INFO 515 Lecture Notes – Val Yonker
Formatted by Glenn Booker – September 23, 2005

These notes should be used with the Action Research handout also by Valerie Yonker. The latter is cited; e.g. Yonker, p. 5 refers to the hand-written page #5 in that handout.

Table of Contents

Week 1 - INTRODUCTION TO ACTION RESEARCH ................................................................. 3
  Purpose of Course: ........................................................................................................... 3
  Importance of Course: .................................................................................................... 3
  INTRODUCTION TO ACTION RESEARCH ...................................................................... 4
  Research Strategies/Methodologies: ................................................................................ 7
  Quantitative Vs. Qualitative Research: .......................................................................... 10
  INTRODUCTION TO STATISTICS .................................................................................. 12

Week 2 - DESCRIPTIVE STATISTICS ................................................................................. 14
  Levels of Measurement: ................................................................................................. 14
  Frequency Distributions: ............................................................................................... 17
  Frequency Distribution Graphs: ....................................................................................... 18
  Continuing with Descriptive Statistics, but Still Focusing on Types of Distributions: ... 20
  Measures of Central Tendency: ...................................................................................... 23
  Theoretical Distributions: ............................................................................................... 26
  Measures of Variation or Dispersion: ............................................................................. 28
  Coefficient of Variation: ................................................................................................. 31

Week 3 - STANDARD SCORES .......................................................................................... 32
  Standard (z) Scores: ....................................................................................................... 32
  Transformed z Scores (T Scores): .................................................................................... 35
  SAMPLING METHODOLOGIES ................................................................................... 36
  Random Sampling Strategies: ........................................................................................ 38
  Non-random Sampling Strategies: ................................................................................ 41
  SURVEYS AND DATA COLLECTION .......................................................................... 42

Week 4 - DESCRIPTIVE STATISTICS--MORE GRAPHS ......................................................... 46
  Stem and Leaf: ................................................................................................................ 46
  Boxplot: .......................................................................................................................... 48
  Frequency Polygon: ........................................................................................................ 49
  INFERENTIAL STATISTICS ............................................................................................. 50
  Standard Error of the Mean: ............................................................................................. 52
  Inferential Statistics, the z test: ....................................................................................... 53
  Inferential Statistics, Types of Errors: ............................................................................ 55
  Inferential Statistics, Confidence Intervals: ................................................................. 57

Week 5 – t TESTS ................................................................................................................. 59
  More Inferential Statistics, Single Sample t Test: ......................................................... 59
  Inferential Statistics, Independent t Test: ....................................................................... 61
  One- and Two-Tailed Tests of Significance: ................................................................. 65
  Inferential Statistics, Dependent t: ............................................................................... 66
Week 1 - INTRODUCTION TO ACTION RESEARCH

Purpose of Course:

1. This course will introduce you to the basic concepts of statistics with a focus on interpretation in the context of Action Research. This is a gentle approach to statistics that is aimed at the beginner, or perhaps someone who needs a refresher course in statistics. We will focus on how these techniques are useful as tools in Action Research and how to interpret results. There will not be a heavy emphasis on calculations (although you will be doing some in your homework) and certainly you will not be deriving formulas. You will also be gaining some experience with a widely used statistics software package, SPSS.

2. This course is also intended to introduce you to the types of research strategies in general, with a focus on Action Research specifically. You will learn about some basic research methodologies that are appropriate for Action Research, such as sampling methods and survey methods; however, this course focuses more extensively on statistics used in Action Research rather than the methodologies for conducting Action Research.

Importance of Course:

1. As a graduate student, you need to have a grasp of statistics to be a good consumer of literature. In a multidisciplinary field, such as Information Science, you need to recognize the statistics used in the research literature, and have a basic understanding of how and why they are used. And, if they are used properly!

2. As an information worker, you will undoubtedly be called upon to organize or present information. I’ve had the pleasure of hearing “statistics stories” from many of my former students who--

   created frequency distributions to use in presentations
   calculated a median or mean from usage data (such as circulation) in a library
   calculated a chi square from tabular data
   etc., etc., etc….

3. As a potential researcher, this course will provide you with the basics of statistics and how to use and interpret these techniques in research. The course will also provide a good base for understanding more advanced statistics should the need arise in your future.
INTRODUCTION TO ACTION RESEARCH

Introduction to Research/Terminology:

What is research? Research describes what or explains why. It is a method for finding answers to questions or a strategy for explanation.

Research is:

1. Empirical, meaning based on evidence or data-centered.
2. Systematic, meaning by use of a method.
3. Objective, meaning it is presumably conducted and interpreted by the researcher without bias.

We can distinguish between two very broad types of research: basic and applied research.

**Basic research** usually refers to laboratory research and is well known from experimental psychology. In basic research, the researcher is testing theory and ideas without necessarily applying the results to practical problems. A classic example is from learning experiments where subjects were asked to recall a series of paired-associate nonsense syllables.

**Applied research** can be called field research, evaluation research, and action research. This type of research is often used to influence policy and decision-making, and is conducted to solve problems (often immediate problems), sometimes only within one organization (and, results are only applicable to that organization).

Examples:

The effects of violence on TV

The effects of TV on reading achievement in children

User studies in Library Science—

Does the type of user influence service received?

What is the success rate for answering reference questions?

What are user information needs (an assessment study)?

We need to go over a few terms that will come up again and again in this course, so they are defined up front. The terms are not necessarily presented in logical order. One of the handouts for this week accompanies these notes and lists each term and its basic definition. A couple of PowerPoint slides also accompany some of the terms.
Cases and Variables

- **Cases** = units of analysis
  - people, things, records, etc.
  - AKA: entities, respondents, subjects, items
  - Become the rows in your data matrix
- **Variables** = things that vary! (not constant)
  - Example: Achievement, Intelligence, Attendance, Income, Agression
  - AKA: measures, attributes, features
  - Become the columns in your data matrix
- **Discrete** = Counting Units
  - Example: Attendance
- **Continuous** = Measurement
  - Example: Intelligence Tests
- **Independent Variables**
  - influences other variables
- **Dependent Variables**
  - influenced by (or consequence of) independent variable.

---

**Data**: Information or observations collected in order to **measure or describe** a situation or problem of interest. Data can be numbers (like percentages), documents (government documents), observations of behavior, questionnaires or interviews. As part of your homework assignment for this week, I am asking you to complete a survey. This is your first example (in this class) of a data collection instrument. Data is collected on a case. See PowerPoint slide (above).

Data is collected to describe one or more **variables**. Variables are objects or concepts that must have a value or a definition assigned to them in order that they can be measured and analyzed. Variable is a good name because variables vary—they take on different values for individuals and groups.

**Examples:**

- Achievement, Intelligence, Attendance, Income, Agression, Types of Books, Types of Computers, Databases, Staff (professional, nonprofessional)

Just by thinking about these different variables you would probably be able to define some more easily that others, such as Income vs. Aggression.

How a variable is defined or measured is the **operational definition**. The operational definition really results in the data you collect. Variables can be defined with two very broad types of data: discrete and continuous.
**Discrete** data can take on only a finite number of values and can be characterized by counting units. Attendance can be defined/counted as the number of days absent from school per month (discrete, or in counting units). Other examples include number of books checked out of the library per day, number of registered students in a class, or amount of money earned in a year (income).

**Continuous** data can take on an infinite number of values and is characterized by some type of measurement, instrument, or scale. You measure height, weight (Does anyone ever know exactly how much they weigh?), speed, and intelligence via some type of instrument (measure).

Once you have your variables, you relate them based upon theory—

**Theory:** A set of concepts (variables) plus the interrelationships that are assumed to exist among those concepts. A theory is a possible explanation of the relationships among variables. It explains what the variables are, how they relate, and why they relate (the underlying logic). Some typical theories, which have been important in information science, include theories of leadership, motivation, learning, and, of course, information theory. A researcher can use existing theory, such as the preceding examples, to lend credence to a study, or could build their own theory by borrowing from various reports in the literature or doing exploratory studies. A hypothetical theory of college achievement is included in Yonker, p. 5. A good theory should be testable.

The consequences of our theoretical assumptions or the statements we submit to testing are called **hypotheses**. The theory is why we think a hypothesis should be true. From our theory of college achievement (Yonker, p. 5), a hypothesis could be that the higher the intelligence the greater the achievement in college.

To clarify our relationships or our hypotheses between variables we can define two types of variables:

**Independent Variable (IV):** A variable which is thought to influence another variable. Some variables precede other variables in hypotheses and theories—these variables seem to come first (the “causes”). Intelligence is an independent variable in our theory of college achievement.

**Dependent Variable (DV):** A variable which is influenced or is the consequence of the independent variable. It is dependent on the variable that precedes it (the independent variable) in a theoretical sense. The dependent variable is what you would most often think of as the results of the study or the “effects.” Achievement is a dependent variable in our theory of college achievement.
Examples illustrating independent and dependent variables:

Effects of (inconsistent punishment) **IV** on (aggression) **DV** in children.

Effects of (wage inequities) **IV** on (work performance) **DV**.

The value of (grade reports) **IV** on (class achievement) **DV**.

Short-term analytically oriented psychotherapy versus behavior therapy (**IV**):
Assessment of target symptoms (**DV**).

**Hypothesis testing** will be covered in more detail when we get into inferential statistics later in the course, but for now, there are two major types of hypotheses:

**Research hypothesis**: Our hunch or assumption—what we plan to test through our research. For example, a researcher may have reason to believe that there is a difference among educational levels and the amount of library use. A researcher could be even be more specific and suggest that higher educated people use the library more.

**Null hypothesis**: A statistical hypothesis—an assumption that no relationship exists. Essentially null says that there is no difference, no influence, no effect. The null hypothesis would state that there is no difference among educational levels and the amount of library use.

This concludes our discussion of basic terminology.

**Research Strategies/Methodologies**:

As mentioned earlier in the orientation to this course, we will cover some research strategies/methodologies at various times during the course. We will start with nine types of research methods. A listing of these methods and the basic purpose of each are included in the handouts for this week. It is important to note that these strategies of research are not clear-cut categories and they can overlap. For example, some of the other strategies (such as descriptive) could be part of an action research study, or lead to an action research study. Probably the clearest distinction among the strategies is the motivation for its use. A couple of these strategies will be revisited at later times during the course in order to illustrate a specific statistic/analysis.

1. **Historical**: to reconstruct the past systematically, objectively, and accurately, often in relation to a hypothesis or theme.

   Example—A researcher could use this strategy in library science to, historically, examine manuscripts which have unknown authors. The researcher could hypothesize about the identity of an author and use
patterns in the writing style and patterns in the content to either support or refute the hypothesis.

2. **Descriptive**: to describe systematically a situation or area of interest factually and accurately. This type of research is usually not conducted for the purpose of hypothesis testing.

   Examples—public opinion studies, census studies, task or job analysis studies. A public library could do a needs assessment study to get some facts and opinions of how the community views the library’s services.

3. **Developmental**: to investigate patterns and sequences of growth and/or change as a function of time.

   Examples—Follow a specific user group over time (longitudinal study), say 10 years, to discover how their needs for and uses of the library change.

   Examine trends in the recent past using existing data that could show regular patterns over time to aid in planning. What materials have been consistently used or not used? What are growth patterns in user groups or growth patterns in the community at large?

4. **Case and Field**: to study intensively the background, current status, and environmental interactions of a given social unit: an individual, group, institution, or community. This type of study is regularly used in anthropological research, but can be used to do an information requirements analysis for an information system.

   Examples—Conduct detailed and in-depth interviews with a specific user group in an organization, such as high-level managers, to determine requirements for a decision support system.

5. **Correlational**: to investigate the extent to which variations in one factor correspond with variations in one or more other factors based on correlation coefficients. Later, we will look at basic statistics of correlation, such as Pearson’s r.

   Examples—Do kids who are high in reading achievement have more skills when using the library? Or, is there a relationship between the size of the student body in a college or university and the size of the periodical collection?

6. **Causal Comparative, or Ex Post Facto**: to investigate possible “cause-and-effect” relationships by observing some existing consequence and searching back through the data for plausible causal factors. Cause-and-effect relationships are very difficult to determine in this type of study. There is already a consequence so there is no direct control by the researcher.

   Example—identify a group of high school students as readers versus nonreaders. What type of variables might we hypothesize as plausible causal factors? Were they read to as a young child? Educational background of parents, number of books in the home, reading level
7. **True Experimental Research**: to investigate possible cause-and-effect relationships by exposing one or more experimental groups to one or more treatment conditions and comparing the results to one or more control groups not receiving the treatment (random assignment being essential). The researcher has control of the group assignment and does so without bias (randomization). This is in contrast to ex post facto studies where the researcher has no control of the group assignment (such as readers versus non-readers). Experimental research is often difficult to do in a natural setting.

   I will use an example of an experiment in a library setting, later on, to illustrate a method of analysis called Analysis of Variance.

8. **Quasi-Experimental Research**: to approximate the conditions of the true experiment in a setting which does not allow the control and or manipulation of all the relevant variables. This type of study is more flexible and adaptable to a natural setting than a true experiment because it does not assume assignment to groups (such as treatment and control) without bias (meaning at random).

   Quasi-experimental studies can be used in more sophisticated forms of action research to test the effectiveness of some type of program (such as a library use training program for junior high school students—this example will be used later in the course).

9. **Action Research**—to develop new skills or new approaches and to solve problems with direct application to the classroom, library, or other applied setting. Action Research is a well-known and widely used strategy in education. It enables problem solving and decision making in the natural setting, but usually lacks the control of a true experiment. The results usually apply only to the setting where the study is conducted, in other words, the results are not generalizable to other groups, settings, or situations. Because action research often leads to change in organizations, it is important to get the support of individuals/groups who would be impacted by the change (participatory action research). See

   [ActionRes1.ppt]

---

**Action Research**

- A strategy in Educational Research
- Enables problem solving in the natural setting
- Participatory action research
- Connect theory with practice

For types of questions in library and information science, which could lead to an action research study, see
Action Research Questions in Library and Information Science

- How much does the library spend?
- How much do potential users actually use the library?
- How productive is the library staff, is the staff the right size?
- How are users served by the library?

Quantitative Vs. Qualitative Research:

To round off the discussion of methodologies it is important to contrast Quantitative and Qualitative Research. We will be focusing on the quantitative in this course. In the 3rd class session, we will return to a discussion of methodologies by covering sampling and survey methods.

The difference between quantitative and qualitative research is the type of data you collect (the result) and the tools (instruments and strategies, such as interview, questionnaire, observation, document analysis, tests) employed to collect it. In regard to tools or strategies:

1. The same data collection strategies can be qualitative or quantitative.
   a. The structure of an interview or questionnaire can determine whether you have qualitative or quantitative data. A structured interview or questionnaire has predetermined answers to questions or includes some type of scale. Examples:

   Have you ever had any experience with quantitative methods (including previous coursework) or participated in a research project?

   _____yes _____no

   What are your feelings when you hear the term statistics? Please circle the point which indicates your feelings.

   1 2 3 4 5
   negative neutral positive

   In the above examples, the frequency (and percent) of responses for each answer could be counted; therefore this is quantitative data.
In contrast, if a question in an interview or questionnaire allows the respondent to freely answer without predetermined categories, this results in qualitative data (the respondent’s own words). A comments section after a structured question will result in qualitative data.

b. If you are observing and writing descriptions of the behavior of a group of people, that is qualitative. If you are looking at the frequencies of certain types of behavior, that is quantitative.

2. Qualitative data can often become quantitative. When you find patterns or categories in an interview or a document and determine the frequency of those categories, then qualitative data has become quantitative.

Example: The table below (Table 1) was taken from my dissertation and illustrates information types used by clinical psychologists in making diagnoses about their clients. These information types (and, the decision as well) were isolated via quotes found in the interviews with the psychologists. There were 15 examples (incidents) of the diagnostic decision found in the interviews. So, qualitative data became quantitative!

3. Pure quantitative data cannot become qualitative. You can interpret quantitative data, but cannot change its form—think of a standardized achievement test, results of the GRE exams, or scales used in a questionnaire.

4. Often in research, it’s good to use both quantitative and qualitative data in the same study. Qualitative data provides rich support or explanation to quantitative data. Note in the table below from my dissertation that the qualitative data explains the importance of the psychological evaluation!

Table 1: Types of Information Required for Diagnostic Decision (N=15 incidents)

<table>
<thead>
<tr>
<th>Information Types</th>
<th>Percent of Incidents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Psychological Evaluation</td>
<td>73%</td>
</tr>
<tr>
<td>Behavior History</td>
<td>53%</td>
</tr>
<tr>
<td>Family/Social History</td>
<td>40%</td>
</tr>
<tr>
<td>Early and Developmental History</td>
<td>13%</td>
</tr>
<tr>
<td>Previous Diagnosis</td>
<td>40%</td>
</tr>
<tr>
<td>Current Behavior</td>
<td>33%</td>
</tr>
<tr>
<td>Psychiatric Evaluation</td>
<td>20%</td>
</tr>
<tr>
<td>Organic/Neurological Evaluations</td>
<td>33%</td>
</tr>
</tbody>
</table>
**Qualitative Data:**

"The psychological is very important and that it be done in depth….to make some inferences about a person's character. We try to understand things like how does this person view the world, how does this person view themselves….."

Basically, you can think of qualitative data as words, and quantitative data as numbers. So, that launches us right into our discussion of statistics!

**INTRODUCTION TO STATISTICS**

**What is statistics?**

Methods for arranging data or pieces of information and drawing conclusions about them—a language with special jargon and symbols. In Yonker, p. 4, you will notice a listing of some common symbols in statistics. These symbols are not for you to memorize, although I suspect you will become used to most of them as you progress through the course.

**There are two major functions of statistics:**

1. **Descriptive statistics:** Numbers that describe a situation or problem of interest and provide an efficient summary. Examples--
   
   average family size
   literacy rates
   unemployment figures
   batting average
   average daily circulation
   percentages, averages, rates of change

   We will spend about a session and a half on this function.

2. **Inferential statistics:** Tests which we use in research to draw conclusions about groups of data. In research you make hypotheses, collect data, apply tests (inferential statistics) to this data, and make decisions/conclusions concerning the results of these tests. Inferential statistics are more powerful than descriptive statistics because you can draw conclusions; however, this function involves probability and some risk. There is always a certain probability you could be wrong in your conclusion. We will get into inferential statistics in more detail in session 4.
Sources Used For This Course:


Week 2 - DESCRIPTIVE STATISTICS

Levels of Measurement:

Last week we left off by introducing descriptive and inferential statistics. The type of statistic you use, whether descriptive or inferential, depends on the scale of measuring a variable, also known as levels of measurement.

There are four levels of measurement. The levels of measurement will be described from the lowest level of measurement, which is a nominal scale, to the highest level of measurement, which is a ratio scale. Throughout the course, you will see that certain statistics can only be used with higher levels of measurement, whereas some statistics are used exclusively with lower levels of measurement. See PowerPoint slide that supplements the discussion of levels of measurements.

