researchers to use the sample data to draw conclusions about the broader populations, researches need to employ random selection processes.
- Specifically, random sampling techniques are those in which all members of a population have an equal probability of being selected for inclusion. Examples of random procedures typically used are as follows:
  - Simple Random Samples
  - Stratified Samples
  - Multi-staged, Stratified, Cluster Samples (don't worry about this type)
- Random Samples are preferred because, in the long run, they are more representative of the population than non-random samples

Sampling Error

- Although researchers must rely on sample data, it has been shown that sample characteristics, such as the mean, are not likely to be identical to the corresponding population characteristics.
- For example, it is generally the case that:

$$\bar{x} \neq \mu$$

- This observed disparity, that is, the difference between the sample and population values is known as sampling error.
- Essentially, sampling error is present regardless of how careful a researcher is in designing a study.

Central Limit Theorem

- The above discussion indicates an inherent difficulty:
  - Researchers are forced to relay on sample based data to draw conclusions about the population.
  - However, due to sampling error, the sample characteristics are not likely to match the population characteristics of interest.
- The central limit theorem provides the link which allows us to use the sample data for inferential purposes.

Central Limit Theorem

- The central limit theorem is a theoretical construct that tells us:
  - If we were to take an infinite number of samples of size "N" from any population, and then create a distribution of a population characteristic of interest, say the mean, the result is what is called the sampling distribution.
  - The sampling distribution has a number of important characteristics:
    - The distribution is normal in shape.
    - The mean of the sampling distribution, that is, the mean of means, is equal to the true population mean (μ).
    - The standard deviation of the sampling distribution is called the standard error, and is smaller than the population standard deviation.

Sampling Error

- Class average (μ)= 23.21 years
- Number samples (sample size=6):
  - 50 samples= 23.67
  - 100 samples= 23.49
  - 500 samples= 23.25
  - 1000 samples= 23.22
- As is evident in this example, as the number of samples increases, the mean of the sampling distribution approaches the population mean.
- Also note that the mean of the sampling distribution converges with mu after a relatively small number of samples are drawn.

The sampling distribution of means

- Because the sampling distribution of means is normal in shape, we can measure the area under the curve between any two points, using a technique very similar to the ones learned in the previous chapter.
- However, because the distribution is not comprised of raw scores, but rather sample means, we are going to calculate a standard score (z-score) that indicates the distance from the mean in standard errors.
- Keep in mind that the calculation is going to be nearly identical in structure to the one we used previously, but it differs conceptually.
- It is also important to remember that the sampling distribution is also a probability distribution, and as such, we can see that it is more likely (i.e., higher probability) that sample means will fall near the population mean, and less likely (i.e., lower probability) that a sample mean will fall far from the population mean simply by chance.

Z-scores from sampling distribution

- To calculate a z-score from a sampling distribution, we use the following equation:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

- As you can see, the numbers in the numerator will be provided as part of the problem. However, we will have to generate the denominator, which is the standard error of the mean using the following equation:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

Z-scores from a sampling distribution Example

- Following an example provided by the authors, imagine that a report indicated that the average income for university alumni (mu) is $20,000, and the standard error for this distribution was $700.
- If you wanted to verify the university claims, you could draw a random sample of alumni on your own and calculate the average income. If the results from your sample indicate that the average alumni annual income is $18,500, what does this say about the validity of the university report?
- In other words, what is the likelihood of obtaining an average income of $18,500 if the true population average is $20,000?

Z-scores from a sampling distribution Example

- To answer this question, we first need to determine the distance of our sample mean from the hypothesized population mean:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

All of the necessary information was provided for us in the problem:

sample mean (xbar)=$18,500
population mean (mu)=$20,000
standard error of means=$700

$$z = \frac{18{,}500 - 20{,}000}{700}$$

- Therefore, we know that our sample value is -2.14 standard errors below the mean.

