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.

13

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.

14

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

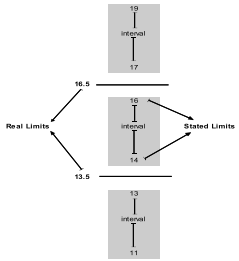
15

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)