[Levels_of_Measurement.ppt]

Levels of Measurement

<table>
<thead>
<tr>
<th></th>
<th>Nominal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>names different categories, not ordered, not ranked: Male, Female, Republican, Catholic..</td>
</tr>
<tr>
<td></td>
<td>Ordinal</td>
</tr>
<tr>
<td></td>
<td>Categories are ordered: Low, High, Sometimes, Never,</td>
</tr>
<tr>
<td></td>
<td>Interval</td>
</tr>
<tr>
<td></td>
<td>Fixed intervals, no absolute zero: IQ, Temperature</td>
</tr>
<tr>
<td></td>
<td>Ratio</td>
</tr>
<tr>
<td></td>
<td>Fixed intervals with an absolute zero point: Age, Income, Years of Schooling, Hours/Week, Weight</td>
</tr>
<tr>
<td></td>
<td>Age could be measured as ratio (years), ordinal (young, middle, old), or nominal (baby boomer, genex)</td>
</tr>
<tr>
<td></td>
<td>Level of measurement affects (sometimes determines) type of stats. That you can use to analyze.</td>
</tr>
</tbody>
</table>

Nominal: A scale of measurement that involves distinct categories or classification. The categories may differ (not be equal), but you can't say one category is greater than (in terms of quantity) then another category. A nominal scale may use numbers as categories (such as room numbers), but the numbers do not have quantitative properties. So you would not add room 205 to room 209!
Examples:

- Types of computers—IBM, Macintosh, Compaq, Dell
- Library classification—the Dewey Decimal system or the Library of Congress system both use numbers for classifying different sections of the library, but the numbers do not tell you that one section of the library is greater or better than another!
- Gender—male, female (not necessarily in this order!)
- I think you would all agree that one category of gender is not greater than or lesser than another category!
- Academic majors—psychology, sociology, business
- Numbers on hockey players' jerseys (a classification of the player)
- Volume numbers on a Library and Information Science journal

**Ordinal:** A scale of measurement that possesses an inherent order or different ranking. The rank or the scale points indicate greater than or lesser than, but we don't know the magnitude (amount) of the difference between the ranks or scale points or "how much" difference.

Examples:

- Choices on a questionnaire: Under 18, 18-45, 46-65, over 65
  
  I am in the 18-45 range, so you know that I am older (greater in age) than someone who is under 18. You also know that I am not as old (lesser in age) as someone who is 46-65. But, you have no idea how much older I am than an individual who is under 18 because from the choices (the ordinal scale), you do not know my exact age!

  - Class Ranks of three students: 99 average (ranked 1), 98(2), 97.5(3)
    
    The ranks (the ordinal scale) are in the parentheses. Once again, you can see that the rank does not show the actual differences (how much) in the averages among the three individuals.

- A scale which indicates the wear and tear on library books: 1=least, 5=most

**Interval:** A scale of measurement that provides fixed units of measurement (equal intervals) and an arbitrary zero point. An arbitrary zero means that zero does not indicate a complete absence of the characteristic being measured. Because an interval
scale is characterized by equal intervals, you can determine the amount of difference between two numbers (unlike ordinal), but because you don't have a base of zero, you cannot say one number is twice as great as another. You can use addition, subtraction, etc. with an interval scale. Interval scales are commonly used to measure concepts, such as intelligence and achievement.

**Examples:**

Intelligence test—theoretically, zero intelligence has not been defined, so zero is arbitrary. An intelligence test does have equal intervals, so Einstein with an IQ of 200 was 100 points above the average person (with an IQ of 100). But, because zero is arbitrary and not absolute, we cannot say that Einstein was twice as intelligent as someone with IQ=100.

Achievement tests, GRE'S, SAT'S

The calendar as a measure of time.

**Ratio:** A level of measurements that has the same characteristics of an interval scale (fixed units of measurement/equal intervals) with the additional characteristic of an absolute zero point. With a ratio scale, zero means total absence of the characteristic being measured. Anything that can be measured from a base of zero is a ratio scale. Because you are dealing with a base of zero, you can make the statement “twice as much.” A student who is absent from class 4 days has been absent twice as much as a student who only missed 2 days.

**Examples:**

- number of students in a class (class size)
- number of days absent (attendance)
- number of relevant citations received from a search
- number of books checked out of the library per day
- height, measured by a tape measure
- number of reference materials which have to be reshelved by the library staff per day.
**Frequency Distributions:**

We are going to change topics and focus on ways of organizing or describing data. When you have a number of observations (data) from cases (such as people), you need to organize or present the data so you can say something about it. So, we need to start talking about tools that help us with making descriptive statements about data. One way of presenting data is to prepare a **frequency distribution**. A **frequency** is defined as the number of times a given value of a variable occurs. Frequency is symbolized by “f” or “F.”

![BookMobile_1.ppt](image)

<table>
<thead>
<tr>
<th>Case/Bookmobile</th>
<th>Value of Var. No. of Stops</th>
<th>X No. of Stops</th>
<th>F No. of Bookmobiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
<td>17</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>9</td>
<td>16</td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td>10</td>
<td>14</td>
<td>5</td>
</tr>
<tr>
<td>D</td>
<td>14</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>E</td>
<td>16</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>F</td>
<td>17</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>16</td>
<td></td>
<td>N = 17</td>
</tr>
<tr>
<td>I</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>16</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The first two columns of the PowerPoint slide represent an unordered list of data on 17 bookmobiles (bookmobiles are labeled by a letter). The variable of interest (in the second column) is the number of stops (made by each bookmobile). An action researcher would want a good presentation of this data, so some descriptive statements can be made concerning the “behavior” of these bookmobiles. Some questions to ask are: Do we need so many bookmobiles? Can we economize? How many stops can each bookmobile reasonably handle?

A **frequency distribution** will give order to our data. The last two columns represent the frequency distribution. A frequency distribution is a table that lists the individual observations for a variable and the number of times a given value occurs. As you can
see we took our variable (labeled X) and listed the number of stops from highest to lowest (lowest to highest is also an option). Beside this list, we placed the frequency (symbolized as F), or the number of times each observation/value occurs.

Important points--
1. The frequencies should total 17, which is the N (or n). “N” represents the number of bookmobiles (cases).
2. From this frequency distribution you can make descriptive statements about the bookmobiles. You can see that quite a number of bookmobiles made 14 stops or over!

As an extra piece of information (FYI)—When a researcher has a lot of data, often the numbers are grouped by intervals (range of values). The frequency is the number of values in a distribution that fall within an interval.

Say you have N=500 test scores, you may want to group scores with an interval of 5.
96-100
91-95
86-90

We will not be working with any distributions that have an incredibly large N, so in this class, list each number individually (in other words, do not group the data) when you construct a frequency distribution.

**Frequency Distribution Graphs:**

From a frequency distribution, you can take this same data and display it graphically by a technique known as a **histogram. A histogram is known as a frequency distribution graph** and gives you a “picture” of the distribution. We will cover this and one other graphic technique during this session, but will cover more techniques in week four.

[Histogram.ppt]
The histogram (Bookmobile distribution) in the PowerPoint slide was created by SPSS. A histogram consists of a set of bars—the position of a bar is over the value of the variable, the height of the bar represents the frequency of the value. Some general rules apply for constructing histograms, and SPSS follows these rules.

1. The frequency is always located on the vertical axis (called the Y axis), and the variable is always located on the horizontal axis (X axis).
2. In designing the histogram, the “3/4 high rule” applies. This rule states that the vertical or Y axis should be constructed so that the height of the maximum point (the highest bar) is approximately equal to ¾ the length of the horizontal or X axis. This rule is important for providing a standard height to the bars. It is very difficult, but necessary, to create standards for constructing graphs, considering variables have different units of measure, scales, etc. Standards are to keep “well-meaning” individuals from stretching the truth with statistics! A good book on this topic is How to Lie with Statistics, by Darrell Huff.
3. Always label the X (Number of Stops in this case) and Y (f or F) axis, the O point or origin, and title the graph.
4. In SPSS, each bar usually represents a range of values (an interval) that is calculated by an algorithm (you do not need to know the algorithm!). The midpoint of the interval becomes the value of the variable. In this particular histogram, 2.5 is the range of the interval with 1.25 on either side of the midpoint.

So what does a histogram do for us? Very simply, we can see very quickly the frequency of a value. It also enables us to see the shape or pattern of the distribution. This shape or pattern will take on more significance when we go over the normal distribution (the bell-shaped curve) later on in this session.

From your homework assignment, you may have seen the following graph:
Looks like a histogram, but it’s really a **bar graph**! A bar graph consists of bars (not surprisingly) like a histogram, but is used only for nominal and ordinal level of measurement.

So, to summarize frequency distributions and graphs see PowerPoint slide.

[Frequency_Distributions.ppt]

---

**Frequency Distributions**
- A frequency distribution is a tabulation that indicates the number of times a score or group of scores occurs
- Bar charts best used to graph frequency of nominal & ordinal data
- Histograms best used to display shape of interval & ratio data

---

**Continuing with Descriptive Statistics, but Still Focusing on Types of Distributions:**

There are three other distributions that can be created along with the frequency distribution. These distributions will be illustrated, once again, through the bookmobile data in the PowerPoint slide.

[BookMDist.ppt]

---

**Bookmobile Distributions**

<table>
<thead>
<tr>
<th>Stops</th>
<th>f</th>
<th>%</th>
<th>CF</th>
<th>CF</th>
<th>C%</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>2</td>
<td>11.8</td>
<td>17</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td>23.5</td>
<td>15</td>
<td>6</td>
<td>88</td>
</tr>
<tr>
<td>14</td>
<td>5</td>
<td>29.4</td>
<td>11</td>
<td>11</td>
<td>64</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>11.8</td>
<td>6</td>
<td>13</td>
<td>35</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>17.6</td>
<td>4</td>
<td>16</td>
<td>23</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>5.8</td>
<td>1</td>
<td>17</td>
<td>6</td>
</tr>
</tbody>
</table>

1. **Percent** (3rd column of PowerPoint slide)  
2. **Cumulative Frequency** (4th and 5th column)  
3. **Cumulative Percent** (6th column)

In order to calculate a **percent**, you first have to calculate a **proportion**, which is a fraction, usually expressed as a decimal, comparing a small number of cases to a total number of cases.

**Example:**  
5 bookmobiles made 14 stops; therefore, 5/17=.294
.294 is the proportion
A **percent** is a proportion multiplied by 100. \(0.294 \times 100 = 29.4\%\)
So, 29.4\% of the bookmobiles made 14 stops.

**When you look at an SPSS printout, you will see percent and valid percent.** See table in PowerPoint slide below.

[Percent_Table.ppt]

<table>
<thead>
<tr>
<th>Employment Category</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td>363</td>
<td>76.6</td>
<td>76.6</td>
<td>76.6</td>
</tr>
<tr>
<td>Custodial</td>
<td>27</td>
<td>5.7</td>
<td>5.7</td>
<td>82.3</td>
</tr>
<tr>
<td>Manager</td>
<td>84</td>
<td>17.7</td>
<td>17.7</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>474</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

**Valid percent** excludes missing values in the computation. A missing value, for example, occurs when an individual does not answer a question in a survey and is therefore not included in the distribution for that question (the “N” is reduced). If there are no missing values (as in the table above), percent and valid percent are the same.

The 4th and 5th columns in the PowerPoint slide of the bookmobile data both represent a **Cumulative Frequency Distribution**. Cumulative frequencies (CF) are successive additions or "accumulations" of entries from the frequency (f) column. **How to do this from the 4th column starting at the bottom of the frequency column:**

The first entry in f (bottom) is 1, so enter 1 in cumulative frequency.
The second entry in f is 3, so add 3 to 1 to get 4 in CF
The third entry in f is 2, so add 2 to 4 to get 6 in CF
And so on until the top number in your CF column should equal your N (17 bookmobiles).

So, why do we care about a cumulative frequency?! This distribution provides the total frequencies associated above or below each score. A descriptive statement that illustrates is: From the 4th column, only 6 bookmobiles made 10 or fewer stops!

A cumulative frequency can also be created from the top of the frequency column down (either way is fine!). This is illustrated in the 5th column. Work through it! A descriptive statement is: Eleven bookmobiles made 14 or greater stops!

The last column from this PowerPoint slide is called the **Cumulative Percent**. It shows the percentage of cases above or below each score. A cumulative percent can be obtained by the following formula:

\[ C\% = \frac{\text{CF}}{N} \times 100 \]
From the bottom of the 4th column
1/17 x 100 = 5.88 rounded to 6%
4/17 x 100 = 23%

**THE TOP NUMBER SHOULD EQUAL 100%**

A descriptive statement that illustrates the function of cumulative percent is: 35% of the bookmobiles made only 10 or fewer stops.

You can create both cumulative frequency and cumulative percent from either direction (top or bottom) in your distribution.

Still continuing with Descriptive Statistics:

Two Useful Descriptive or Summary Measures are--

1. **Percentage Rate of Change**
2. **Ratio**

A **percentage rate of change** measures changing conditions over time and can show either a percentage increase or decrease.

**Example:**

Historical study of the change in a library’s collection size:

- 1965—collection size was 50,000 volumes
- 1975—collection size was 150,000 volumes

\[
\text{Rate of change} = \frac{\text{Later Value} - \text{Earlier Value}}{\text{Earlier Value}}
\]

\[
\frac{150,000 - 50,000}{50,000} = 2
\]

2 X 100 changes to 200%

The library’s collection size increased by 200%

A **Ratio** is the number of cases with property X compared to the number of cases with property Y.
Example:

Property X—on a given day, a library checked out 150 fiction titles.

Property Y—on the same day, a library checked out 300 nonfiction titles

The ratio of fiction/nonfiction is 150/300

This fraction can be reduced by dividing both sides of the fraction by 150. So, the ratio becomes $\frac{1}{2}$ or 1:2—fiction is $\frac{1}{2}$ of nonfiction.

SPSS changes this to a decimal value, .5.

**Measures of Central Tendency:**

The focus is still on descriptive statistics, so we'll move on to three types of summary measures called **Measures of Central Tendency**:

1. **Mode**
2. **Median**
3. **Mean**

Central Tendency refers to a number that is located at the center of a group of scores. Keywords include, center value, average value, and most typical value.

The **Mode** is the value of the variable that occurs most frequently. From the bookmobile data (see PowerPoint slide below), the mode is 14 because it is the most frequent value (it occurs 5 times). The mode can be reported with any level of measurement (nominal, ordinal, interval, ratio).

[BookMDist.ppt]

<table>
<thead>
<tr>
<th>Bookmobile Distributions</th>
<th>Stops</th>
<th>f</th>
<th>%</th>
<th>CF</th>
<th>CF</th>
<th>C%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>17</td>
<td>2</td>
<td>11.8</td>
<td>17</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>4</td>
<td>23.5</td>
<td>15</td>
<td>6</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>5</td>
<td>29.4</td>
<td>11</td>
<td>11</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>2</td>
<td>11.8</td>
<td>6</td>
<td>13</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>3</td>
<td>17.6</td>
<td>4</td>
<td>16</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>1</td>
<td>5.8</td>
<td>1</td>
<td>17</td>
<td>6</td>
</tr>
</tbody>
</table>
Example with nominal data:

<table>
<thead>
<tr>
<th>Pets(X)</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dogs</td>
<td>12</td>
</tr>
<tr>
<td>Cats</td>
<td>10</td>
</tr>
<tr>
<td>Fish</td>
<td>4</td>
</tr>
<tr>
<td>Ferrets</td>
<td>2</td>
</tr>
</tbody>
</table>

The mode = Dogs (occurs most frequently, 12 times)

A distribution can have more than one mode. A bimodal distribution contains two modes, but there can be trimodal or multimodal distributions.

The **Median** is a point in a distribution where half of the cases are above and half of the cases are below when your values are ordered highest to lowest (or lowest to highest). The median divides the distribution in half, or two equal parts. Sometimes you will see the median with ordinal data, but usually it is used with interval or ratio level of measurement.

There are two methods to finding the median, depending on whether you have an odd or even number of cases. For an odd number of cases, see PowerPoint slide:

[Median1.ppt]

<table>
<thead>
<tr>
<th>HOLDINGS IN 7 DIFFERENT LIBRARIES</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
</tr>
<tr>
<td>7,400</td>
</tr>
<tr>
<td>6500</td>
</tr>
<tr>
<td>6200</td>
</tr>
<tr>
<td>5900</td>
</tr>
<tr>
<td>5100</td>
</tr>
<tr>
<td>4300</td>
</tr>
<tr>
<td>3800</td>
</tr>
</tbody>
</table>

In this example, the median is the 4th value from the top or bottom of the distribution. Median = 5900

For an even number of cases, See PowerPoint slide:
An **important note** about finding the median with a frequency distribution: The median is at the midpoint of all of the scores. In our bookmobile distribution (**see PowerPoint slide**),

\[(17 + 1)/2 = 9\]

The 9th value in this distribution is the median (you can count up or down the frequency column, \(f\), to find this value); therefore, the median is 14.

**The Mean** is what you most often think of as the average score. It can only be computed with interval or ratio level of measurement. Because the mean takes into account all of the values in a distribution--the average--it is most often used in inferential statistics. Consult your symbol list (in Yonker, p. 4) for common symbols for the mean. You will also see on your symbol sheet, the “sum of” symbol, which is used in the formula for computing the mean.

If you have values listed without a frequency distribution, see PowerPoint slide below for the computation of the mean.
HOLDINGS IN 7 DIFFERENT LIBRARIES

<table>
<thead>
<tr>
<th>X</th>
<th>Mean = ( \frac{\sum X}{N} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>7,400</td>
<td>5600</td>
</tr>
<tr>
<td>6500</td>
<td>( \frac{39200}{7} )</td>
</tr>
<tr>
<td>6200</td>
<td></td>
</tr>
<tr>
<td>5900</td>
<td></td>
</tr>
<tr>
<td>5100</td>
<td></td>
</tr>
<tr>
<td>4300</td>
<td></td>
</tr>
<tr>
<td>3800</td>
<td></td>
</tr>
<tr>
<td>( \sum X = 39200 )</td>
<td></td>
</tr>
</tbody>
</table>

For values in a frequency distribution (which is very common), see Yonker, p. 6 for the computation of the mean (this was an actual distribution of IQ scores from a class at IST!). Note that the fX column takes into account that some of the values occur more than once (Remember Frequency!), and the mean is the average of all of the values. So, for example, the first number at the top of the fX column was calculated by taking 140 (value of variable) times 2 (the frequency of that value). The sum of fX is equal to 2597 and that becomes the numerator in the formula for the mean.

Considering we have three measures of central tendency, why would we use one as opposed to the others? You have already seen that you use these measure of central tendency with different levels of measurement. But, there is one other characteristic that determines their use. The mode and median are not influenced by extreme (high or low) values in a distribution. The mean is most definitely influenced by extremes. See the example in Yonker, p. 7, which shows a distribution of part-time salaries. You can see the values for the mode, median, and mean. You can also see that it would not matter if the last salary was 100,000 (instead if 20,000), the median and mode would be the same values as calculated. The mean is higher than the median and mode due to the extreme value of 20,000.

In cases where you have an extreme value (high or low) in a distribution, it is important to report both the median and the mean. Reporting both values gives some indication (through comparison) of a skewed distribution. A very high or low score in the distribution could skew the mean in the direction of the extreme. Skewing often occurs in distributions like income and salary.