Z-scores from a sampling distribution Example

- To answer the question, we refer to Table A, and that shows that 48.38% of all sample means falls between the mean and 2.14 standard errors below that mean.

- In other words, only 1.62% of all possible mean values fall below our sample mean of $18,500.

- As the book suggests, because the probability of selecting a sample with a mean of $18,500 is .02 (2 chances in 100), if the true mean was $20,000, we can conclude that the evidence does not support the claims made in the university report.
- Are we correct in making this assertion?

Confidence Intervals

- For the reasons discussed above, it is the case that researchers are likely never going to know the true population mean (μ) for a given indicator.
- However, it is possible to *estimate* a range of possible values in which the true population mean is likely to fall. Similarly, we can also estimate the probability that the population mean falls within a given range.
- The statistical procedure that we use to accomplish this objective is called a confidence interval.
- Essentially, the confidence interval allows us to build a “margin of error” around our sample mean.
- Although the concept of a confidence interval may be new, in practice we are frequently presented with such information (i.e., political polls, attitudinal polls).

Confidence Intervals

- Intervals can be constructed for any level of confidence, however, by convention, researchers rely on confidence levels of either 95% or 99%.
- It is important to recognize that we construct confidence intervals because sampling error indicates that our sample mean is not likely to equal the population mean (Recall: $\bar{x} \neq \mu$).
- When we construct a confidence interval, we are essentially recognizing the fact that the sample mean likely does not equal the population mean.
- In return for the loss of precision in knowing where the true population mean is, the intervals provide confidence about its approximate location. We will never know what μ is, but we will conclude with a high degree of certainty its *relative* position.
- Keep in mind that even though we will be 95% (or 99%) sure that μ falls within our confidence interval, there is still a 5% (or 1%) chance that μ falls somewhere outside of our interval.

Confidence Intervals: Large Samples

- When we have information for large samples, for our purposes N>100, Confidence intervals are calculated using the following equations:

95% confidence level

$$CI = \bar{x} \pm 1.96\sigma_{\bar{x}}$$

99% confidence level

$$CI = \bar{x} \pm 2.58\sigma_{\bar{x}}$$

$\bar{x}$ = sample mean

1.96 and 2.58 are the z-scores associated with 95% and 99% confidence, respectively.

$\sigma_{\bar{x}}$ = standard error of the mean

Confidence Intervals:
Large Samples (Example)

- As a violence researcher, you randomly select a sample of 150 cities. The average homicide rate for these cities is 4.1 homicides per 100,000 population. The standard deviation for your sample is 2.6. Using this information, construct a 95% confidence interval for the true national average homicide rate.
- First, you need to gather the three important pieces of information:
  - $\bar{x}$ = 4.1 homicides per 100,000
  - N=150
  - s=2.6
- Because we have a large sample size, again N>100, we know that we will be using the large sample equation. As such we know that we will use a z-value of 1.96.

Confidence Intervals: Large Samples (Example)

- Our equation is still missing one piece of information, the standard error:

$$CI = 4.1 \pm 1.96\sigma_{\bar{x}}$$

- To generate the value to be used for the standard error, we use the following equation:

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{N}}$$

- Please note, we use s in this equation because the true population standard deviation (σ) is not known.
- To calculate the standard error, we plug the known sample information into the equation:

$$\sigma_{\bar{x}} = \frac{2.6}{\sqrt{150}} = .21$$

Confidence Intervals:
Large Samples (Example)

- We now have all of the information we need to construct the confidence interval:

$$CI = 4.1 \pm 1.96 * .21$$

$$CI = 4.1 \pm .42$$

$$CI = 3.68, 4.52$$

- What do these values tell us?
- Funny you should ask. These values are the upper and lower boundaries of our 95% confidence interval.
- Specifically, these values indicate that although we don't know what the true homicide rate (μ) is, we can be 95% sure that it falls somewhere between3.68 and 4.52 homicides per 100,000 population.
- Now construct a 99% confidence interval using the same data.