**Theoretical Distributions:**

Up until this point in the session, we have been dealing with actual data, called empirical distributions. There are also theoretical distributions based on the way data generally behave. One theoretical distribution you may be familiar with is the Normal Distribution or the Normal Curve. The normal curve is a bell-shaped or symmetrical curve where scores gather or cluster around a central value like the mean. And, as you
will see when we cover measures of dispersion/variation, these values vary about the mean in different ways. The normal distribution is actually a family of distributions made up of:

1. **Mesokurtic curves**—the classic bell-shaped curve
2. **Leptokurtic curves**—indicated by scores that are very closely clustered around the mean
3. **Platikurtic curves**—indicated by scores that are spread out around the mean

![Leptokurtosis and Platykurtosis](http://www.contingencyanalysis.com/glossarykurtosis.htm)

A measure of how a curve is “peaked” (leptokurtic) or “flat” (platykurtic) is called the **Kurtosis** of the distribution and is illustrated in the figure above. The index for kurtosis is zero for Mesokurtic curves (shown as normal, above), positive for leptokurtic curves, and negative for platikurtic curves. Kurtosis is also influenced by the length of the curve’s tails. A positive kurtosis shows that the tails are longer than normal. A negative kurtosis shows that the tails are shorter than normal. Generally, a kurtosis value above two (positive or negative) is evidence that the curve is not classically normal (a mesokurtic curve). So, a high positive value of kurtosis indicates a peaked curve with a long tail(s). The measure of kurtosis should be viewed with a graph such as a histogram.

Getting back to the normal distribution, because this curve is perfectly symmetrical, 50% of the distribution falls above the mean, and 50% of the distribution falls below the mean. In all normal or symmetrical distributions, the mean, median, and mode should be about equal, and there is no **skewing**.

Sometimes this complete perfection does not occur. As mentioned before, when you have extreme values in a distribution, the data are said to be **Skewed or Assymetrical**.

1. When an extreme value(s) is high, this is called a **positive skew** and is represented in the graph below.
2. When an extreme value(s) is low, this is called a **negative skew** and is represented in the graph below.

The index for skewing is zero when you have a normal curve. If the skewing index is above two, either positive or negative, the skewing is becoming more extreme. Again, graphs, such as the histogram, will show you the shape or pattern of the distribution and whether you have a fairly normal or skewed distribution.

**Measures of Variation or Dispersion:**

I mentioned earlier that other than central tendency we also have measures that tell us how data is spread out or dispersed around a central value like the mean. These descriptive statistics are called **Measures of Variation or Dispersion.** They are:

1. **Range**
2. **Variance**
3. **Standard Deviation**

The **Range** is defined as the largest score in a distribution minus the smallest score in the distribution. From the salary data in Yonker, p. 7, Range = 20,000 – 4100 = 15,900

The range is not a very reliable measure of dispersion because it only reflects the extremes. You do not know what the middle values are and you do not know if the distribution is skewed.

The most common measure of dispersion is called the **standard deviation**. Basically, the standard deviation shows how other scores in a distribution deviate around the mean. It shows the spread of scores or how the scores are gathered around the mean. It also indicates whether the scores are close to the mean as in a **leptokurtic** curve, or flattened out around the mean as in a **platikurtic** curve.

In order to compute the standard deviation we must first compute another measure called the **variance**. The variance is usually a very large number that is not easily
interpreted. So, we will compute the variance, but not attempt to go over it conceptually. We will concentrate on the standard deviation.

The first formula I will show you for the variance and ultimately the standard deviation is a definitional formula. You can actually see the meaning of the standard deviation from the calculations because you subtract the mean from each score; therefore, you can see the “dispersion” of scores from the mean. The formula is called the Mean-Deviation Formula. You can check your symbol sheet in Yonker, p. 4 for the symbols for variance and standard deviation. Also, in Yonker, p. 8, you will find the Mean-Deviation Formula for the variance and the step-by-step procedure to use on the data below:

We have circulation figures for 6 weeks (N = 6). Mean=1600

<table>
<thead>
<tr>
<th>X</th>
<th>Step 1 (X – mean)</th>
<th>Result Step 1 (X - mean)²</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>2000-1600</td>
<td>=400</td>
</tr>
<tr>
<td>1900</td>
<td>1900-1600</td>
<td>=300</td>
</tr>
<tr>
<td>1800</td>
<td>1800-1600</td>
<td>=200</td>
</tr>
<tr>
<td>1600</td>
<td>1600-1600</td>
<td>=0</td>
</tr>
<tr>
<td>1200</td>
<td>1200-1600</td>
<td>=-400</td>
</tr>
<tr>
<td>1100</td>
<td>1100-1600</td>
<td>=-500</td>
</tr>
<tr>
<td>9600</td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

Also note:

1. The sum of the deviations from the mean is equal to zero.

2. You need to get rid of the negatives so you square the deviations from the mean (X – mean)² and then add the squares to give you the Sum of Squares, which is 700,000.

Step 4, The variance = (700,000)/6 = 116,666.6

Once you have the variance, it is just one easy step to the standard deviation—the standard deviation is simply the square root of the variance. So, if you put a square root sign over the Mean-Deviation formula, then it will be the formula for the standard deviation. You need to return back to the original units of measurement before you squared your deviations from the mean. Standard Deviation = 341.6

The formula I just covered for the variance and standard deviation is a good definitional formula as it shows you how scores deviate from the mean, but it is very tedious to calculate (and, anytime you use it, you should round the mean to a whole number). A more convenient computational formula is called the Raw Score Formula and is illustrated completely in Yonker, p. 9, using the bookmobile data.
Important points in using formula:

1. The first three columns (the score—X, the frequency—f, and fX—all used to calculate the mean) you are familiar with.

2. \(X^2\) requires each value of X to be squared, so, from top, 17 (X) squared is equal to 289 (\(X^2\)).

3. The fifth column, \(fX^2\), is a misleading symbol. Always think of it as \(f(X^2)\). So, from the top, 2(frequency) times 289 (our \(X^2\) value) is equal to 578.

4. Check your symbol sheet for the “sum of” sign in Yonker, p. 4. The sum of \(fX = 221\), the sum of \(fX^2 = 3061\), and \(N = 17\). You can see those numbers “plugged in” to the formula.

5. Then follow the order of operations. Remember “My Dear Aunt Sally?” This, of course, refers to the order of operations—multiplication, division, addition, and subtraction. So, you can see that, in the formula, 221 squared divided by 17 is equal to 2873.

6. 3.3 is the standard deviation.

In your homework, you can use either the Raw Score Formula, or the Mean-Deviation Formula, whichever you like the most! But keep in mind for future reference, if you use the Mean-Deviation formula with data, like the bookmobile distribution just illustrated, you must subtract the mean from each of the 17 scores. REMEMBER FREQUENCY—sometimes a score occurs more than once! So you would subtract the mean from the value of “16” four times!

Now, for the meaning of the standard deviation! You can think of the standard deviation as a measure of variability or precision. Whenever a distribution covers a small range of scores about the mean, the standard deviation will be small. The closer the scores cluster about the mean the more homogeneous (similar) the scores. As the distribution spreads out, the standard deviation becomes larger, and the scores are heterogeneous (varied).

In order to expand on the concept of the standard deviation and how you can use it to compare across groups, I'll give you an example which emphasizes the idea of precision using the mean as a reference point.

In Yonker, p. 10, you will see hypothetical scores made by two artillery batteries when firing at a designated target: a positive score means the shell went beyond the target, a negative score means the shell fell short of the target. Zero is right on target. The frequency distributions (the number of times each value occurs) for both batteries are listed. The means indicate that both batteries, on average, were on target! But, the standard deviations are clearly different. Using the mean as a reference point, which
artillery battery has the best precision? **Battery B has smaller standard deviation**; therefore, has more precision when firing at the designated target. Even though both batteries have the same mean accuracy, the shelling of battery B showed less dispersion around the target!

**Coefficient of Variation:**

This is a measure of variability that is useful for comparing distributions when both the mean and standard deviation are different between the distributions (unlike the artillery batteries above that had the same mean). The Coefficient of Variation has two functions:

1. Shows how representative the mean is of the distribution.
2. Shows the difference in the variability in one data set when compared to another.

The formula for the Coefficient of Variation and function number one are illustrated in the following PowerPoint slide:

[Coefficient_of_Variation.ppt]

<table>
<thead>
<tr>
<th>Coefficient of Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>n Divide the standard deviation by the mean.</td>
</tr>
<tr>
<td>n The smaller the decimal fraction this produces, the more representative is the mean for the total distribution</td>
</tr>
<tr>
<td>n The larger the decimal fraction, the worse job the mean does of giving us a true picture of the distribution.</td>
</tr>
</tbody>
</table>

The second function is very important in cases where comparing the standard deviations of two different distributions is not realistic. Consider the following:

Which group of pups show more variability in weight gain?

**Great Danes**—
Mean = 30 pounds
Standard Deviation = 10 pounds
Coefficient of Variation = Standard Deviation/Mean
10/30 = .33

**Chihuahua**—
Mean = 3 pounds
Standard Deviation = 1.5 pounds
1.5/3 = .5

The higher ratio is found in the Chihuahua pups (.5 compared to .33 for the Great Danes); therefore, the Chihuahua pups show more variability in weight gain!
Week 3 - STANDARD SCORES

Standard (z) Scores:

Last session, we looked at the Standard Deviation and you saw how it could be used to compare two (or more) groups that have the same units of measure and the same reference point or mean (remember the artillery batteries!).

We also covered the Coefficient of Variation and you saw how you could compare two (or more) groups, which have different reference points (mean) and different standard deviations. The Coefficient of Variation is used when just comparing standard deviations is not convenient or realistic (remember the pups!). Both the standard deviation and coefficient of variation are general comparison measures of variation between groups.

Now, suppose you want to find the standing of an individual within a group. A descriptive measure called a z score allows you to do either one of the following:

1. Find where one or more individuals stand in reference to the mean of a single distribution on one unit of measure (one variable).
   Example--Where an individual is located relative to a distribution of test scores.
2. Find where one or more individuals stand in reference to the mean of two (or more) different distributions that may have different units of measure.
   Example--Where an individual stands relative to two tests, each given in a different class (different distributions).

The measure used to make these comparisons is called a Standard Score or z score. A z score is a way of standardizing measures and is very useful when you have distributions with different units of measure and you want to make comparisons between distributions.

A z score tells you how far above or below the mean any given score is in standard deviation units. A standard deviation unit is another term for a z score. You can compute a z score for any score in a distribution and compare it to another score in the same or a different distribution. But, z scores are most useful when the shape of your actual distribution of scores is nearly normal (remember the bell-shaped curve!) The normal distribution serves as the model for standardizing scores. It is a mathematical model that contains constant percentage areas (these are given!) within which you can find where a certain score is located given the mean and standard deviation. See Yonker p. 11.

As you can see, the area under the normal curve is divided into standard deviation units with corresponding percentages. The numbers at the bottom of the curve (-3, -2, +1, +2, etc.) are the values of z, and each one represents a standard deviation unit. The mean has a z score of 0. If we go one standard deviation unit (a z of +1) to the right of the
mean, we always find 34.13% of the cases in a distribution. If we go to the left, we find
the same thing at one standard deviation unit below the mean (a z of –1). Therefore,
about 68.26% of our cases will lie within + or - 1 standard deviation unit above
and below the mean.

A bit over 95% of our cases fall within + or - 2 standard deviation units above and below
the mean.

And, about 99% of our cases fall within + or - 3 standard deviation units above and
below the mean.

The rest of the cases in a distribution fall beyond a z score of +3 or -3. And, because the
normal curve is a theoretical curve, the curve does not completely touch the X axis,
meaning it goes on to infinity!

Using the standard normal curve and z scores we can see where any score “fits” or is
located relative to the mean of a normal distribution. We do this by transforming a score
(X) to a z score. The following is the formula for computing z:

\[ Z = \frac{\text{Score} - \text{Mean}}{\text{Standard Dev.}} \]

Example: In a university, where does the graduate assistant with the larger stipend (X)
“fit” in a distribution of other students with stipends?

\[ \text{Mean} = \$17,940 \]
\[ \text{Standard Deviation} = \$4,960 \]
\[ X = \$25,000 \]

Look at the normal curve (in Yonker, p. 12) to get an idea where 25,000
would be located.

17,940 is the mean of the distribution and has a z score of 0.

17,940 + or – 4960 (one standard deviation) gives stipends between
12,980 and 22,900.

17,940 + or - 2(4960) gives stipends between 8020 and 27,860.

**25,000 (the stipend of interest) is located between 1 and 2 standard
deviation units (z scores) above the mean.**

Let's find the exact location of this stipend by calculating z.
\[
\frac{25,000 - 17,940}{4,960} = \frac{7,060}{4,960} = 1.42
\]

So, why does the mean have a z score of 0? You can see by the calculations below!

\[
17,940 - 17,940 = 0
\]

Did you ever wonder where those percentages come from when you get your GRE results (or other educational test)? These percentages are found through z scores. By using z scores we can determine the percentage of cases:

1. between the mean and a z score.
2. beyond (above) the z score, and
3. below the z score.

In order to do this, go to p. 37 in Yonker and you will find a z score table. Using the example, above, locate the z score (in column A) of 1.42.

**Column B** will answer number 1 above when changed to a percentage. The percentage of the cases that fall between the mean and z is 42.22%.

**Column C** will give you the answer to number 2 above when changed to a percentage. The percentage of cases that fall above z is 7.78%. So, not many students (7.78%) have achieved a higher stipend than our student of interest. Notice that column B plus column C always equals 50% (0.5000).

In order to find number 3 above, the percentage of cases below z, you must do a minor calculation. Since you know from the discussion of the normal distribution that 50% of the cases fall below the mean (all Z less than zero), and you also know that 42.22% of the cases fall between the mean and z, then 50 + 42.22 =92.22. So, 92.22% of the students get a stipend below $25,000. This is descriptive information about the distribution. Quite a few students have lower stipends!

This example with the stipends showed how you can compare an individual value to the mean of a distribution with one unit of measure (dollars). A very common use of z is when you have distributions with different units of measure and you want to make comparisons between distributions.

A very good example of this is in a book called *The Geography of Reading*, by Louis Brown Wilson (1938, University of Chicago). Wilson did and extensive study of the distribution of library resources in the states of the nation (see Yonker p. 13).
Specifically, he compared states on four factors considered important for library development. He had four distributions of the states on

1. percent of population served
2. per capita (for each person) circulation
3. per capita expenditure
4. per capita volumes

In order to compare the states on these variables, he did a couple of things:

1. He standardized the units of measure on all variables to z scores for each state.
2. He averaged the z scores for each state on 3 factors (circulation, expenditures, volumes) to get an index of library development.
3. He then ranked the states on this index of library development.

You can see where Pennsylvania "fits" on the index compared to all of the other states, and, you can see where PA "fits" on each variable compared to the other states. You can also compare PA to itself among the different variables.

If you look very closely at this table, you will see we had only 48 states! Remember this was an old study, but it shows a nifty use for z scores!

**Transformed z Scores (T Scores):**

For another use of z scores, see PowerPoint slide below—

[TScores.ppt]

Transformed z scores

- AKA: Standardized or "T" scores
- Z scores are transformed
  - How to find T Scores: Multiply each z score by the desired standard deviation and add the desired mean (usually 10 and 50, respectively)
- This is used a lot in Psychology and Education
- Benefits: gets rid of negative connotations of negative and zero scores.
SAMPLING METHODOLOGIES

At this point we are going to leave descriptive statistics and get into sampling methodologies. It is very important to understand sampling strategies before we get into inferential statistics. Sampling strategies have much to do with what types of conclusions you can make with your data.

In order to understand the meaning of a sample, and why we need sampling methodologies, you should know the meaning of a population. A population is formally defined as any aggregate or complete set of persons or things that have traits or characteristics which we want to measure. Essentially a population is the totality of a defined group that you (as a researcher) are interested in studying.

Examples:

- All students at the College of Information Science and Technology, or all graduate students at IST, or all online students.
- All reference books at Hagerty Library
- All microcomputers being used for educational purposes in US high schools.
- Would you believe a study of the informational content of soap operas (all soaps would be the population)? Yes, there was a study like this!!

Values, which are associated with populations, are called parameters. And, you can check your symbol sheet (Yonker, p. 4) for the symbols associated with population parameters: mean, variance, and standard deviation.

Many times when doing research we may want to make inferences to a population from the information we gather on a sample. A sample is a subgroup of the population chosen to represent it. There are two important, but related reasons why we sample.

1. population is too large
2. not enough time or money to study the whole population

When you decide to take a sample and collect measures on that sample, the numbers generated from those measures are called "statistics." From your symbol sheet, you can see the symbols associated with sample statistics: mean, variance, and standard deviation.

Essentially, a researcher wants to make inferences about population parameters from sample statistics; the purpose of inferential statistics. In order to make these inferences, a good sampling strategy is needed because the sample should be an accurate representation of the population. A good sampling strategy is based on the Principle of Random Selection.
Random selection is a procedure by which each member of the population has an equally likely chance of being chosen as any other member of the population. In other words, the sample should be representative of the population and not biased! Consider the following example:

An action researcher wants to study the attitudes towards service of the population of library users at a university (with a sample of 100 library users). The researcher decided to go to the library every Tuesday and Thursday at 11:00 A.M. for one term and ask the people "who looked most willing" about their feelings towards the library's service.

Obviously, this is NOT a RANDOM, representative sample!! Think about why.

In the discussion below, you will read about four random sampling methodologies, and then we will briefly discuss three other techniques that are not random.

Random techniques:

1. Simple random sample
2. Stratified random sample
3. Cluster sample
4. Systematic sample

Not random:

1. Judgement
2. Quota
3. Accidental

Random sampling methodologies are known as probabilistic techniques and non-random methodologies are known as non-probabilistic techniques. See PowerPoint slide as supplement.

[Prob_and_NonProb.ppt]

- Probabilistic sample - sampling in which the probability of each element in the population being selected is known and can be specified. Each element has the same chance
- Non-probabilistic sample - probability not known a priori. Convenience samples, available samples
Random Sampling Strategies:

Simple Random Sample:

Example: We have a population of 480 graduate students in the science and technology division of a university. We want to take a sample of 80 students and examine their microcomputer use. The following are the steps in selecting a simple random sample:

1. Number the students from 001-480. Because the largest number is three digits, lead zeros must be used.
2. Consult a table of random numbers. See Yonker, p. 39.
3. Although a researcher usually goes to a random place in the table, for simplicity, start with the first three digits of the number at the top left of the table (ignore row number). This number is 100.
4. The first student in the sample is student 100. As one consistently moves down the column (and, one should be consistent!) the next student in the sample is student 375. Any number “hit” in the random numbers table must be 480 or lower (the size of the population).
5. One would keep moving through the table, until 80 (unduplicated) numbers have been selected—these numbers represent the students in the sample.

The good news is that some computer programs will generate random numbers automatically for you, and even the sample itself, if you enter the population size, and your sample size.

In selecting a simple random sample, the two biggest problems encountered are an accurate list and a numbering system. Considering an accurate list:

1. Cases should be included only once in the list.
2. The list may be outdated—often the case with a telephone directory.
3. The list may not include everyone—once again, the telephone directory.

Considering a numbering system: A population of 10,000 people identified only by social security number would pose problems for the researcher. The social security number would be very difficult to match to a table of random numbers! So, another unique identifier would have to be provided!

There are two other random techniques that break up a population into subgroups within which random samples are drawn: they are stratified samples and cluster samples.

Stratified Random Samples:

In stratified samples you classify or divide your population into categories or homogeneous groups that are relevant for study. Relevant is a key word! If you were studying high school students using the library, it is doubtful that you would stratify them
according to hair color, but you might stratify them by program, such as academic, business, etc.

There are two types of stratified random samples: proportional and disproportional. The technique you choose depends on the purpose of your research.

Example of proportional stratified sampling:

An action researcher works in the office of computing services and is interested in the microcomputer experience of incoming freshman by the major they have chosen. A freshman orientation class with 100 students is the focus of the study and consists of the majors listed in the table (Yonker, p. 14). The researcher wants a sample of 20 students that is proportional to the population and will survey (interview) each student in the sample on his/her computing experience. An accurate representation (like a profile, or snapshot) of the group is needed, so a sample should not over represent or under represent each major from the population. The following are the steps to obtaining the sample:

1. Calculate the percentage of each major in the population (you should review getting a percentage from last session, but an example is shown at the bottom of the table). The percentages are shown in the table.
2. Take the percentage of each type of major times 20 (the sample size). You can see the number in each major represented in the sample (last column).
3. Randomly select within the strata. For instance, randomly select 10 out of 50 education students.

Suppose a researcher wants to equally represent these types of majors in the sample and disproportional to the population. As you might expect this is called disproportional stratified sampling. This type of technique could be used if a program, such as a microcomputer-training program, was evaluated per major on errors in computing. An equal number of students in each group would be useful for determining if the training program is more effective per different type of major.

In this example, a researcher would over sample some strata and under sample others. With a sample of 20 students, you would have 5 students randomly selected from each category. Of course, all of the Science and Technology majors would be selected.

Cluster Sampling:

This type of sampling is often used for the purpose of a needs assessment study (a Descriptive Study) over a large geographic area. A public library, public school system, or community health facility may undertake a needs assessment study and use cluster sampling to obtain a random sample.
Example:
A planning commission, such as Delaware Valley, can provide a researcher with a list of Census Tracts for a city or region, such as Philadelphia or Camden, New Jersey. For each Census Tract, city blocks are listed and addresses can be obtained. The PowerPoint slide shows the steps in taking a random sample in clusters.

Cluster Sampling
- Randomly select n (certain number) of census tracks
- From randomly selected census tracks, randomly select n blocks
- From randomly selected blocks, randomly select addresses
- Interview the family--unit of study

Cluster sampling is different than stratified sampling in that the clusters are often regional areas and not homogeneous groups or strata relevant for study. A researcher hopes that the clusters are a heterogeneous group to get a good representation of a region; however, sometimes clusters, like census tracts, are unequal in size and could contain a concentration of a certain group of individuals.

Systematic Sampling:
This is a convenient technique if you have a large unnumbered list with no categories or clusters in which to random sample. With systematic sampling, you randomly select the first element in the population as a starting point and then move ahead systematically at a fixed sampling interval. The sampling interval depends upon the size of the sample you would like and the size of the population.

Example:
You want to sample 40 cards from an order file (card catalog) of 1000 cards to get a sense of the types of subjects covered. The following are the steps to obtaining a systematic sample.

1. Calculate the sampling interval:
Size of population/Size of sample. 1000/40 = 25

2. Select the first element at random from within the sampling interval (the first 25 cards). By looking at a random numbers table, let's say the number 10 is found.

3. The tenth card is the first card in the sample. From the tenth card you move ahead systematically by the sampling interval (25) until you reach the desired sample size. The second card is number 35, the third card 60, and continued.
An important note: Systematic sampling assumes a random order in the list of cards, people, etc. So, if there is an ordering principle to the list (as in a card catalog, or any alphabetical list), it is important to do a quality assurance check of the resulting sample to make certain that a particular group of people, etc., is not over represented (as could happen through a concentration at a certain place in the alphabetical list!).

When Systematic Sampling goes wrong—the Corner House Phenomenon! When using systematic sampling along a city or town block, often the sampling interval lands right on the corner house. Think about it—what is different about the corner house when compared to the other houses on the block?

In general, although random sampling is the best strategy for obtaining a representative sample, it is always a good idea to do a quality assurance check of any sample—even random!

Non-random Sampling Strategies:

These strategies are useful for studies that do not require generalization to larger populations. Fact-finding, or quick marketing studies may use these strategies. However, when using statistics that assume a random sample, a researcher must proceed with caution in making conclusions if the sampling strategy used was not random! Some of the analyses that assume a random sample will be discussed in this class. The following is a decidedly brief account of the non-random techniques:

Judgment Sampling:

A judgment is made concerning individuals who have knowledge about a topic of interest. Because the individuals have the required knowledge on a topic, they are asked to be in the sample. When I worked in community mental health, we conducted a key informant study to assess the mental health service needs for children and youth. Our “key informants” were individuals who knew something (assumed) about youth services, such as staff from the school system. These individuals were not selected at random!

Quota Sampling:

This technique seems like the non-random version of stratified sampling. With this strategy, characteristics of the population are specified (e.g. age, gender, etc.) and quotas are specified for each group. Quotas may represent (be proportional) a guesstimate of the number in the population for each group. The key is that the quotas should be satisfied, but not by selecting at random.

Interview 20 females and 20 males over the age of 65!
Accidental Sampling:

The best way to illustrate this technique is with an example that probably all of you have experienced: the individuals with the clipboards ready to grab you in a mall! Anyone who comes along is a potential target and you may receive a few dollars for complying. Most of the studies I have experienced are quick marketing studies (I remember a study regarding brands of pet food.). Of course, you can refuse to be interviewed!

SURVEYS AND DATA COLLECTION

A survey is a general type of methodology that can be used for action research. Most of the sampling techniques we just covered could be used to conduct a survey depending on the purposes of the data collection. A survey is often thought of as a Descriptive Study (remember the nine strategies of research from the first week?).

A survey usually refers to a large data collection effort:

What it involves—personal interviews, telephone interviews, a questionnaire sent through the mail, document survey, literature survey, social area analysis (observation and description of different areas of the city)

“Who” it involves—community, customers, users, employees, literature

Purpose—information gathering and fact finding to

Describe what exists (such as public library services)

Establish need

Identify problems

Imply possible solutions

In this discussion of surveys, we will briefly cover:

Three survey methodologies—the advantages and disadvantages of each method

The structure of a survey methodology

Content and wording of questions

Ordering of questions
Survey methodologies:

Mailed Questionnaire:

Advantages--

1. More economical than the other techniques. As quoted in Selltiz (1976, p. 294), “Questionnaires can be sent through the mail; interviewers cannot.”
2. Faster if response rate is not a major issue and follow-ups are not necessary.
3. Can cover a wide range of issues.
4. Can cover a widely spread sample because there are no geographic limitations.
5. Avoids interviewer bias.
6. Anonymity

Disadvantages--

1. No opportunity for respondents to request question clarification.
2. No opportunity for probing for more detailed responses.
3. Could include responses from more than one person, or not the intended person.
4. No opportunity for observation of the respondents.
5. Response rate can often be less than 50%.

Personal Interview:

Advantages--

1. Can explore complex issues through rich qualitative information.
2. Can clarify questions and responses.
3. Can establish rapport with respondent.
4. Characterized by a higher response rate because, generally, people enjoy talking.
5. Allows for observation of respondents.

Disadvantages--

1. Subject to interviewer error and bias, so training of interviewers is important.
2. Subject to problems of question uniformity because of different interviewers, so training is important once again.
3. No anonymity
4. Difficult to analyze.
5. More time consuming.
Telephone interview:

Advantages--

1. Some anonymity when compared to a personal interview, but less than a questionnaire.
2. Economical
3. Can be completed fairly rapidly.
4. Characterized by fairly high response rates, again, because people like to talk.
5. No travel time.
6. No geographic boundaries.

Disadvantages--

1. People may be difficult to reach (especially individuals with a Caller ID system!).
2. Can only include people with phones, or perhaps those with accessible phone numbers.
3. No opportunity for observation of respondents.
4. Some interviewer error and bias is possible.

Structure of a Survey Methodology:

You should review the discussion of qualitative and quantitative data in week one concerning the structure of an interview or a questionnaire.

Generally, open-ended or unstructured questions are called for when the topic of a study is complex and qualitatively rich responses are needed. A personal interview is usually designed with unstructured or semi-structured questions. In an unstructured interview, only the first question is standard for all respondents. The remaining questions are determined by the answers of each respondent. In a semi-structured interview, the questions are open ended, but all of the respondents receive the same questions.

When a topic is certain or fairly well explored, a structured methodology is called for. Structured methodologies are generally a questionnaire sent through the mail or a telephone interview.

Content and Wording of Questions:

See Yonker, p. 15, for some typical problems in asking questions. Most of these items “speak for themselves;” however, I have included some additional comments about some of the items below.

Items a. and b. should be designed using structured questions, with responses in categories or on a scale.
Item d. could also be designed using a structured question that included a scale of "strongly agree" to "strongly disagree" at either ends of the scale (Likert-type scale).

Item e. is called a double-barreled question, meaning two questions are included in one.

Ordering of Questions:

The following are some recommendations on ordering questions in an interview or a questionnaire:

1. Start with easy questions that the respondent will enjoy answering. You want to prevent boredom early on while building rapport and putting the respondent at ease.
2. Try for an easy and natural flow over topics. Place like items together and give a brief explanation when a topic breaks.
3. Within topics, go from the general to the specific. For example, start with questions on use of the Internet in general, then move on to specific questions about the use of search engines.
4. Put open-ended or difficult questions (if any) at the end of the interview or questionnaire.
5. Put questions on “sensitive” matters (such as age or income) at the end of the interview or questionnaire. Otherwise, the interview may be over before it has started!
Week 4 - DESCRIPTIVE STATISTICS--MORE GRAPHS

Through your homework, you have seen what the frequency distribution of exam scores looks like. Now a couple of additional graphic techniques will be introduced to show how this distribution looks in "picture" form.

**Stem and Leaf**

The first graphic technique is called the **Stem and Leaf**, which was developed by John Tukey. The Stem and Leaf is a frequency distribution graph. Why is the stem and leaf useful when you can just as easily use a histogram? Well, often distributions which are plotted using the histogram are in class intervals (I mentioned class intervals in week two--FYI) because the number of cases is so large. When you have intervals, you lose the value of the original score. Remember the bookmobile histogram? In this histogram (see PowerPoint slide below), a bar is placed over the midpoint of the class interval, so you lose the original value of the score.

[Histogram.ppt]

**HISTOGRAM OF BOOKMOBILE STOPS**

![Histogram](image.png)

The **stem-and-leaf** technique is a method whereby you can retain the original score (for smaller numbers), its frequency of occurrence, and still see the shape of the distribution. The PowerPoint slide shows a stem-and-leaf plot of the exam scores. You may want to
look at the frequency distribution of the exam scores (from your homework) along with the Stem and Leaf.

[StemLeaf.ppt]

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Stem &amp; Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.00</td>
<td>Extremes (=&lt;46)</td>
</tr>
<tr>
<td>1.00</td>
<td>5 . 2</td>
</tr>
<tr>
<td>.00</td>
<td>5 .</td>
</tr>
<tr>
<td>3.00</td>
<td>6 . 013</td>
</tr>
<tr>
<td>3.00</td>
<td>6 . 678</td>
</tr>
<tr>
<td>4.00</td>
<td>7 . 0004</td>
</tr>
<tr>
<td>11.00</td>
<td>7 . 5667778899</td>
</tr>
<tr>
<td>8.00</td>
<td>8 . 11112234</td>
</tr>
<tr>
<td>4.00</td>
<td>8 . 6778</td>
</tr>
<tr>
<td>3.00</td>
<td>9 . 224</td>
</tr>
<tr>
<td>1.00</td>
<td>9 . 8</td>
</tr>
</tbody>
</table>

Stem width: 10
Each leaf: 1 case(s)

1. The numbers to the left of the point (.) are called the stem (the numbers are under the word Stem.). Since our scores are all units of ten, meaning 50, 60, 70, etc., the stem width is 10. A stem width could represent hundreds or even thousands.

2. The numbers to the right of the point are called the leaves (under word Leaf) and each number represents a single unit (one case), so look at the 4th row from the top, you can see we have a frequency of 3 (and there are 3 leaves). The three numbers are 60, 61, 63. In row 7, with a frequency of 11, you can spot one of the modes of the distribution—77, which occurs 4 times.

3. You can see the note at the bottom of the graph showing the stem width and that each leaf represents one case.

4. Notice that values that stand apart from other scores in the distribution are not identified as a number in the stem, but are called extremes. There are two extremes that are less than or equal to 46.

5. SPSS uses an algorithm that looks at the variability of the values in the distribution and forms a scale for the stems that best displays the shape of the distribution when the leaves are added. (You do not need to know this algorithm, which I am sure is a relief!)
**Boxplot**

Another graphic technique, which will be illustrated using the same exam data, is called a **Boxplot** (also developed by Tukey). The Boxplot is not a frequency distribution graph. A Boxplot is a way to visualize a distribution when you do not need as much detail about the individual values. It is very compact, and gives you a quick view of the distribution. What does it tell you?

![Boxplot](BoxPlot.ppt)

1. The box represents the middle 50% of the cases--those that are in the center of the distribution. The edges of the box are called "hinges" and they show scores that range from the 25th percentile, which is the lower edge of the box, to the 75th percentile, which is the upper edge of the box (this is called the **interquartile range**). The length of the box, from top to bottom, gives you a sense of the range for the center of the distribution of scores, or the spread of the central values. Note that the width of the box does not represent anything.

2. The line inside the box represents the median. Remember that the median was 77.5 and you can see that from the Boxplot.

3. The lines that go from the edges of the box are called "whiskers" and they show the largest and smallest values that are not extremes. The largest value is 98 and the smallest value, which is not an extreme, is 52.

4. The O’s (the number beside the O is the record number) below the whiskers are called outliers and represent values that are from 1.5 to 3 box lengths from the edges of the box. We had two outliers in this distribution-- 46 and 40.

5. Values greater than 3 box lengths from the edge of the box are called extremes and are marked with an asterisk. We do not have any extremes in this distribution.

6. A Boxplot is very useful when you have more than one distribution and you want to make quick comparisons among the distributions. See the slide below that shows you Boxplots of the variables from the Medical Libraries study.
**Frequency Polygon**

One more graphic technique is called a **frequency polygon**. This is another frequency distribution graph. See I.Q. distribution in the attached EXCEL document

The polygon is constructed much the same as a histogram, including the "¾ high rule," except a dot instead of a bar is placed at the frequency with which each value occurs and a line is drawn between the points. You can see this polygon shows the actual value of the score, its frequency, and the shape of the distribution, but the intervals are

---

**Frequency Polygon of Class I.Q.**

Data set is:

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>110</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>120</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>123</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>125</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>126</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>128</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>130</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>132</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>135</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>140</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

N=21
unequal between data points. If you look at the horizontal (X) axis, you can see, for example, the numbers go from 101 to 110, and then from 120 to 123 (unequal intervals!).

So, if a histogram and a polygon are constructed in a similar way (except one has bars and the other has dots and lines), why do we need both? Generally, a **Histogram** is used for discrete data (counting units, like the bookmobile stops) and the **Polygon** is used for continuous data (measurement, like the I.Q. distribution).

**INFERENTIAL STATISTICS**

Inferential Statistics, an Introduction:

At this point we are going to leave descriptive statistics and move into **inferential statistics** and talk about hypothesis testing and confidence intervals. Inferential statistics involves taking sample estimates and making inferences to usually unknown population parameters. There are two types of sample estimates: 1. A point estimate 2. An interval estimate.

An interval estimate refers to a confidence interval and will be discussed later in this session. Point estimates are used in hypothesis testing. A common point estimate is the sample mean. The sample mean is one number (a point) that is used as an estimate of usually an unknown population parameter. We can test hypotheses about a population mean through use of a sample mean—this is accomplished through inferential statistics.

During the first session, two types of hypotheses were mentioned and you should review your notes on that discussion. The two types of hypotheses are:

1. **Research hypothesis**—what the researcher assumes

   Computer-aided instruction will influence math test scores in comparison to standard instruction.

2. **Null (statistical) hypothesis**—an assumption that no differences exist.

   There is no difference between computer-aided instruction and standard instruction and the influence on math test scores.

**Important note!** We collect sample data and apply inferential statistics in order to reject the null hypothesis if it is probably false. But, because these tests are based on probability (which indicates some uncertainty or risk), we can never prove (100%) that the null hypothesis is false. We can just provide results that would probably not occur by chance.

Hypothesis testing involves three interrelated distributions:

1. **Sample Distribution:** you have seen that this is the distribution of data in a sample or the individual values/scores that make up a sample.
Example—you ask a sample of 2000 high school students the number of books they read in a year and plot the numbers in a frequency distribution graph, such as a histogram. This is the sample distribution.

2. **Population Distribution:** you have see that this is a distribution of all cases in the population on the variable of interest. Usually this distribution is unknown.

Example—you ask every high school student in the US how many books they read in a year. If you were physically able to plot the results in a frequency distribution graph you would have the population distribution!

3. **Sampling Distribution:** this is probably a new concept for you. It is the theoretical probability distribution of a sample statistic like the mean. This is a model created by statisticians for use in hypothesis testing. It allows you to take one sample and calculate a mean (a point estimate) and make inferences to the population from which the sample was drawn. I have said before that we will not discuss probability theory in detail, but we will use it, so key words to describe this distribution are theoretical probability! So, how would a statistician create the sampling distribution for the mean (the sample statistic).

Example—creating a sampling distribution involves randomly selecting many, many large samples of a fixed size from a population. A statistician could take thousands of large samples (N over 30 and of a fixed size, say 45) at random from the population of high school students and ask every student in each sample how many books they read in a year. For each sample, a mean could be calculated resulting in thousands of means (mean of sample 1, mean of sample 2……mean of sample 1000, etc). If all of these means are plotted in a frequency distribution graph, there you have the sampling distribution for the mean!

The theory of the sampling distribution states, once again, that if your samples are large (greater than 30) and of a fixed size, your plotted means will take on a highly predictable form from which you can test hypotheses. This highly predictable form looks like the normal distribution except the plot is not of individual scores, but of individual means from samples (see Yonker, p. 16). The plot shows that some means will cluster around the center and occur more frequently. Other means are on the tails of the curve and do not occur very frequently. And, remember the percentage areas under the normal curve that you learned when we discussed z scores? Well, those same percentage areas under the normal distribution apply here as well. So, about 68% of the sample means will cluster near the center of the sampling distribution.