Confidence Intervals: Small Samples

- Often researchers analyze data for small samples (N<100). In such cases, we can still calculate confidence intervals, but we have to make some adjustments for the small sample size.
  - Most notably, instead of using z-values associated with a given confidence level, we will have to learn about the t-distribution. This distribution is necessary because small samples (and especially with very small samples) do not have the same properties as the standard normal curve (i.e., the z-distribution).
  - We are also going to calculate our estimate of the standard error slightly differently. Because the size of sample standard deviations (especially with very small samples) are biased downward. As such, we need to correct for this known problem, so that the resulting estimate of the standard error is unbiased.

Confidence Intervals: Small Samples

- When we have information for small samples, for our purposes N<100, Confidence intervals are calculated using the following equation:

$$CI = \bar{x} \pm t_{\alpha, df} s_{\bar{x}}$$

$\bar{x}$ = sample mean
$s_{\bar{x}}$ = small-sample estimate of the standard error
$t_{\alpha,df}$ =the value we obtain from the t-distribution

- The specific t-value that we will use in our equation is based on two additional pieces of information:
  - alpha-level (α)=this calculated as follows:
    α=1-confidence level
    for example the alpha-level associated with a 95% confidence interval is 1-.95, or .05. The alpha-level associated with a 99% confidence interval is 1-.99, or .01.
  - degrees of freedom (df) this is an indicator of how different the sample distribution is from the standard normal curve.
  - The number of degrees of freedom for a given problem is calculated as follows: df=n-1.

Confidence Intervals: Small Samples (Example)

- As a violence researcher, you are interested in studying the violent criminal history of inmates. Thus, you randomly select a sample of 25 inmates. The average number of violent crimes committed in the one year prior to the current incarceration for your sample is 13.4. The standard deviation for your sample is 3.2. Using this information, construct a 95% confidence interval for the true average for the number of violent criminal acts committed prior to incarceration.
- To solve this problem, you again need to gather the three important pieces of information:
  - $\bar{x}$ =13.4 violent crimes
  - n=25
  - s=3.2
- The small sample size indicates we are going to need to use the small sample equation.

Confidence Intervals: Small Samples (Example)

- When you have small samples, here are the steps to follow to obtain the t-value of interest:

1. Determine degrees of freedom.
   df=25-1, or df=24

2. Determine alpha-level.
   α=1-.95, or α=.05

3. Next, use these pieces of information and refer to Table C, which contains "Critical Values of t." In this case, we are going to look in the table that contains values associated with a "two-tailed" levels of significance.
   The t-value associated with an alpha-level of .05 (values listed along the top of the table) and 24 degrees of freedom (values listed on the far left side of the table) is 2.064.

   Note that this is the value that we will plug into our equation.

Confidence Intervals:
Small Samples (Example)

- Our equation is still missing one piece of information, the standard error:

$$CI = 13.4 \pm 2.064 s_{\bar{x}}$$

- To generate the value to be used for the standard error, we use the following equation:

$$s_{\bar{x}} = \frac{s}{\sqrt{n-1}}$$

- Note that when we use small samples, we subtract 1 from the sample size to correct for the downward bias associated with small samples. In essence, this procedure "inflates" the standard error slightly.
- To calculate the standard error, we plug the known sample information into the equation:

$$s_{\bar{x}} = \frac{3.2}{\sqrt{25-1}} = .65$$

Confidence Intervals: Small Samples (Example)

$$CI = 13.4 \pm 2.064 * .65$$

$$CI = 13.4 \pm 1.35$$

$$CI = 12.05, 14.75$$

- What do these values tell us?
- Funny you should ask. These values are the upper and lower boundaries of our 95% confidence interval.
- Specifically, these values indicate that although we don't know what the true number of prior violent crimes (μ) is, we can be 95% sure that it falls somewhere between 12.05 and 14.75 in the year prior to their current incarceration.
- Now construct a 99% confidence interval using the same data.