The theory of the sampling distribution also states that not only will the plot of sample means be relatively normal, but the true population mean will occur at the center of the distribution--the mean of the means! This theory that the central most commonly occurring sample mean will be a very good estimate of the true population mean is called the Central Limit Theorem and is described in more detail in Yonker, p 17.

There are sampling distributions for all statistical analyses, and again, these distributions were created by statisticians. We will concentrate only on the sampling distribution for the mean. The whole point of a sampling distribution (and this one in
particular) is that if you know how sample means distribute (behave) then you can make
generalizations/inferences to the population from which they were drawn and test
hypotheses about population parameters.

How sample means distribute is our sampling distribution of the mean based upon the
central limit theorem. It is a theoretical probability distribution. It allows us to take
one random sample, calculate a sample mean (a point estimate) and test a hypothesis
about a population parameter. Using the sampling distribution we can assess which
population parameters are likely or unlikely (probability!) given one sample. So, there
are mathematical principles of probability behind this distribution that we will not dive
into; therefore, when using this sampling distribution for hypothesis testing, you will
have to accept the distribution and related concepts at face value. You should have
some sense of what a sampling distribution is and how to use it rather than why it
works. But, if you are interested in the mathematical backing behind these concepts a
good (not overly technical) discussion is in H. Loether and D. McTavish, Descriptive and
Inferential Statistics.

**Standard Error of the Mean:**

Now, if you consider a normal distribution of individual scores, every person does not
fall on or near the mean!! And, the standard deviation is a measure of the variation of
individual scores around the mean and each standard deviation unit represents a z
score. So, do you think that every point estimate (sample mean) that we calculate will
give us the true value of the population mean, or occur near the center of the sampling
distribution? Of course not--a sample mean could fall anywhere on the curve (just like
an individual score). A sample does not represent every member of the population, so
there will be some sampling error (errors produced by chance) in each sample. The
measure of sampling error is called the standard error (specifically, the standard error
of the mean). The standard error is formally defined as the standard deviation of
the sampling distribution of means. It functions much like the standard deviation, but
for a different distribution (the sampling distribution). And, the sampling distribution is
divided into standard error units instead of standard deviation units.

The following is the formula for the standard error of the mean:

\[
\text{standard error of the mean} = \frac{s \text{ (standard deviation)}}{\sqrt{N}} \text{ (or lower case 'n').}
\]

As you can see from the formula, the standard error is a function of N and the standard
deviation. The larger the N, the smaller the standard error. The larger the standard
deviation, the larger the standard error.

Up until this point, the only difference mentioned for the population standard deviation
and the sample standard deviation is the symbols used for each measure. In calculating s (sample standard deviation) to be used in inferential statistics there is a
slight modification in the standard deviation formula which you will find in Yonker p.
18. You can see both the mean-deviation formula and the raw score formula as covered in session two (top row, first and third formulas). The middle formula is the raw score formula without the f (frequency)—we did not cover this form of the formula, but you may see it in some stat books. The bottom row illustrates a slight modification to all of the formulas and you can see that modification is N – 1.

FYI (I will not test you on this): N – 1 is based on the concept of Degrees of Freedom (df) (SPSS uses this modification, which is why the figures were a bit different from your “by hand” calculations in Assignment 3). Degree of freedom is another mathematical principle that refers to the number of restrictions placed on your data each time you make an estimate based on a sample statistic. Basically, this means that anytime you make an estimate based on a sample you use up some information in the sample; therefore you lose some “freedom” with your data. When you get to the point of calculating the standard deviation, you have already made a sample estimate in calculating a sample mean, so a degree of freedom has been lost! Without N - 1 you would have an underestimate of the standard deviation based on that loss of information (freedom). N – 1 gives you an unbiased estimate of the standard deviation. During this course, you will see degree of freedom show up in different forms, but the concept is always similar as described above.

**Inferential Statistics, the z test:**

Now that we have the concept of the sampling distribution and standard error, you will see how they work in hypothesis testing. We are going to use a technique called the z test (one of our inferential statistics) to illustrate hypothesis testing. The z test assumes a sampling distribution constructed with large samples (over 30).

**Example:**

A librarian in a high school is interested in the reading habits of seniors in the academic program. The librarian has been at this school for a long time. Fifteen years ago, a survey was conducted of all seniors in the academic program and the population mean for the number of books read in a year was 30. It is present time and the librarian is still interested in the reading habits of seniors in the academic program, but the number is too large to survey the entire population. So, a random sample is taken to see if the average has changed from fifteen years ago. The question is, are the students still reading an average of 30 books per year?

The null hypothesis states that there has been no change—students are still reading about 30 books per year. Or, there is no difference between the sample mean and the population mean.
What would our librarian’s research hypothesis be? A distinct possibility is that high school seniors in the academic program are reading fewer books per year (think of the Internet, video games, cable TV, and many other “opportunities!).

A random sample of 35 students (N=35) results in the following data.

Population mean, as previously mentioned = 30

Sample Mean = 11.3

Sample Standard Deviation (s, calculated using N –1 in the denominator of the standard deviation formula) = 22.35

From this data the question is asked--Does this sample mean of 11.3 really come from a population with a mean of 30.? The null hypothesis says yes, the research hypothesis says no.

The following are the steps for calculating the z statistic:

1. Calculate the standard error of the mean, \( \frac{s}{\sqrt{N}} \)

\[
\frac{22.35}{\sqrt{35}} = 3.78
\]

In Yonker, p. 19, you can see that the sampling distribution is divided into standard error units (each representing a value of the z test, +1, +2, -1, -2, etc.). Since we are testing to see if 30 is still a likely population parameter, that number is placed at the center of the sampling distribution. On each side of 30, the standard error is added and subtracted to provide a range of possible sample means (remember the standard error is like the standard deviation of the sampling distribution); therefore, 30 + 3.78 = 33.78 and 30 – 3.78 = 26.22, and so on.

2. The sample mean of 11.3 is beyond 18.66 (or a z value of –3). We can find the exact location of our sample mean on the sampling distribution with the formula for the z test (can only be used when N>30):

\[
z = \frac{\text{sample mean} - \text{population mean}}{\text{standard error of the mean}}
\]

\[
z = \frac{11.3 - 30}{3.78} = -4.95
\]

Given this value of z, can we (and our librarian) reject the null hypothesis? We will not know until we have compared our obtained value of z, which is –4.95, to a critical
value of z. Critical values correspond to a significance level, which is specified before data collection. A significance level allows us to specify the amount of risk we are willing to take in rejecting the null hypothesis that may be true. You should remember that hypothesis testing involves some uncertainty or risk. There are two standard significance levels for the z statistic (and other tests as well):

1. The 5\% (.05) significance level corresponds to a critical value of z = 1.96. This significance level is generally suitable for social and behavioral sciences. In research reports you will see p < .05.
2. The 1\% (.01) significance level corresponds to a critical value of z = 2.58. This significance level is used in drug studies, clinical trials, or other medical research, and is a more conservative, rigorous test of significance. In research reports you will see p < .01.

Suppose the significance level was set at 5\%. Look at the sampling distribution in Yonker, p. 20. The "shaded" portion in the tails of the sampling distribution is called the critical region of rejection corresponding to the 5\% significance level. You can see that each tail corresponds to 2.5\% of the curve (2.5 + 2.5 = 5\%), which is called a two-tailed test of significance. The region of rejection is an area in which if your obtained value of z falls, you can reject the null hypothesis.

The obtained value of z = -4.95 is beyond our critical value of 1.96 (on the negative end of the tail) and falls into the region of rejection. Therefore, we (and our librarian) reject the null hypothesis at the 5\% significance level. Since we have rejected null, we can conclude that it is unlikely we would get a sample mean as low as 11.3 if the sample truly came from a population with a mean of 30. It is not likely that the population mean is still 30. It seems seniors in the academic program are reading fewer books as was assumed in the research hypothesis.

In technical terms, these results tell us that there is a significant difference (statistically significant) between our population mean of 30 and our sample mean, and these results are not likely to have been produced by chance. But, at the 5\% significance level, there is a five percent chance we could be wrong in rejecting the null hypothesis--This is the Risk or Uncertainty.

**Inferential Statistics, Types of Errors:**

Part of the risk of rejecting or failing to reject (accepting) the null hypothesis is the possibility of one of two types of errors. They are a Type I Error and a Type II Error:

A Type I Error happens when the null hypothesis is rejected when it is probably true. This is saying there is a significant difference when there probably is not.

At the 5\% significance level, which we just used in the example, we could mistakenly reject the null hypothesis 5\% of the time (a 5\% chance of making a Type I Error). So, if the significance level is set at 1\%, you decrease your risk of making a type I error (a 1\%
chance of a wrong conclusion). In the preceding example, if the sample mean of 11.3 really did come from a population with a mean of 30, a Type I Error has been made.

There is a trade off between significance levels. The obtained value of z was equal to -4.95, so it was beyond the 5% significance level (critical value = 1.96) and the 1% significance level (critical value = 2.58). But, what if the obtained value had been -2.08? We could do one of two things:

1. We could reject null at the 5% significance level and risk making a Type I Error, or
2. We could fail to reject null at the 1% significance level and risk a Type II Error.

A Type II Error happens when the null hypothesis is not rejected (accepted) when it is probably false. This is saying there is not a significant difference when there probably is.

**Note:** A Type II Error is probably more common among researchers. Significant results are usually those published, so if a Type I error is made, it is published for interested parties to see! Type II Errors are less likely to be found in the literature because results that are not significant are usually not published. So, researchers are willing to make an error by not claiming a result rather than claiming a false result—hence Type II errors are more common!

There are times when it would be more serious to make one error as opposed to the other. The following illustrate two extreme examples:

1. **Type I:** Saying there is a significant difference when there is probably not. A new drug to treat migraine headaches was compared to a standard drug and tested. The new drug is accompanied by a couple of uncomfortable side effects. But, tests show that the new drug is more effective. Why would a Type I error be unacceptable? Because a new drug, which is not more effective, could be on the market with nasty side effects.

2. **Type II:** Saying there is not a difference when there probably is. We have the same scenario as above, but the new drug does not have side effects. Tests show that the new drug is not more effective. Why would a Type II error be unacceptable? The new drug would probably not hit the market and it could have helped people. And, the drug has no side effects.

**To be very conservative and avoid making a Type I Error, a researcher can also set the significance level at .001. With this significance level, there is only 1 in 1000 chance of making the wrong conclusion. This is the most rigorous of the three significance levels.**
Inferential Statistics, Confidence Intervals:

In the previous example we used our z statistic to test a hypothesis about a population parameter based on a sample statistic. We used a point estimate, which is a sample mean, to draw conclusions about the population parameter.

Often, the primary purpose of data collection is not to test a hypothesis, but to obtain an estimate of a population parameter, like a mean. A point estimate used to test a hypothesis will give you a good idea of what the population mean is not (in other words, not likely to be 30), but it does not tell you what it is. That is why we have interval estimates, otherwise known as confidence intervals.

A confidence interval is an estimated range of values with a given probability of including the true population value. These are limits within which we would expect repeated sample means to fall. The usual probability level is 95%, but we can also use 99%. At 95%, we are saying that if we took repeated samples at random, constructed confidence intervals around the sample means, the true population mean would be captured in 95 out of 100 of the intervals. In other words, 95 times out of 100 we would probably be correct in stating that the population mean is within the intervals.

When N>30, we use the z statistic to calculate our confidence interval. Using the same data as in the example above (when we used the z test) the procedure for calculating a confidence interval is: sample mean +/- z (value at .05 or .01) times the standard error of the mean. At the 95% probability level:

\[ 11.3 \pm 1.96 \times (3.78) \text{ Note: } 1.96 \text{ is z at } .05 \]
\[ 11.3 \pm 7.41 \]

Confidence interval = 3.89—18.71 (the limits are the upper and lower numbers)

The confidence interval can be interpreted by saying there is a 95% chance that this is one interval that would include the true population mean. So seniors in the academic program would probably not read less than about 4 books or more than about 19 books. Notice that 30 is not within the confidence interval.

At the 99% probability level:

\[ 11.3 \pm 2.58(3.78) \text{ Note: } 2.58 \text{ is z at } .01 \]
\[ 11.3 \pm 9.75 \]

Confidence interval = 1.55—21.05

There is a 99% probability that this is one interval that would include the true population mean.
You can see that the confidence interval is influenced by the probability level we select as well as the standard error of the mean. As you make your probability level more conservative (99%), the interval becomes wider giving us “more room to be correct.” As your standard error increases, the confidence interval widens, and is not as precise.

To compare a point estimate with an interval estimate:

1. A point estimate is more likely to be reported to the public (average income in a township) because people tend to understand averages.

2. An interval estimate is useful in decision-making and is often used with the test of hypothesis.
Week 5 – t TESTS

INFERENTIAL STATISTICS CONTINUED

In this session, we are going to run the gamut of “t tests,” which are also tests of hypotheses—single sample, independent t, and dependent t.

More Inferential Statistics, Single Sample t Test:

In the example with the single sample z test, which we looked at last week, we had a fairly large sample—N over 30. We used the z test and the two standard critical values, either 1.96 (5%) or 2.58 (1%).

When N is less than or equal to 30, we cannot always assume that our distribution of sample means is so classically normal, so we use another mathematical distribution for testing hypotheses called the t distribution. This is a sampling distribution based on the probability of how sample means distribute with small samples. The good news is that you calculate the single sample t test with the same formula as the z test. The difference is that the critical value of t varies as a function of df (degree of freedom), which is N -1.

So, if N is greater than 30, you use the z test and the two standard critical values. If N is less than or equal to 30, you use the t test and df to find a critical value.

Example—

Let’s say it is budget time and you are a librarian wanting to purchase materials, such as videos, and you want to see if you can ask for an increase in your budget. The following represents the steps/data for this study:

1. Our librarian knows that the population price of videos last year was $20.00. The research hypothesis is that the prices have gone up. The null hypothesis says they have stayed the same—no change.

2. Set significance level at .05 (5%).

3. Random sample of N=16 video titles

4. Sample Mean=$24.00

5. Standard Deviation (s--using N -1 in denominator)=$3.00
6. Calculate standard error of the mean: $s/\sqrt{N}$

$$3.00/\sqrt{16} = .75$$

7. Calculate $t$: \[\text{sample mean} - \text{population mean} = 24 - 20 = 5.33\]

\[
\text{standard error of the mean} \times .75
\]

**5.33 is the obtained value of $t$.**

8. Calculate df: $N - 1$ \[16 - 1 = 15\]

9. Go to Yonker p. 40 to find the critical value of $t$.

You can see there is a different probability distribution for each df. For level of significance for a two-tailed test, find where .05 meets 15 df. For 15 df, the **critical value is 2.131**.

10. Our obtained value of 5.33 is beyond the critical value of 2.131 on the positive side of the curve (see Yonker, p. 21), so we can reject the null hypothesis. It is unlikely we would get a sample mean as large as 24 dollars if our population mean was really 20 dollars. So, it is likely that prices have gone up!

To review the steps necessary for testing a hypothesis about a population mean, using the single sample $z$ or $t$ test, see Yonker, p. 22.

You can also set a **confidence interval** around the sample mean using $t$, at 95% or 99%. At 95%, the procedure is:

\[
\text{sample mean} +/- t \times \text{at .05 times the standard error of the mean}
\]

\[24.00 +/- 2.131(.75)\]

\[24 +/- 1.6\]

**Confidence interval = 22.4—25.6**

There is a 95% probability that this is one interval that includes the true population mean. Notice that 20 dollars is not in the interval! Prices of videos probably fall between about 22.5 dollars and 25.5 dollars.
Inferential Statistics, Independent t Test:

With a single sample t test we were trying to see if there was a change in the population value using a sample mean (point estimate). We were testing to see if there is a significant difference between the sample value and the population value. A more common use of t tests is when you are testing for significant differences between two sample means. There are two types of t tests which do that. See PowerPoint slides below:

[Dependent_t.ppt]

Two kinds of t-tests

- All t-tests compare two means or two proportions
- Related sample, paired sample, correlated sample--these are all names for dependent means designs for t-tests.
  - Each subject in one sample has a corresponding subject in the second sample...perhaps the same person

[Independent_t.ppt]

Independent sample t-test

- Second variety of t-test compares two independent means
  - two groups measured using the same instrument, for example

Let's start with an example that involves two independent random samples (Independent Samples t test). Independent means that no individual or object in one sample can be in the other as well.

Let's say an elementary school librarian wants to compare the average cost of workbook prices for different subjects to make budgeting decisions—specifically, comparing social studies and science workbooks. The librarian takes two independent random samples of workbook prices, one sample from science and the other from social studies. The average price or mean is calculated for each sample.

Science—sample mean 1   Social studies—sample mean 2

Using these two sample means, the librarian wants to test to see if the two samples come from two populations that have fairly equal means, or if the
samples come from two populations with different population means. **Again we are inferring from samples to populations.** The **null hypothesis** states that there is no difference between the population mean prices of workbooks from the sciences and social studies. Another way of stating null is that the difference between population means is equal to zero. **A research hypothesis** is that there is a difference in the book prices, and it may even say that science workbooks are more expensive than social studies workbooks. So, the **t test (called independent t)** will test to see if there is a significant difference in the average prices of the two workbooks.

Remember with our single sample tests we were testing hypotheses using a sampling distribution of many sample means (called the sampling distribution for the mean). For two samples, we also have a sampling distribution, but this sampling distribution is for all possible differences between sample means—again this is a **theoretical probability distribution** (you may want to review the notes from last week concerning sampling distributions).

How was this sampling distribution developed?

1. By taking many pairs of random samples (with equal or unequal N's) from two populations and calculating sample means for each sample (average prices of science and social studies workbooks).

2. And, determining the difference between the sample means. The sampling distribution for the differences between means would form a frequency distribution graph of the differences.

Pair 1 Sample mean 1 – Sample Mean 2 = difference
Pair 2 Sample mean 1 – Sample Mean 2 = difference
……
……
Pair 1000 Sample mean 1 – Sample Mean 2 = difference

What does this sampling distribution look like?
Since we are testing a null hypothesis which states that the population means are equal—if the population means are really equal, we would expect to obtain a zero difference between sample means most of the time. Sometimes we would get a positive difference, sometimes we would get a negative difference, and very rarely we would get large differences between the sample means. So the **sampling distribution for differences between means** is a normal distribution with zero at the center as the most frequent value!
So, when the t test is calculated and a large difference is found between the two sample means (taken from the two populations we are studying), we will probably reject the null hypothesis that the difference between the population means is really equal to zero.

Remember also, that with our single sample tests, the sampling distribution of the mean was divided into standard error units—called the standard error of the mean, which is \( \frac{s}{\sqrt{N}} \). We also have the concept of the standard error for this sampling distribution for differences between means and it is called the **standard error for differences between means**. We will run through the example we have already started about the workbook prices and illustrate the concepts of the sampling distribution and the standard error for differences between means.

<table>
<thead>
<tr>
<th>Sciences</th>
<th>Social Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Mean 1  = $17.90</td>
<td>Sample Mean 2  = $16.40</td>
</tr>
<tr>
<td>Standard Deviation (s) 1 = $2.30</td>
<td>Standard Deviation 2 = $2.10</td>
</tr>
<tr>
<td>N1 = 20</td>
<td>N2 = 30</td>
</tr>
</tbody>
</table>

We must use the t test for this problem because the sample sizes are 30 or under. You can only use the z test if both samples have an N greater than 30.

We will set the significance level for t at .05 (5%).

The **standard error for the differences between means** measures the combined sampling error for the two samples. The formula follows:

\[
\sqrt{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)}
\]

\[
\sqrt{\left(\frac{2.30^2}{20} + \frac{2.10^2}{30}\right)} = \sqrt{(.265 + .147)} = .641
\]

So the **standard error for differences between means** = .641

In Yonker, p. 23, you will see that the standard error for differences between means functions like the standard deviation of the sampling distribution for differences between means (be sure to review the notes from last session which originally defined a standard error). You can see that the curve, once again, is divided into standard error units—otherwise known as the values of t, -2, -1, 0, +1, +2, etc. Using the standard error you can get a range of possible differences between means for each value of t.
So, at t = 2, the difference between means is equal to 1.28, or two standard error units above 0.

The t test (Independent t) for the above example (workbook prices) is calculated with the following formula:

\[
\frac{\text{Sample Mean 1} - \text{Sample Mean 2}}{\text{Standard Error for Differences Between Means}}
\]

\[\frac{17.90 - 16.40}{1.5} = 2.34\]

\[.641 \div .641\]

On the curve, in Yonker, p. 23, you could place the difference between means of 1.5 at t = 2.34, which is where this difference is located on the sampling distribution. The difference is quite a distance away from zero, but is it far enough away from zero to reject the null hypothesis? We will know the answer when we compare the obtained value of t = 2.34 to a critical value of t at .05.

Since we are using t, the next step is to find degree of freedom. The formula for degree of freedom for two samples is:

\[\text{df} = N_1 + N_2 - 2\] (1 df is lost for each sample).

\[20 + 30 - 2 = 48 \text{ df}\]

The next step is to go to the t table in Yonker, p. 40. Use the two-tailed test at .05 with 40 df. In this case, since there is no entry in the table for 48 df, we must use 40. Using 40 is the more conservative df because it provides a larger critical value (a bit larger than the critical value at 60 df). With a larger critical value, it is more difficult to reject the null hypothesis and then a Type I error is not as likely. Even if the actual df had been calculated at 59, we would still use 40 df in the table.

So at 40 df, the critical value = 2.021. And, the obtained value of t, which is 2.34, goes beyond the critical value; therefore, we (and the librarian) can reject the null hypothesis. There is a significant difference between the prices of the workbooks, and by the values of the sample means ($17.90 and $16.40), it seems that the science workbooks are more expensive. At the 5% significance level, there is a
5% chance we could be wrong in rejecting the null hypothesis, or a 5% chance of making a Type I Error!

It seems our librarian would want to budget more funds to the sciences!

See Yonker, p. 22, and go through the steps below which change the outline to accommodate the steps to working with Independent t.

1. State research and null hypothesis about population parameters.
2. Same
3. Take two independent random samples from populations.
4. Calculate 2 sample means, 2 sample standard deviations (unbiased estimates using N -1 in the denominator of the formula), and the standard error for the difference between means (see formula above).
5. Calculate t (see formula above) if one (or both) of the sample sizes is less than or = 30. If both sample sizes are > 30, use z. The formula for z is the same as t, except we use the two standard critical values for determining significance: 
   \[ z=1.96 \] (.05) or \[ z=2.58 \] (.01).
6. Same, except for Independent t, df = N1 + N2 – 2
7. Same

**One- and Two-Tailed Tests of Significance:**

We have been using a two-tailed test of significance, and really, this is most common among researchers. A two-tailed test divides the critical region of rejection into two tails. At the 5% significance level, each tail of the curve has 2.5% in the region of rejection.

A one-tailed test is only used when we are very sure of the direction of our results— that one mean will be greater or smaller than another mean— either a positive or negative direction. In a one-tailed test, the region of rejection is all in one tail, either the positive tail of the curve, or the negative tail of the curve. **A one-tailed test increases the size of the region of rejection on one side and you are more likely to get significant differences and reject the null hypothesis.** If you look at the t table for 15 df, .05, you will see that the critical value for a one-tailed test (1.753) is actually a smaller value than a two-tailed test (2.131). The smaller value makes it easier to get significant results; therefore, you do not want to use a one-tailed test inappropriately because you are more likely to make a Type I Error. Researchers do not want to make a Type I Error, so they are more likely to use a two-tailed test!

When could you use a one-tailed test?

Prices represent a good example. A school system wants to convince the board that prices have gone up enough from last year to warrant an increase in the budget. Since inflation always goes up, the change would be in a positive direction. A sample of prices this year (Mean 1) minus a sample of prices last year (Mean 2) would produce a positive result and a
A one-tailed test could be used. However, the researcher must know the market for prices because, for example, computer prices usually go down!

Two-tailed tests are used when the results could produce a positive or a negative difference. Even if your research hypothesis indicates a certain direction, the results can sometimes go the other way!

Example:

You are a school librarian who wants to test the effect of a library-use training program on a sample of junior high school students. The intent is to prepare these students for high school and ultimately college, so the impact of the program is important. You have two samples of students.

Sample 1: Students get the training program—Experimental group

Sample 2: Students do not get the training program—Control group

You want to compare these samples on the amount of library use as a measure of the dependent variable. The effect of the training program could be positive or negative when compared to the control group. The students in the training program could use the library less because they are able to find what they want more quickly or they could use the library more because now they are more interested in using the library. So, a two-tailed test should be used!

**IMPORTANT:** Unless you are very sure the direction of your results can only go one way, you should use a two-tailed test.

**Inferential Statistics, Dependent t:**

We will finish up this session by looking at a t test for related samples, paired samples, correlated samples, which is called **Dependent t.** See PowerPoint slide to review:

[Dependent_t.ppt]

- All t-tests compare two means or two proportions
- Related sample, paired sample, correlated sample—these are all names for **dependent means** designs for t-tests.
  - Each subject in one sample has a corresponding subject in the second sample...perhaps the same person

Dependent t is used when you have the **same subjects** tested on the **same variable/measure** before and after an experimental procedure, program, or treatment. Or, it is used when you have matched pairs (such as twins) on some variable and one
member of the pair gets the experimental procedure, and the other member of the pair
does not get the procedure.

The calculation of Dependent t is a more sensitive test for dealing with the true effect of
an experimental procedure. Since the same or matched subjects are less likely to differ,
a more sensitive test is needed to detect any real differences that could occur because
of an experiment.

**We will use an example of a before/after test on the same subjects:**

We have ten (N = 10) junior high school students taken at random, and we
want to test the impact of a library-use training program. We asked them
to track the number of separate times they went into the library and used it
for a 6-week period. Then, they received the library-use training program
for one marking period. After the training program, they were asked to
track their library use for another 6 weeks.

A possible **research hypothesis** is that the library-use training program
will increase library use. The **null hypothesis** states that the training
program will not influence library use, or the difference between means
before and after the training program is zero. The significance level is set
at .05.

See Yonker, p. 24, for the formula for Dependent t and the distribution of
library use before and after the training program. The formula represents
the mean of the differences divided by the standard error of the
mean differences. You can see that the “Difference” column (D)
represents the before measure minus the after measure, and, the $D^2$
column represents the Difference squared. The “sum of” each column is
included in the calculations.

The procedures below represent each step of the t formula and the
conclusions of the test:

1. The numerator in the t formula (the mean of the differences) is found by
taking—

   The sum of $D/N = -37/10 = -3.70$

2. The denominator in the t formula (the standard error of the mean
difference) is found by taking—

   $$\sqrt{\frac{SS_D}{N(N-1)}}$$
SSD is the sum of squares of the difference scores—kind of like a variance.

Where SSD = the sum of $D^2 - (\text{the sum of } D)^2$

$$SSD = \frac{511 - (-37)^2}{10} = 511 - 1369 = 511 - 136.9 = 374.10$$

3. So, $\sqrt{\frac{SSD}{N(N-1)}} = \sqrt{\frac{374.10}{10(10-1)}}$

$$\sqrt{\frac{374.10}{90}} = \sqrt{4.16} = 2.04$$

2.04 = the standard error of the mean difference

4. $t = \frac{\text{The mean of the differences (found through step 1)}}{\text{The standard error of the mean differences}}$

$$-3.70/2.04 = -1.81 \text{ (the obtained value of } t)$$

5. Since we are using $t$, the critical value must be obtained from the $t$ table using degree of freedom. For Dependent $t$, $df = N - 1$

$$df = 10 - 1 = 9$$

6. From the $t$ table in Yonker, p. 40, the critical value of $t$ at .05, two-tailed test = 2.262 (on the positive and negative sides of the curve).

7. Since our obtained value of $-1.81$ does not go beyond the critical value of 2.262 (negative side), we do not reject the null hypothesis; therefore, it seems that the training program did not have an influence on library use!

The good news is that I will not ask you to calculate Dependent $t$ by hand in any of your homework assignments (or, on a test), so this was your first, and only look at the calculations.
One-Way Analysis of Variance:

Today we are going to talk about a type of statistic called Analysis of Variance—AKA ANOVA, AKA the “F test”. Analysis of Variance is still part of inferential statistics.

Let’s contrast Analysis of Variance with the independent “t test” which we have already covered. The following are characteristics of independent t:

1. Two independent samples are taken at random.
2. Can be used with smaller sample sizes, N less than or equal to 30.
3. Looks for differences between means.
4. Assumes a fairly normal distribution.
5. Used with interval or ratio level of measurement.

The t test is very useful for two groups with small sample sizes, but what if we want to get closer to the real world and test more than 2 groups on some variable and look for differences between means—that’s where ANOVA comes in. ANOVA has similar characteristics to t, as listed above, except ANOVA can be used with two or more groups. An additional characteristic of ANOVA is the assumption for Homogeneity of Variance. Basically, this means that the variances of the two or more groups are assumed to be fairly equivalent in order to use ANOVA.

Now the question might be, couldn’t we use the “t test” for all possible comparisons/combinations between more than two groups and test for differences between means? We could, but imagine how tedious that would be, and imagine that with each t test we would have to worry about making a type I error. ANOVA solves these problems by showing an overall indication if there is a significant difference among the 2 or more groups in regard to means (it functions the same as t for 2 groups).

If we look back again to the “t test,” the ratio for finding “t” was to see if the effect between means (the difference between means in the numerator) was larger than the sampling error (known as the standard error for differences between means in the denominator):

\[ t = \frac{(\text{sample mean1} - \text{sample mean2})}{(\text{standard error for diff. between means})} \]

We have a similar ratio with ANOVA but the terminology is different. ANOVA compares the between group variances in the numerator (which is like the effect of whatever research variable we are trying to study) with the within groups variance in the
**denominator** (kind of like the sampling error within the groups). When you see the calculations for ANOVA, this ratio of between groups variance to within groups variance will give you the result of “F.” See PowerPoint slides as supplement to the characteristics of ANOVA.

F-tests and ANOVA

- Analysis of variance (ANOVA) allows comparison of several means in one test
- Strong assumptions about the nature of the data are made--the groups being compared are assumed to be random samples from normal populations with the same variance

One-Way ANOVA

- Compares the means of two or more unrelated samples.
- An estimate of the variance between groups is compared to an estimate of the variance within groups -- groups are made up of one variance.
- The null hypothesis in n-way analysis of variance is that the means of the n groups are equal (do not differ) in the population

As an example, a library experiment will be used to illustrate how to calculate and interpret ANOVA. We discussed experimental research as part of the nine research strategies from the first session.

Noise in the library is a factor that has always been an issue—how does noise affect the working habits of the patrons, especially when study areas are near the service desk (such as, in a small school library)?

For the experiment, assume a self-contained laboratory-like classroom under three conditions. A pool of 15 students is randomly assigned to one of the three conditions, so there are an equal number of students in each group. **Random assignment means that each student has an equally likely chance of being in each group.** Five students in each group will make the calculations of ANOVA less tedious. But, although ANOVA can be used with small numbers in each group, usually you would want to have at least 10 in each group. The three conditions for the experiment (the independent variable) are as follows:
Group 1, Low Noise: room quiet, librarian sits at desk and does not speak

Group 2, Medium Noise: librarian visited by students (different students than the five in each group) one at a time and conversation is kept to a low level.

Group 3, High Noise: students and other librarians visit the librarian (more than one at a time) and talk in normal speaking voices.

The five students in each group are all sitting at desks working on a math test. The dependent variable is the number of correct answers on the set of math problems—all students have the same number of problems and have the same amount of time to complete the test.

The research hypothesis is the more noise, the fewer correct answers on the math test.

The null hypothesis is that noise does not affect problem solving.

As with other tests of significance, we have to set a significance level. It is set at .05. We will find our critical value of “F” in a table, just like we did for “t”.

In Yonker, p. 25, you will see the distribution of math test scores for the three groups and an ANOVA summary table which was obtained via SPSS. You can see, for each group distribution, the sum of X and the sum of $X^2$ were calculated. A mean for each group was also calculated. The ANOVA summary table shows steps, one through four, which will be covered to find the result of “F.” We will calculate each value in the summary table. The good news is that you will see how to manually calculate ANOVA in this session, but SPSS will do it for you in your homework. And, I will not ask you to manually calculate ANOVA in a test!

STEP 1: Calculate degrees of freedom

$dfbg$ (degree of freedom between groups) = number of groups – 1
3 – 1 = 2

$dfwg$ (degree of freedom within groups) = Ntotal – number of groups
15 – 3 = 12

$dftot$ (degree of freedom total) = Ntotal – 1
15 –1 = 14

STEP 2: Calculate sum of squares-- SSbg, SSTot, SSwg
**n = size of a group

SSbg (sum of squares between groups) =

\[
\frac{(\text{sum of } X_1)^2}{n_1} + \frac{(\text{sum of } X_2)^2}{n_2} + \frac{(\text{sum of } X_3)^2}{n_3} - \frac{(\text{sum of } X_1 + \text{sum of } X_2 + \text{sum of } X_3)^2}{N_{\text{total}}}
\]

\[
\frac{(78)^2}{5} + \frac{(65)^2}{5} + \frac{(37)^2}{5} - (78 + 65 + 37)^2/15 =
\]

\[
1216.8 + 845 + 273.8 - (32400)/15 =
\]

\[
2335.6 - 2160 = 175.6
\]

SStot (sum of squares total) =

Sum of \(X^2\)1 + sum of \(X^2\)2 + sum of \(X^2\)3 – (sum of X1 + sum of X2 + sum of X3)\(^2\)/N\(_{\text{total}}\)

**Notice** the last part of this formula (after minus sign) was already calculated above

\[
1232 + 865 + 291 - (2160) \text{ was calculated above}
\]

\[
2388 - 2160 = 228
\]

SSwg (sum of squares within groups) = SS\(_{\text{tot}}\) – SSbg

\[
228 - 175.6 = 52.4
\]

STEP 3: Calculate mean square-- MSbg, MSwg

MSbg (mean square between groups) = SSbg/dfbg

\[
175.6/2 = 87.8
\]

MSwg (mean square within groups) = SS\(_{\text{wg}}\)/dfwg

\[
52.4/12 = 4.37
\]

STEP 4: Calculate F

\[
F = MSbg/MSwg
\]

\[
87.8/4.37 = 20.1
\]

As mentioned above, “F” is obtained through the ratio of the between group variances in the numerator (which is like the effect of “noise”) to the within groups variance in the denominator (the sampling error within the groups).
In order to determine the **significance of F**, go to Yonker, p. 26 and you will find Table D—Critical Values of F. Look at the top row of the table where it says “Degrees of freedom for numerator.” That is the *degree of freedom between groups* (dfbg), which is 2. Look at the left side of the table where it says “Degrees of freedom for denominator.” That is the *degree of freedom within groups* (dfwg), which is 12. Where 2 meets 12 degrees of freedom, look at the top value, which is the critical value for .05 (the bottom number is the critical value at .01). The critical value of F is equal to 3.88. Our obtained value of 20.1 is beyond the critical value of 3.88; therefore, we reject the null hypothesis. You will also note that the SPSS table on p. 25 in Yonker, shows a highly significant F as well (.000 in Sig.). **Noise level does seem to have an effect on students’ performance**—there is a significant difference among the groups in math test performance.

You can see by the three means on p. 25 in Yonker, that the greater the noise level, the lower the math test scores, which is especially so in the high noise group (group 3). ANOVA (the F test) will tell you that there is an overall significant difference, but it does not tell you where that significant difference lies. You can find exactly which means are significant from other means by a **post hoc test**. You only use a post hoc test if F is significant. A common post hoc test is called the **Tukey Test**. You will explore this post hoc test as you complete your homework assignment (number five) for next week.
Week 7 - MEASURES OF ASSOCIATION

INFERENTIAL STATISTICS: MEASURES OF ASSOCIATION

Measures of Association:

We will start this session by comparing Measures of Association to a statistic we have already covered—Independent t. In session five, using t, we had two independent random samples tested on the same measure, book prices. Measures of Association involve taking two different measures (otherwise known as variables) for every individual in one random sample.

Measures of Association are used to determine how strong the relationship is between two measures/variables, and how we can predict such a relationship. So Measures of Association have two functions:

1. Strength
2. Prediction

As an example, we could take a random sample of graduate students from IST and collect two measures from each student: 1. GRE scores, and 2. GPA (grade point average)

We can use our techniques called measures of association to determine:

1. How strong the relationship is between these two measures. Do the measures of GRE and GPA "go together" in a specified way?
2. And, if it is a strong relationship, how well can we predict a value of one variable when the value of the other variable is known?

There are two types of analyses used to determine strength and prediction:

1. Correlation: Describes the degree (strength) to which the two variables are related and indicates to a certain extent how (meaning direction) they are related.
2. Regression: Used to predict the values of one variable when values of the other variable are known.

These two techniques of correlation and regression are ultimately linked--the ability to predict one variable when another variable is known depends on the degree to which the two variables are correlated. For instance, if GPA and GRE scores are highly correlated, we could use one to predict the other. So, the higher the correlation, the better the prediction; therefore, correlation is the first step.
**Correlation:**

There are different types of statistical measures that result in an analysis of correlation and they are known as **correlation coefficients**. A very common correlation coefficient and the one we will focus on in this session is called **Pearson's Product Moment Correlation**, AKA **Pearson's r**.

**Pearson's r** can only be used with interval or ratio level of measurement—this is an assumption for using "r."

The result of r is a decimal value ranging from -1 to +1. As stated previously, correlation (in this case Pearson's r) measures the **strength** and **direction** of a relationship:

1. **Strength** is determined by how close your decimal value of r reaches +1 or -1. Both would indicate a strong relationship. As your decimal value approaches 0, this means very little or no relationship.
2. **Direction** is determined by the sign of the correlation coefficient, either positive or negative.

In the example below, you can see two variables, GRE (variable X) and GPA (variable Y) taken from five students selected at random. The Pearson's r calculated from these two variables results in a perfect positive correlation (r = 1). In a positive correlation, as one variable increases, the other variable increases, so a student with a high GRE score has a high GPA. You can also see that a student with a low GRE score has a low GPA.