Confidence Intervals: Proportions

- Statisticians are not always asked to calculate confidence intervals using sample means.
- There are times when a researcher will want to draw inference based on proportions, rather than averages.
- Commonly the use of proportions is related to questions about attitudinal measures, such as the share of the population that has a favorable perception of their local police department.
- The commensurate calculation for constructing a confidence interval for proportions is as follows:

  $$CI = P \pm z s_p$$

  Keep in mind that when we perform this calculation, we will always use values derived from the z-distribution. Again, this means that when constructing a 95% confidence interval we will use a z-value of 1.96, and a value of 2.58 for a 99% confidence interval.

Confidence Intervals: Proportions

- As you can see, we will only have to calculate the standard error of the proportion. To generate this value, we will use the following equation:

$$s_p = \sqrt{\frac{P(1-P)}{N}}$$

- Where P is the sample proportion provided in the problem.

- As an example, imagine that you drew a random sample of 150 UML undergraduate students, and determined that 36% were in favor of the death penalty. Based on this information, construct a 99% confidence interval for the true population proportion.

Confidence Intervals: Proportions

- In order to complete this problem, we need 3 pieces of information:

  P=.36 (54/150)
  Z=2.58

- The final value we need is the standard error of the proportion, which we will have to compute

$$s_p = \sqrt{\frac{P(1-P)}{N}} \qquad s_p = \sqrt{\frac{.36(1-.36)}{150}} \qquad s_p = \sqrt{\frac{.36(.64)}{150}} \qquad s_p = .039$$

- Thus the confidence interval is as follows:

$$CI = .36 \pm 2.58(.039) \qquad CI = .36 \pm .101 \qquad CI = .259, .461$$

  Or, based on this sample information, we can be 99% sure that the true population proportion that supports the death penalty falls somewhere between 26% and 46%.

Chapter 7: Testing Differences Between Sample Means

- In this chapter we build on the logic from the previous chapter. However, the method discussed in this chapter is one that allows us to compare values from two different samples.
- Moreover, this chapter introduces the concept of *hypothesis testing*, which allows researchers to draw conclusions about the differences between two (or more) sample means.
- As the authors suggest, there are a number of steps involved in a formal hypothesis test.
- It is important to keep in mind that the results from the hypothesis test will allow us to conclude whether the observed difference between sample means is likely a product of random chance (i.e., sampling error), or if is more likely to reflect a true difference in the population.
- By convention, if the observed difference is large we will conclude that it is not likely due to sampling error, but instead that there is a statistically significant difference between the sample means.

Testing Differences Between Sample Means

- Below is a scenario when testing whether a difference between sample means would be useful statistically.
- Imagine that you wanted to know whether the life expectancy of smokers was significantly different from that of non-smokers. To investigate this question, you collect data for 110 individuals (55 smokers and 55 non-smokers). The sample characteristics are as follows:

	Smokers	Non-Smokers
Mean	76.5	81.5
Std. Dev.	4.1	3.7
N	55	55

- Based on these data, we can see that the average life expectancy between these groups differs. What we do not know is to what do we attribute this difference. The two possibilities are:
  - Sampling error (that is, random chance).
  - A true difference in the population (that is, the difference is statistically significant).

Testing Differences Between Sample Means

- In order to make a statistically informed decision regarding this question, we are required to perform a hypothesis test.
- It is important to keep in mind that the steps involved are a means to answer a very specific question—in this case, whether there is a significant difference between the life expectancy of smokers and non-smokers.
- We will be working through a number of calculations, but at the end we are going to make a single decision about one specific question.