<table>
<thead>
<tr>
<th>Student</th>
<th>GRE (X)</th>
<th>GPA (Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1500</td>
<td>4.0</td>
</tr>
<tr>
<td>2</td>
<td>1400</td>
<td>3.7</td>
</tr>
<tr>
<td>3</td>
<td>1250</td>
<td>3.5</td>
</tr>
<tr>
<td>4</td>
<td>1050</td>
<td>3.1</td>
</tr>
<tr>
<td>5</td>
<td>950</td>
<td>2.9</td>
</tr>
</tbody>
</table>

In this next example, the resulting "r" shows a perfect negative correlation (r = -1). In a negative correlation, as one variable increases, the other variable decreases, so a student with a high GRE score has a low GPA. And, a student with a low GRE score has a high GPA.
Both of these examples would form a straight line if plotted on a graph. These are linear relationships (a straight line).

**It is important to note** that a correlation coefficient, like Pearson’s r, does not give us any information about causation. In other words, high GRE's don't cause high GPA's and vice versa.

As an example: One summer in the city, a high correlation was found between crime rate and ice cream sales! Would any one of us dare to say that ice cream causes crime?! Probably not, but a third variable could be involved. Yes, you guessed it—heat!

**Correlation and Scatter grams:**

So how would we get an idea if two variables are related (other than calculating Pearson’s r, of course)? One of the clearest ways is to visually examine the relationship by a graphic technique known as a **Scatter gram**, AKA **Scatter plot**. A scatter gram is a two dimensional graph with an axis for each variable, X and Y. To construct the scatter gram, you place a point for each case (person, library, etc.) that corresponds to the value the case has on both variables. In Yonker, p. 28, you see a scatter gram representing ten libraries on two variables, per capita expenditure (variable X) and per capita circulation (variable Y). You should count ten points on the scatter gram, one for each library. The lowest point, to the left, represents library E with a per capita expenditure of 3.00 and a per capita circulation of 2.1. We will come back to this graph and this data again. We will use this data to calculate Pearson’s r later.

Scatter grams look different for different types of relationships or correlations. In Yonker, p. 27, you see five scatter grams representing different types of relationships with the corresponding values of Pearson’s r included. Pearson’s r is the summary statistic for the data in the scatter gram. The Pearson’s r and the scatter gram should be interpreted together.

1. Scatter gram a. represents a perfect linear relationship (a straight line)—as one variable increases the other also increases. The r that summarizes this data is 1 (a perfect positive correlation).
2. Scatter gram b. also represents a perfect linear relationship—as one variable increases the other variable decreases. The r that summarizes this data is -1 (a perfect negative correlation).

3. Scatter grams c. and d. both represent moderate correlations as shown by the values of Pearson’s r, .65 and -.62. You do not see a perfect straight line in either graph, but the points are fairly well grouped together in a linear trend. In c., as one variable tends to increase, the other increases (a positive correlation). In d., as one variable tends to increase, the other decreases (a negative correlation).

4. In scatter gram e., there is no relationship and Pearson’s r = 0. The points are very scattered, no pattern is observed, and no relationship exists.

Go back to Yonker, p. 28, and think about the direction and strength illustrated in this scatter gram. Would you say this is a positive or negative relationship? Do you think this relationship is strong? We will find out with more certainty later when we calculate Pearson’s r.

**Important:** The scatter gram is extremely useful in showing us the nature of the relationships; it will give a sense of the direction of the relationship (positive or negative), and whether the relationship is linear (approaches a straight line). **Linearity** is important to detect because it is an assumption which must exist in order that Pearson’s r can be calculated. Earlier, I stated another assumption for calculating Pearson’s r—you must have interval or ratio level of measurement.

**Calculating Pearson’s r:**

In Yonker, p. 29, we revisit the ten libraries (as labeled by a letter) that were used to illustrate the scatter gram. Included on this page is the distribution of the ten libraries on the two variables, per capita expenditure (variable X) and per capita circulation (variable Y). The formula and computation of Pearson’s r is also included. The formula looks complicated, but does not require any calculations that you are not familiar with. The following steps take you through the calculations. **First the columns:**

1. You can see that the numbers in the X column are added to get the sum of X, and the numbers in the Y column are added to get the sum of Y. **Check your symbol sheet in Yonker, p. 4 for the sum of symbol.**

   **Sum of X = 50.30**

   **Sum of Y = 52.7**

2. Each value in the X column is squared to get $X^2$. So, at the top, for library A, $4.60^2 = 21.16$. You sum all of the values in the $X^2$ column.

   **Sum of $X^2 = 269.61$**
3. Each value in the Y column is squared to get $Y^2$. So, at the top, for library A, $4.5^2 = 20.25$. You sum all of the values in the $Y^2$ column.

**Sum of $Y^2 = 326.19$**

4. Each value of X is multiplied by the corresponding value of Y to get XY. So, at the top for library A, $(4.6)(4.5) = 20.70$. You sum all of the values in the XY column.

**Sum of XY = 292.66**

Now the formula:

1. N is equal to the number of libraries (N = 10). Just because two variables are taken from the 10 libraries does not mean that N doubles—this is a common mistake that people make.

2. Match each number from the column totals to the appropriate quantity in the formula and “plug in” the numbers.

3. Remember the order of operations (“My Dear Aunt Sally”). In the numerator:
   a. $10(292.66) = 2926.6$
   b. $(50.30)(52.7) = 2650.81$

Then $2926.6 - 2650.81 = 275.79$

4. In the denominator:
   a. In the first brackets for the X variable—
      $10(269.61) = 2696.10$
      $(50.30)^2 = 2530.09$
      Then (not shown on page 29 in Yonker)
      $2696.10 - 2530.09 = 166.01$
   b. In the second brackets for the Y variable—
      $10(326.19) = 3261.9$
      $(52.7)^2 = 2777.29$
      Then (not shown on page 29 in Yonker)
3261.9 – 2777.29 = 484.61

c. (166.01)(484.61) = 80450.106

**Note:** In your homework assignment for this week—when you get to this point (immediately above) in your by-hand calculation of Pearson's r, take the square root of each number in parentheses first. After you take the square root of each number, then multiply them. You do not have to take the square root again. Just finish the calculation of Pearson's r. The reason for doing this is because in your homework, the two numbers in parentheses, when multiplied (without taking the square root first), result in a very large number which your calculator may not handle. Taking the square root first, reduces the size of each number in the parentheses.

5. 275.79/square root of 80450.106 = .97

Our resulting r = .97. What does this result say about our data? There tends to be a high positive correlation between a library's per capita expenditure and per capita circulation. As one variable tends to increase linearly, so does the other.

The following list gives you some guidance on the interpretation of different values of r. These numbers work for both positive and negative values of r.

If r is greater than .90: very strong correlation

.7--.9: strong

.4--.7: moderate to substantial correlation

.2--.4: low to moderate

.2 and below: slight to almost negligible

**Coefficient of Determination:**

Another useful measure, which can be calculated from Pearson’s r, is known as the **Coefficient of Determination, or r²**. The result of r² tells us the amount of variance the two variables share (explained variance). In the previous example of the ten libraries, r = .97, so .97² = .94.

The value of r² is changed to a percentage, so in this example, that is 94%. The coefficient of determination is interpreted as the amount of the variance the two variables share or have in common—called common or explained variance. Therefore, 94% of the variance in per capita circulation (Y) is explained, accounted for, or shared with per capita expenditure (X).
If 94% of the variance is explained, what number represents the unexplained variance between the two variables. That’s right—6%!

**The Significance of r:**

As with other statistics, like z and t, we can test for the significance of Pearson’s r. The null hypothesis is that the correlation between the two variables (per capita expenditure and per capita circulation) is equal to zero--there is no relationship. To test for significance we have to set a significance level and then consult a table of critical values for r. This table is found on p. 41 in Yonker.

The significance level is set at .05 (and remember, this is set before you begin the study). The degree of freedom (df) for checking the significance of Pearson’s r is equal to **N – 2**. We have one random sample with two measures, so a degree of freedom is taken from each measure.

\[ df = 10 - 2 = 8 \]

Look at the r table, at .05 (two-tailed test). We find that a minimum correlation of .632 would be needed to reject the null hypothesis. **Since the obtained r of .97 is beyond the critical value of .632, we reject the null hypothesis.** We state that there is a significant positive correlation between per capita expenditure and per capita circulation.

**Note:** A significant r does not necessarily mean you have a very strong correlation (as you can see from the critical value of .632 above). Essentially, significance is evidence of the stability of r. A significant r indicates that you are likely to get the same results again, in other words, the results are not a fluke or not due to chance. So, when interpreting r you should look at the **significance** and the **strength** of the relationship. Is the relationship significant, and if so, how strong is it? In the example of the 10 libraries (above), \( r = .97 \); therefore, we have a **significant** r, but it is also **very strong** and **positive**.

**Regression Analysis:**

Once we have found the variables that are strongly related, then we are likely to turn to regression analysis. Regression analysis allows us to predict the likely values of one variable from knowledge of another variable.

As mentioned previously, in order to be able to predict one variable from another variable, the two variables must be **fairly strongly correlated**, (close to a straight line). The mathematical expression used for prediction is called the regression equation. **The regression equation is an equation for a straight line.** The following is the regression equation:

\[ Y = a + bX \]
In this equation, you are predicting Y values from known values of X (called the regression of Y on X). Y is the dependent variable, and X is the independent variable. Since the regression equation is an equation for a straight line (linear), the stronger our correlation, or the closer our variables fall on a straight line (a perfect linear relationship), the better the prediction.

The "b" in the regression equation is called the slope or the regression coefficient—it tells you how much to multiply X in order to get Y. A formal definition of the slope is the rate of change in Y with a change in X. To illustrate the slope, go to Yonker, p. 30. These ten pairs of observations represent a perfect positive correlation or a perfect positive linear relationship—as X increases linearly, so does Y. All of these points would fall on a straight line in a scatter gram. How would you get the values of Y from the values of X? If you look closely at all of the values of X, you can see that if you multiply X by 2, you will obtain the corresponding value of Y.

\[ Y = bX \]

Y = 2X Example from the top: 10 = 2(5)

The slope, AKA “b” is 2. You must multiply X by 2 to get Y. Or, as X changes by 1 unit, Y would change by two.

Look at the graph (Yonker, p.30)--the line that runs through the graph is called the regression line and it is always a straight line. It is the line of prediction. If you have a perfect correlation, your actual values of X and Y will fall on the regression line and you will have perfect prediction.

FYI: Try, if you can, to remember high school geometry (I know this is difficult for me!). The graph shows that the slope is equal to Rise over Run.

**IF YOU DO NOT REMEMBER THIS, THAT IS OK!**

In this example, it was fairly easy to find the slope by observation because we have perfectly correlated data. In the real world, we usually do not have perfectly correlated data, so we have to use a formula (which I will show you later) to obtain the “best estimate” of the slope.

Let’s turn our attention to “a” in the regression equation. The value of “a” is called the Y-intercept. It is the point where the regression line intercepts the Y axis. Or, another way of defining “a” is the value of Y when X is equal to zero. The example, in Yonker, p. 30, shows that the regression line intercepts the Y axis at the zero point or origin (when \( X = 0, Y = 0 \)) of both X and Y. But, this is not always the case. The regression line can intercept the Y axis at any point.
Look at Yonker, p 31, for another example of ten pairs of observations that illustrate a perfect positive correlation. But, in this example how would we obtain $Y$ from $X$. The keen observer should notice that all values of $Y$ are obtained by the following:

$$Y = a + bX$$

Looking at this graphically (bottom of Yonker, p. 31), you can see that the regression line intercepts $Y$ at 2; therefore, 2 (or the $Y$ intercept) is our value of "a." If we add this constant of "a" equal 2 to our regression equation, then we can get our value of $Y$ by taking 2 (the slope) times $X$.

Why is the $Y$ intercept also defined as the value of $Y$ when $X$ equal zero?

$$Y = a + bX$$

$$Y = 2 + 2(0)$$

$$2 = 2 + 0$$

In both of the situations where we found “a” and “b” previously, we had perfect correlations and it was fairly easy to look at the distribution of $X$ and $Y$ and to find the intercept and the slope. And, in both cases, all of our points of $X$ and $Y$ fall right on the regression line and we have perfect prediction. When $X$ and $Y$ are not perfectly correlated (which is usually the case) and not exactly on a straight line, you have to estimate “a” and “b” with formulas. The formulas (which you will see soon!) provide us with the “best estimates” of the intercept and the slope from data that is not perfectly correlated and does not fall on a straight line. The ultimate goal of the regression equation is to provide the “best fitting” straight line through values of $X$ and $Y$ that, again, are not perfectly correlated.

**WHAT DOES THIS SAY ABOUT CORRELATION IN RELATION TO REGRESSION?**

The higher the correlation, the closer the points of $X$ and $Y$ fall to a straight line, and the better the prediction will be.

Now in both the relationships I just showed you (Yonker, p 30-31), $X$ and $Y$ were positively related. In a negative relationship, the regression equation stays the same, however the sign for "b" will be negative and the regression line will look as illustrated in Yonker, p 32. You can see that an increase in $X$ means a decrease in $Y$.

Let’s go back to our expenditure and circulation data from Yonker, p. 29, and I will finally show you the formulas for “a” and “b.” I am assuming you remember how to calculate means and standard deviations, so when the formulas call for these numbers, I will give the values with the idea that you know what they are.

$$b = (r) \text{ standard deviation of } Y = (.97) \frac{2.2}{1.29} = (.97)(1.71) = 1.66$$
standard deviation of X

**b (slope) = 1.66** So, as X increases by one value, Y increases by 1.66!

\[ a = \text{mean of } Y - (b)(\text{mean of } X) = \]

\[ 5.27 - (1.66)(5.03) = \]

\[ 5.27 - 8.35 = -3.08 \]

**a (Y intercept) = -3.08** So, this is the point where the regression line intercepts the Y axis!

**Using the Regression Equation:**

Now that we have “a” and “b,” let’s see what our regression equation looks like.

\[ Y = a + bX \]

\[ Y = -3.08 + 1.66(X) \]

We already know what our Y values (per capita circulation) are (from Yonker, p. 29), but let’s see how closely we can predict Y with our known values of X using the regression equation. The predicted Y has a different symbol: a \(^\wedge\) is placed over the Y to get a Ŷ. The symbol over the Y is called a *circumflex*.

We will take two values of X--the highest and the lowest--and predict the value of Ŷ. The lowest value of X is 3.00, and 7.2 is the highest value of X.

\[ \hat{Y} = -3.08 + (1.66)3.00 \]

\[ \hat{Y} = -3.08 + 4.98 \]

\[ \hat{Y} = 1.9 \]

\[ \hat{Y} = -3.08 + (1.66)7.2 \]

\[ \hat{Y} = -3.08 + 11.95 \]

\[ \hat{Y} = 8.87 \]

These are just two values of X used to predict Y. We can also predict Y from each of our values of X in the distribution. From Yonker, p. 33, you can again see the distribution of the 10 libraries on Per Capita Expenditure (X) and Per Capita Circulation (Y). You can also see the predicted values of Y from each value of X in the distribution.
ARE THE PREDICTED VALUES OF Y THE SAME AS OUR ACTUAL VALUES OF Y?
The predicted values of Y are not the same as the actual values of Y, but they are close. The predicted values of Y are close to the actual values because the correlation between X and Y is very strong (.97); therefore, prediction of Y from X is very good. But, because the correlation was not perfect, the predicted values of Y are not the same as the actual values of Y! We have some errors in prediction!

Errors in Prediction:

We can visualize the concept of errors in prediction by drawing the regression line through the original scatter gram of this data (see Yonker, p.34). The easiest way to draw the line is by taking the lowest value of X (3.00) and the highest value of X (7.2) and plotting our points where these values meet the corresponding predicted values of Y (1.9) and (8.87) that were calculated above. Drawing a line through these points includes all the other predicted values of Y from X. We can see how our actual values deviate from the regression line—the errors in prediction.

Also, imagine the regression line crossing the X axis and intercepting Y. Where do you think the regression line will intercept Y? That's right, at -3.08!

Another way to address errors in prediction is to calculate a measure called the Standard Error of Estimate. It is defined as the standard deviation around the regression line. It is a measure of how the actual values of Y cluster around the regression line that includes our predicted values of Y.

From Yonker, p 33, you can see the column \(Y - \hat{Y}\), which is a step in the calculation of the standard error of estimate. To obtain the values in this column, you subtract each predicted value of Y from the actual values. As an example, from the top, 4.5 – 4.56 = - .06. The numbers in this column are a mix of both positive and negative numbers, which when added, result in zero (or very close to zero).

The column, \((Y - \hat{Y})^2\) eliminates the negative numbers, so we can continue with the calculation of the standard error of estimate. As an example, from the top, \((-0.06)^2 = .0036\).

The sum of the column \((Y - \hat{Y})^2\) = 2.641. This number is used to calculate the standard error of estimate. The formula follows (note that the symbol for the standard error of estimate, \(s_{Y/X}\), refers to the standard error in Y predicted by X):

\[
s_{Y/X} = \text{square root of: } \frac{\text{sum of } (Y - \hat{Y})^2}{N - 2}
\]

\[
\text{square root of: } 2.641 = .574
\]

\[
10 - 2
\]
.574 is the standard error of estimate

This measure is similar to the standard deviation only it is around the regression line—it shows the overall error on either side of the regression line. You can add and subtract the standard error of estimate from each predicted value of Y to get a constant percentage area around the regression line within which you would expect actual values to cluster. You can think of the standard error of estimate as forming normal curves around each predicted value on the regression line (If you need, review the normal curve, and the percentage areas under the normal curve from session 3).

Again, look at Yonker, p. 33, and you will see the predicted Y of 4.56 at the top of the column. Using the standard error of estimate—

\[ 4.56 \pm .574 = \]
\[ 3.986 - 5.134 \]

**About 68% of the actual values of Y will fall within these limits of 3.986 to 5.134.**

\[ 4.56 \pm (2)(.574) = \]
\[ 4.56 \pm 1.148 = \]
\[ 3.412 - 5.708 \]

**About 95% of the actual values of Y will fall within these limits of 3.412 to 5.708.**

\[ 4.56 \pm (3)(.574) = \]
\[ 4.56 \pm 1.722 = \]
\[ 2.838 - 6.282 \]

**About 99% of the actual values of Y will fall within these limits of 2.838 – 6.282.**

It is important to note that the smaller the standard error of estimate, the closer the actual values of Y will cluster around the regression line, and the more confident you can be in the prediction of Y. The larger the standard error of estimate, the more the actual values of Y spread out around the regression line, and the less confident you can be in the prediction of Y.

**A VERY IMPORTANT NOTE: THE STRONGER THE CORRELATION (r), THE SMALLER THE STANDARD ERROR OF ESTIMATE, AND THE BETTER THE PREDICTION OF Y FROM X.**

**Why do we need regression?**
In this example of our libraries, we had the actual values of Y (per capita circulation), so why would we want to go through the ordeal of predicting Y? Suppose we had an 11th library, and we knew the value of per capita expenditure (X), but not the value for per capita circulation (Y). Knowing that there is a strong correlation between these two variables, we could construct the regression equation, and get a very good estimate of per capita circulation for that 11th library.