Formal Hypothesis Test

- The first step involved in a hypothesis test is to establish the hypotheses that will be tested.
- The primary hypothesis is referred to as the Null Hypothesis. The Null Hypothesis assumes that the observed difference between the two sample means is due to *sampling error.* In other words, the Null Hypothesis asserts that there is no true difference in the population means for the two groups in the study. The Null Hypothesis is denoted as follows:

$$H_0 : \mu_1 = \mu_2$$

- The secondary hypothesis is called the Research Hypothesis. The Research Hypothesis makes the counterclaim. That is, the Research Hypothesis argues that the observed difference between the two sample means reflects a *true* difference in the population. The Research Hypothesis is denoted as follows:

$$H_1 : \mu_1 \neq \mu_2$$

- At the conclusion of this calculation, we will make a decision regarding which of these competing hypotheses is better supported based on the data. Ultimately, we will either Reject, or Fail to Reject the Null Hypothesis.

Formal Hypothesis Test

- The second step requires us to determine the level of confidence we will use to test the Null Hypothesis. By convention, we will use either an alpha-level of .05 (95% confidence) or .01 (99% confidence). For this specific example we will use an alpha-level of .05.
- The third step requires that we obtain the key information from each sample that will allow us to complete the hypothesis test. We will need to gather three key pieces of information from each sample:

	Smokers	Non-Smokers
Mean	76.5	81.5
Std. Dev.	4.1	3.7
N	55	55

- The fourth step requires that we determine the test statistic that we will use as a comparison. If we have large samples (i.e., N> 100), we can use z-values. When small samples used (i.e., N< 100), the test statistic will be a t-value. In this case because we have a total sample size ($N_1+N_2=110$), we can use a z-value.

  As such, the critical test statistic, or t_{crit}, is 1.96

Formal Hypothesis Test

- In the fifth step, we will generate our computed t-value, or t_{comp}, which we will compare to t_{crit}.
- The formula we use for this calculation is:

$$t_{comp} = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}}$$

- Before we can complete this equation, we will need to calculate the standard error for the difference between sample means. Below is the formula used to generate the necessary value:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\left(\frac{N_1 S_1^2 + N_2 S_2^2}{N_1 + N_2 - 2}\right)\left(\frac{N_1 + N_2}{N_1 * N_2}\right)}$$

Formal Hypothesis Test

- This formula looks daunting, but keep in mind that it is really comprised of many values that are given to you, specifically the sample standard deviations (s) and sample sizes (n).

$$s_{\bar{x}_1-\bar{x}_2}=\sqrt{\left(\frac{55*4.1^2+55*3.7^2}{55+55-2}\right)\left(\frac{55+55}{55*55}\right)}$$

$$s_{\bar{x}_1-\bar{x}_2}=\sqrt{(15.53)(.04)}$$

$$s_{\bar{x}_1-\bar{x}_2}=.75$$

- Now that we have our estimate of the standard error, we need to plug this value into the t_{comp} equation.

Formal Hypothesis Test

- The completed equation is now:

$$t_{comp} = \frac{76.5 - 81.5}{.75} \qquad t_{comp} = \frac{-5}{.75} \qquad t_{comp} = -6.67$$

- In the final step we compare our computed t-value to the critical t-value. Specifically, we want to know if the absolute value of t_{comp} is greater than the absolute value of t_{crit}.
- There are only two possibilities:
  - If $|t_{comp}|>|t_{crit}|$ we will reject the Null Hypothesis. When we reject the Null Hypothesis, we will accept the Research Hypothesis.
  - If $|t_{comp}|<|t_{crit}|$ we will fail to reject (or retain) the Null Hypothesis, and in so doing, we are going to reject the Research Hypothesis.

Formal Hypothesis Test

- Because t_{comp} (6.67) is greater than t_{crit} (1.96), we will reject the Null Hypothesis.
- Our decision to reject (or failing to reject) the Null Hypothesis is based on the comparison of t_{comp} and t_{crit}. However, this decision by itself does not answer the question of interest.
- Once we make our decision, we still need to provide an interpretation of what this decision means with respect to the particular question.
- In this case, we are rejecting the Null Hypothesis which assumes that there is no difference in the life expectancy of smokers and non-smokers. Instead, the data indicate that there is a *statistically significant* difference, at an alpha-level of .05, in the life expectancy of smokers and non-smokers.
- In other words, we are concluding that it is unlikely that we would observe a mean difference this large simply as a product of sampling error. Instead, we are concluding that there is truly a difference in the life expectancies of smokers and non-smokers in the population.

Formal Hypothesis Test

- A few final considerations.
  - The decision we make about the Null Hypothesis is binary.
  - Directional Hypothesis tests.
    - If we have knowledge about the direction of the association (based on prior research or theory), we can test whether the difference falls into *one* tail, rather than placing the critical values on each side of the mean.
  - Small samples.
    - As mentioned previously, when our total sample size is less than 100, the value for t_{crit} will be taken from the t-distribution. The degrees of freedom are calculated as:

      $df=n_1+n_2-2$.
    - If the specific number of degrees of freedom are not listed on the table, round *down* to the next lowest value. By convention, we do not round up to the next highest df value.

Formal Hypothesis Test
Small Sample/Directional Example

- In the following problem we will perform a hypothesis test, very much like the first example, with two differences.
  - The sample sizes are small so our critical test statistics (t_{crit}) will be derived from the t-distribution rather than the z-distribution.
  - We will be testing for a specific difference, that is, whether one mean is significantly *greater* or *less* than another. In cases such as this, we will be performing a directional, or a 1-tailed hypothesis test.
  - Recall that in the first example, we were testing for *any* difference. In other words, we were not concerned with the direction of the difference.

Formal Hypothesis Test
Small Sample/Directional Example

As a violence researcher you are interested in examining patterns of suicide among inmates. Based on criminological theory you believe that it is likely that the suicide rates among medium security facilities are likely to be higher than the rates in maximum security prisons. To answer this question, you collect information on suicides committed over the past 2 years from a sample of 65 institutions (35 medium security and 30 maximum security). The results indicate that the suicide rate in medium security prisons is 14.5 per 1,000 inmates, with a standard deviation of 2.3. The rate for maximum security prisons in your sample is 11.7 with a standard deviation of 3.6. Based on this information, does the evidence suggest that suicide rates are *significantly* higher among medium security prisons, as you suspect?

Formal Hypothesis Test
Small Sample/Directional Example

- In order to answer this question, we will follow all of the same steps involved in the hypothesis test.
- First, we will establish our Null and Research Hypotheses:

The Null Hypothesis will be the same as it was previously. Again, in this hypothesis, we are assuming that the observed difference in suicide rates is due to random variability, or sampling error:

$$H_0 : \mu_1 = \mu_2$$

The Research Hypothesis, however, will be slightly different. In this case we want to know if the observed suicide rates are significantly higher in the medium security prisons. Note that we are making an assumption about the direction of the difference (one is *greater* than the other). Thus, the Research Hypothesis is written a little differently:

$$H_1 : \mu_1 > \mu_2$$

Formal Hypothesis Test
Small Sample/Directional Example

- In the second step, we will again determine the alpha-level. In this case we will use an alpha-level of .05.
- The third step requires that we obtain the key information from each sample that will allow us to complete the hypothesis test. We will need to gather three key pieces of information from each sample:

	Medium	Maximum
Mean	14.5	11.7
Std. Dev.	2.3	3.6
N	35	30

- The fourth step requires that we determine the test statistic that we will use as a comparison. Because of our small sample (i.e., N=65), the test statistic will be a t-value. To determine the t-value we need, we have to first determine our degrees of freedom. In this case $df=n_1+n_2-2$, or 63.

  Remember to ***round down to the next lowest df value*** if the table does not contain the exact value we need. In this example we have an alpha-level of .05, and we will use the value associated with df=60.

Formal Hypothesis Test
Small Sample/Directional Example

- In order to determine the value we need, we will also need to refer to the t-table that contains number associated with **one-tailed tests of significance. This is an important step in a directional hypothesis test.** In fact, this value is what differentiates between a directional and a non-directional hypothesis test.

  As such, the critical test statistic, or t_{crit}, is 1.671.

- In the fifth step, we will generate our computed t-value, or t_{comp}, which we will compare to t_{crit}.
- The formula we use for this calculation is:

$$t_{comp} = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}}$$

Formal Hypothesis Test
Small Sample/Directional Example

- Next we will have to generate our estimate of the standard error for the difference between sample means in order to solve the equation above. To do this we use the same formula as before:

$$s_{\bar{x}_1-\bar{x}_2}=\sqrt{\left(\frac{N_1S_1^2+N_2S_2^2}{N_1+N_2-2}\right)\left(\frac{N_1+N_2}{N_1*N_2}\right)}$$

$$s_{\bar{x}_1-\bar{x}_2}=\sqrt{\left(\frac{35*2.3^2+30*3.6^2}{35+30-2}\right)\left(\frac{35+30}{35*30}\right)}$$

$$s_{\bar{x}_1-\bar{x}_2}=\sqrt{(9.11)(.06)}$$

$$s_{\bar{x}_1-\bar{x}_2}=.75$$

- Now that we have our estimate of the standard error, we need to plug this value into the t_{comp} equation.

Formal Hypothesis Test
Small Sample/Directional Example

- The completed equation is now:

$$t_{comp} = \frac{14.5 - 11.7}{.75} \qquad t_{comp} = \frac{2.8}{.75} \qquad t_{comp} = 3.73$$

- In the final step we compare our computed t-value to the critical t-value. Specifically, we want to know if the absolute value of t_{comp} is greater than the value of t_{crit}.
- Note that in directional hypothesis tests, we do not take the absolute value of out t_{comp} value. In this type of hypothesis test, **direction of the difference matters.**
- In this case, because t_{comp} (3.73) is greater than t_{crit} (1.67), we will reject the Null Hypothesis.
- In situations where t_{comp} does not exceed t_{crit} value we established in step 4, we will fail to reject (or retain) the Null Hypothesis.

Formal Hypothesis Test
Small Sample/Directional Example

- Remember that the decision to reject (or failing to reject) the Null Hypothesis is based simply on the comparison of t_{comp} and t_{crit}. However, this decision by itself does not answer the question of interest. Answering the question at hand requires a more detailed interpretation of the results.
- For example, we are rejecting the Null Hypothesis which assumes that there is no difference in the suicide rates between medium and maximum security prisons. The data indicate that the levels of suicide in medium security facilities are significantly higher than maximum security prisons.
- Now, I would like you to determine if you would make the same decision if we used an alpha-level of .01 rather than .05.

Formal Hypothesis Test
Proportions

- Just as we discussed in the last chapter, researchers are not always going to be interested in testing for differences between sample means. It is also possible to apply the same logic to determine if there is a significant difference between sample proportions.
- When we conduct this hypothesis, keep in mind that it involves a very similar process as we used in the previous analysis. However, the calculations of the standard error for the difference between proportions is somewhat different.
- Again, when we perform calculations using proportions, the critical values will always be taken from the z-distribution.

Formal Hypothesis Test
Proportions

- For this example, I will borrow a problem from the text (#36, p. 265), which asks about whether there are gender differences in preferences for stricter gun controls. The data from the sample are as follows:

	Males	Females
Favor	92	120
Oppose	74	85
N	166	205

- In this case, the question asks if there is a significant difference between men and women with respect to their attitudes regarding gun control. Thus, we will be performing a two-tailed test.

Formal Hypothesis Test
Proportions

- We begin by establishing the Null and Research hypotheses. We use different symbols because we are interested in population proportions, rather than population means:

$$H_0 : \pi_1 = \pi_2$$
$$H_1 : \pi_1 \neq \pi_2$$

- We will test for difference using an alpha level of .05. Thus, we know that the critical value used in this comparison will be 1.96.

- Next, we will have to compute the proportion of males and females who are in favor of stronger gun control laws:
  Males (p1)=.554 (92/166) Females (p2)=.585 (120/205)

Formal Hypothesis Test
Proportions

- As it was for the previous problems, the most involved computation involves generating the sample-based estimate of the standard error. In this type of problem we will compute the standard error in two steps.
- First, we need to establish a value identified as P*, which is essentially a weighted average of the two proportions of interest. P* is calculated as follows:

$$P^{*}=\frac{N_1 P_1+N_2 P_2}{N_1+N_2}$$

$$P^{*}=\frac{(166)(.55)+(205)(.59)}{166+205}$$

$$P^{*}=.57$$

Formal Hypothesis Test
Proportions

- Once we have the P* value, we can proceed to the calculation of the standard error:

$$s_{p_1 - p_2} = \sqrt{P^*(1-P^*)(\frac{N_1 + N_2}{N_1 N_2})}$$

$$s_{p_1 - p_2} = \sqrt{.57(.43)(\frac{166+205}{166*205})} \qquad s_{p_1 - p_2} = \sqrt{.245(.011)}$$

$$s_{p_1 - p_2} = .052$$

- Based on this calculation, we can now complete the test to determine if there is a significant gender difference in support for stricter gun control.

Formal Hypothesis Test
Proportions

- In the next step, we will have to convert the observed difference between our sample proportions into a standard score:

$$t_{comp} = \frac{p_1 - p_2}{s_{p_1 - p_2}} \quad t_{comp} = \frac{.554 - .585}{.052} \quad t_{comp} = \frac{-.031}{.052} \quad t_{comp} = -.599$$

- It is clear that the absolute value of our computed z-score (.599) does not exceed the critical value (1.96). Therefore, we will fail to reject the Null Hypothesis. Based on this decision, how would you interepret the result of this hypothesis test?

- Although the above example is a two-tailed hypothesis, keep in mind that this procedure also applies to one-tailed tests, depending on the research question.

Formal Hypothesis Test
The same sample measured twice

- There are instances where a researcher will examine the same sample of data measured at two points in time. For example, if the research question is interested in examining the impact that exposure to a particular stimulus or policy has on a given outcome.
- Because these paired samples are not chosen independently, we cannot simply perform a hypothesis test using the same procedure as mentioned previously.
- Instead, the calculation must explicitly take into consideration the fact that the lack of independence, or autocorrelation, between observations.
- Much of the logic informing this test mirrors the previous calculations, and thus, we will not go over the calculation in its entirely. Rather, we will concentrate on how this particular calculation differs from previous examples.

Formal Hypothesis Test
The same sample measured twice

- The formula used to calculate the computed test statistic, t_{comp}, is as follows:

$$t_{comp} = \frac{\bar{x}_1 - \bar{x}_2}{s_{\hat{D}}}$$

- As before, you will have to generate an unbiased estimate of the standard error for the difference between sample means. Before you can estimate the standard error, you will first need to calculate the standard deviation (S_D) for the before-after difference scores:

$$s_D = \sqrt{\frac{\sum D^2}{N} - (\bar{x}_1 - \bar{x}_2)^2}$$

- The only value in the above equation that you may not recognize is D^2, which is represents the squared difference between the Time 1 and Time 2 values for every observation. The sigma simply indicates that you will need to sum the squared deviation value across all observations.

Formal Hypothesis Test
The same sample measured twice

- Once you have calculated the standard deviation, we will use this value in the computation of the standard error. The equation used to generate the standard error is as follows:

$$s_{\hat{D}} = \frac{S_D}{\sqrt{n-1}}$$

- As before, we use the resulting value as the denominator in the t_{comp} equation. The final t_{comp} value will then be compared to the t_{crit} value established at the beginning of the problem.
- As always, if our computed t-value exceeds our critical t-value, this suggests that the observed difference falls in the critical region of the distribution, and we have enough evidence to reject the null hypothesis.
- See pages 241-243 in the text for a clear example of this type of calculation.