In other "real life" situations, regression can be used to predict one variable from another. For example, if an admissions office in a university found that there was a strong correlation between GRE scores and Graduate GPA, the office could construct a regression equation from actual GRE and GPA data. In the future, that regression equation could be used to predict Graduate GPA from GRE scores at application.

**Curvilinear Relationships:**

Because regression analysis (as discussed in this session) depends on strong correlations, or linear relationships, we would not use this equation to predict curvilinear relationships. Examples of curvilinear relationships are shown in Yonker, p. 35.

Graph a. shows a **positive curvilinear relationship**. As variable X begins to increase, so does Y (making the relationship positive). But, as Y continues to increase, X remains constant, forming a curvilinear relationship.

Graph b. shows a **negative curvilinear relationship**. As variable Y begins to decrease, variable X begins to increase (making the relationship negative). But, as X continues to increase, Y remains constant, forming a curvilinear relationship.
Week 8 - CROSSTABULATIONS AND CHI SQUARE

Lecture Notes - Week 8

Chi Square Summary.ppt  (20992 Bytes )

Online_515_Winter_04.spo  (207360 Bytes )

OutputChiSq.spo  (23040 Bytes )

CROSSTABULATIONS AND CHI SQUARE

Cross tabulation:

Remember the survey you completed during the first week of the course? Be sure to check out the SPSS output (attached below **If the icon does not work, use the attachment at the top of the lecture notes: Online_515_Winter_04.spo) which contains analyses of the data from these surveys. You should look over the survey results as a review of the types of statistics we have covered in this course. The results show the true personality of the class!

[Online_515_Winter_04.spo]

Cross tabulation:

New Salary * New Height Crosstabulation

<table>
<thead>
<tr>
<th></th>
<th>Dwarves</th>
<th>Giants</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>New Salary</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cheap</strong></td>
<td>9</td>
<td>7</td>
<td>16</td>
</tr>
<tr>
<td>% within New Height</td>
<td>69.2%</td>
<td>53.8%</td>
<td>61.5%</td>
</tr>
<tr>
<td><strong>Expensive</strong></td>
<td>4</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>% within New Height</td>
<td>30.8%</td>
<td>46.2%</td>
<td>38.5%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>13</td>
<td>13</td>
<td>26</td>
</tr>
<tr>
<td>% within New Height</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Look at the table called a Cross tabulation in the SPSS output above (you will have to scroll through the output to find it). A cross tabulation allows you to see the co-occurrence of two variables by their joint frequencies. This table is called a 2 X 2 table (otherwise known as a fourfold table). You can see two variables with two levels (or categories) of each variable. I took the two variables, height and expected salary, and reduced them to two categories (a dichotomy--a variable with two values).

If a person is equal to or below the mean in height, he/she falls in the Dwarves category. If a person is above the mean in height, the individual falls in the Giants...
category. For salary, if you are equal to or below the mean, you are Cheap (please no offense, this is all in fun!!), and if you are above the mean, you are Expensive.

In a cross tabulation, the columns usually include the independent variable and the rows the dependent variable. We may have reason to think that a person’s height (independent variable) may somehow affect their perception of their salary level (dependent variable). The numbers in the center of the cells of the table are called the observed (AKA, obtained) frequencies. You should remember that a frequency is the number of times a value occurs. The frequencies in a cross tabulation express the relationship between the two variables—the differences in one variable at different levels of the other variable.

You also see percentages in this table. The rule is usually to calculate the percentages from the independent variable to the dependent variable. So, the percentages are calculated from the column total, which represents each independent variable. The percentages provide a more accurate picture of categories then the raw numbers. In order to interpret the percentages, you compare across the columns in looking for differences. Hmmm....in our "Cheap Dwarves" table, it does not appear that too much is happening in the percentages--height does not seem to be a major influence in salary expectations!

**Chi Square:**

We have been using frequencies and percentages to interpret the data in the cross tabulation (by observation) for the “cheap dwarves.” Suppose we want to find out if the relationship between the variables is significant. We can test for significance with a statistic called Chi Square (also called Pearson Chi Square, or Chi Square Test of Independence). Chi Square is a nonparametric statistic. The tests we've used thus far like t, z, F, and r are called parametric statistics. Parametric statistics examine population parameters, such as means and standard deviations, in the calculations of the statistics. Parametric statistics differ from nonparametric statistics in four major ways. Parametric statistics:

1. Are more powerful tests than nonparametric statistics. Parametrics are considered mathematically more powerful.
2. Require more assumptions--like normal distribution and linearity--than nonparametrics.
3. Require data at higher levels of measurement--interval and ratio--than nonparametrics that are used mainly with nominal and ordinal data.
4. Use the full value of data. If interval and ratio level data does not meet the assumptions for parametric statistics, nonparametric statistics can be applied to the data. The data can be changed to nominal or ordinal level of measurement, which involves redefining or recoding variables, similar to the “cheap dwarves” data. This change in the level of measurement produces loss in the value of the data.
Chi Square is a very commonly used nonparametric statistic that is used with nominal level of measurement. Even if you apply Chi Square on ordinal data, Chi square treats the data as nominal; therefore, the greater than/lesser than quality of ordinal data is lost.

See Yonker, p. 36, for an example of the use of Chi Square. In this example, we have data from a random sample of 60 school children and we want to see if there is a relationship between males and females and their preference for reading fiction or nonfiction.

The null hypothesis states that there is no relationship between males and females and their preference for fiction or nonfiction. Chi square will tell us if there is a significant relationship (meaning that the relationship is not a fluke or not due to chance), but not the strength of that relationship. Other statistics of association can tell us about the strength of the relationship and we will discuss some of them later.

Chi square is a measure that examines how obtained/observed frequencies (in the center of the cells) differ from frequencies we would expect if our null hypothesis were true. The symbol for obtained/observed frequencies is fo. The numbers in the upper left corner of the cells are called expected frequencies. The symbol for expected frequencies is fe.

Expected frequencies show how the table would look if the null hypothesis were true. Expected frequencies are calculated from the totals in our columns and rows called Marginal Totals. Expected frequencies must be calculated separately for each cell. The following demonstrates the calculation of expected frequencies.

Expected frequencies (fe) = (total frequency in column) X (total frequency in row)/N

**Examples from Yonker, p. 36:**

Expected frequency for females who prefer nonfiction (cell d):

\[
30 \times 40 = \frac{1200}{60} = 20
\]

60  60

30 is the total frequency in the column for cell d. and 40 is the total frequency in the row for cell d.
Expected frequency for males who prefer fiction (cell a):

\[
\frac{30 \times 20}{60} = \frac{600}{60} = 10
\]

30 is the total frequency in the column for cell a, and 20 is the total frequency in the row for cell a.

Expected frequencies must be calculated for each cell—There are no shortcuts for finding the expected frequencies.

In Yonker, p. 36, you can also see the calculation of chi square. Each column of the calculations (shown below the cross tabulation) represents a portion of the Chi Square formula which follows:

**Chi Square is the sum of \((fo – fe)^2 / fe\) over all cells in the table**

The formula serves as a guide for setting up the columns in the calculations. You will notice that each cell in the cross tabulation is labeled with a letter. The letters serve to organize the calculations for each cell. The following are the steps for calculating Chi Square:

1. For each cell, list the observed frequencies and the expected frequencies (fo and fe columns). For cell a., 6 (fo) and 10 (fe) are shown.
2. For each cell, calculate observed frequency minus expected frequency (fo – fe column). For cell a., 6 – 10 = -4.
3. For each cell, square the result of the fo – fe column, which eliminates the negative numbers. For cell a., \((-4)^2 = 16\).
4. For each cell, take the result of the \((fo – fe)^2\) column and divide by the corresponding expected frequency (fe). For cell a., divide 16, which is \((fo – fe)^2\), by 10 (fe). The result is 1.6.
5. The last step is to add the numbers in the \((fo – fe)^2 /fe\) column. This is Chi Square.

**Chi Square (obtained value) = 4.8**

As with other tests of hypotheses, Chi square also requires a significance level with a corresponding degree of freedom. The significance level was set at .05 and the degree of freedom is obtained as follows:

\[
df = (\text{number of rows}-1)(\text{number of columns}-1)
df = (2-1)(2-1) = 1
\]

Go to Yonker, p. 42, and you will find a table of **Critical Values for Chi Square**. At .05 with 1df, the critical value is 3.84.

**Critical Value = 3.84**
The obtained value of 4.8 is beyond the critical value of 3.84, so we can reject the null hypothesis. We can say that there is a significant relationship between a child’s gender and the kind of books he or she prefers (and the results are not due to chance!), but we cannot say anything about the strength of the relationship. Chi Square will also not give any information about which cells contribute to the relationship—Chi Square is an overall test of significance.

**Standardized Residuals:**

**Standardized residuals** allow you to see which of the cells in the cross tabulation are major contributors to the significance of Chi Square. When the residual in a cell is greater than 2 (absolute value) you can conclude that the cell is a major contributor to the significant Chi Square. If no cell has a residual greater than 2, then you can conclude that no cell “stands out” as a major contributor to the significant Chi Square. Calculation of the standardized residual is accomplished with the following formula and is calculated for each cell in the cross tabulation.

**Standardized residual = (fo – fe)/square root of fe**

The following represents the calculation of each residual from the example in Yonker, p. 36:

Cell a: –4 (which is fo –fe) divided by the square root of 10 (fe)
  \[-4/3.16 = -1.3\]

Cell b: 4 divided by square root of 20
  \[4/4.47 = .89\]

Cell c: 4 divided by square root of 10
  \[4/3.16 = 1.3\]

Cell d: –4 divided by square root of 20
  \[-4/4.47 = -.89\]

Because the residuals are not greater than 2, we do not have any cell that stands out as a major contributor to the significant Chi Square. But, residuals can aid in the interpretation of the significant Chi Square. We can say for example:

1. In cell a., males prefer fiction less than expected because the residual is a negative number, -1.3.
2. In cell c., females prefer fiction more than expected because the residual is a positive number, 1.3.
Assumptions for Chi Square:

Even though Chi Square does not have the strict assumptions of a parametric statistic (remember Chi Square is nonparametric), there are certain assumptions/cautions for using Chi Square.

1. There should be no observed frequencies of zero in the cross tabulation.

   Solutions:

   a. Combine/recode or omit categories, if you can, to eliminate the zero cells. Example of combining categories (recoding): Combine the categories of never and sometimes to “no.” Combine the categories of frequently and always to “yes.”

   b. Or, in a 2x2 table, where categories are not easily combined, use a statistic called Fisher’s Exact. You will see Fisher’s Exact automatically reported for a 2X2 table using SPSS.

2. The cells must be independent, meaning each case should only be in one cell. If you are a female who prefers fiction, you cannot be a female who prefers nonfiction as well.

3. There should be no expected frequencies less than 1.

4. At least 80% of the expected frequencies should be greater than or equal to five. For a 2X2 table, no expected frequencies should be less than five.

   Solutions to 3 and 4 above: again, combine, recode, or omit categories. Or, for a 2X2 table, use Fisher’s Exact.

5. **Chi Square is more reliable with smaller sample sizes.** With very large sample sizes (such as hundreds), Chi Square will likely be significant. In the example illustrated above regarding gender preference for fiction or nonfiction, if you multiplied each observed value by 10, recalculated the expected frequency and Chi Square, the result would be multiplied by ten as well. Our obtained value was 4.8. With this change in the number of observed values, Chi Square would be 48 and still with 1df!

Yates Correction:

Although there is some debate as to whether or not it is necessary, researchers often use what is called the **Yates Correction for Continuity** in the computation of Chi Square for a 2x2 table. SPSS automatically reports this statistic for a 2X2 table, as you will see in your homework. The Yates Correction provides a more conservative value of Chi Square for a 2x2 table. Some statisticians say this is an unnecessary reduction in the power of Chi Square, yet the Yates Correction is usually reported.
**What is Yates correction?** In the numerator of the Chi Square formula, the Yates Correction is simply: \( (fo - fe) - .5\)^2 where the absolute value of \((fo - fe)\) is used.

The correction is .5 and it is subtracted from the absolute value of \(fo - fe\).

Some major points which summarize Chi Square are included in the following PowerPoint Slide (**If the icon does not work, use the attachment at the top of the lecture notes for the Powerpoint slide):

[Chi_Square_Summary.ppt]

---

**Chi-square**

- Measures how much observed frequencies differ from expected frequencies
- Considered *Nonparametric*
  - does not require assumptions about the shape of the population distribution
  - does not require variables be measured on an interval or ratio scale
- Allows us to test for a relationship (not the strength or direction)
- Null hypothesis = no relationship

---

**A Layered Cross tabulation:**

Let’s look at another example of a cross tabulation and the computation of Chi Square when we have a layer variable—you will also see a layer variable in your final homework assignment. View, or print out the attached SPSS table (**If the icon does not work, use the attachment OutputChiSq.spo at the top of the lecture notes.**).

[OutputChiSq.spo]
Here we have a table with 2 independent variables, gender and marital status. Marital status is in the columns of the table and gender is the layer variable, which is really like a control variable. The dependent variable, in the rows, is “how’s life.” You should notice that we have column percentages taken to the base of the independent variable (marital status) for each gender. If you want to see the impact of the independent variables, you can compare the percentages across columns for each gender. Think about a couple of things as you are looking at and interpreting the table. These are just questions for thought, not a homework assignment.
1. Do you see any interesting results when you compare percentages across columns?
2. Look at the Chi Square results. Are the results significant?
3. Look at the Standardized Residuals. Do you see any cells that are major contributors to a significant Chi Square?

**Measures of Association for Nominal Level of Measurement:**

Chi Square tells us if there is a significant relationship and the standardized residuals tell us what cells contribute to the significance, but we may also want to see if the relationship is strong. For nominal variables, only strength is considered since direction is meaningless for such categories. For example, you would not say marital status increases as satisfaction with life increases? Direction is not indicated for marital status.

There are three common measures of association for nominal variables:

1. Phi
2. Contingency Coefficient
3. Cramer’s V

**The result of these three measures is based on a decimal value from 0 to 1, with 0 meaning no association, and 1 meaning a perfect association.** All of these measures are independent of the size of the sample; therefore, it is very good to report the results of these statistics when you have a large sample size. This is because the Chi Square could be significant (due to the large sample size), yet there could be little or no strength to the relationship.

**Phi** is used with a 2x2 table. The formula for Phi is as follows and is calculated using the Chi Square example from Yonker, p. 36:

Take the square root of result-- Chi Square/N

Square root of result-- 4.8/60 = .28

**Phi = .28**

Even though the Chi Square from this example is significant, it is not a strong relationship.

**The Contingency Coefficient** can be used with a 2X2 table or larger table. The drawback to this statistic is that it can never reach perfect 1; therefore, Phi is a preferred statistic for the 2X2 table. The Contingency Coefficient is also calculated from the example in Yonker, p. 36.

Take the square root of result-- Chi Square/(Chi Square + N)
Square root of result—sqr[t\(\frac{4.8}{(4.8 + 60)}\)] = .27

**Contingency Coefficient = .27**

The results are similar to Phi. The relationship is not strong.

A measure which is preferred to the Contingency Coefficient for a table larger than 2X2 is **Cramer's V**. I will use the example of Marital Status/How’s Life, which was shown earlier in this session (and, you should still have the SPSS results), to illustrate the calculation of Cramer’s V. Let’s consider the Chi Square result for females:

Take the square root of result—Chi Square/[N(k – 1)]

Where k is whatever is smaller—the number of rows or number of columns. In this example, there are 4 columns and 3 rows, so k = 3.

Chi Square for females = 19.1

N for females = 524

Cramer’s V = Square root of result—19.1/524(3 – 1)

Square root of result—19.1/1048 = .135

**Cramer’s V = .135**

Again, in this example, we have significant Chi Square results, but the relationship is not strong.

**Measures of Association for Ordinal Variables?**

In the **Reference Statistics Workbook**, which is part of your assignment for next week, you will be introduced to a measure of association for ordinal level variables called **Yule's Q**. Yule’s Q is designed for a 2X2 table. **Gamma** is a measure of association for larger than a 2x2 table (ordinal variables). Both Gamma and Yule’s Q range from –1 to +1 with 0 indicating no relationship. We will focus on Yule’s Q through the Reference Statistics Workbook.

In your homework for this week, you will find Gamma listed as the measure of association for the 2X2 table. The Gamma is really a Yule’s Q—SPSS does not distinguish between the two statistics. Try to interpret the result of “Gamma” in your homework. You should have a sense of this statistic after you read the Reference Statistics Workbook.
Week 9 - REFERENCE STATISTICS WORKBOOK

Introduction:

Read through the Reference Statistics Workbook at the back of your "booklet" (Yonker)-it is located after p. 42. Try to do the exercises. The notes below are supplemental to the workbook, follow it page by page, and provide the answers to the exercises. I will also add my own comments to some of the subjects covered in the workbook. The notes below will, decidedly, emphasize sample size, Yule's Q, and percentages in cross tabulations.

If you feel the reference to "punched cards" makes the workbook out of date, then jump over that section! But I assure you the setting described, the procedures, and the analyses used are not out of date!

So, what will you find in the Reference Statistics Workbook?

1. It shows you the step-by-step procedures in conducting action research in an actual library setting, specifically, evaluation of the reference service.

2. It covers the process of determining sample size, all the way through to the analysis of data.

3. It deals with dichotomous variables. A dichotomy is a variable with two values.

4. It deals with ordinal level variables.

As you look at the variables discussed in the workbook, they may seem like categories, or what you would usually think of as nominal level of measurement. But, the way these variables are defined, gives them ordinal quality. If you look at the definition of ordinal level of measurement on p. 2 of the workbook, it reiterates the "ranking" or "greater than/lesser than" characteristic of ordinal variables. So, if you think of library staff as being professional or nonprofessional, the professionals have greater skill than the nonprofessionals-this makes the variable ordinal. Or, if you consider reference versus directional questions (both defined on p. 1 of the workbook), the reference questions require more skill to answer than the directional questions.

Sample Size:

The Reference Statistics Workbook on p. 3 only discusses statistical factors related to the dichotomy. The two statistical factors discussed are:
1. **Tolerance**: This is defined as the percent of error you are willing to tolerate on either side of your sample estimate. Tolerance involves setting limits around your sample estimate-limits within which you would expect the inclusion of the actual population value.

Say you have sample estimates which show that your reference desk (during a specified time period) receives 75% reference questions and 25% directional questions. Reference questions 75%, and directional questions 25% are called the "split" in the dichotomy. Suppose also, that you set your tolerance at 5%. This means that you would expect the true population value to occur within the limits: 70 to 80 percent reference questions (75 +/- 5) and 20 to 30 percent directional questions (25 +/- 5).

2. **Confidence**: This is the confidence you want to have in your tolerance limits. As we have seen with confidence intervals, usually your probability level is 95% or 99% (Howard White also mentions 90% in the workbook). This means that your tolerance limits will contain the true population value in 95 or 99 out of 100 samples.

The combination of error tolerance and confidence helps you to determine your sample size. A way to determine sample size is to consult a table which takes into account both error and confidence. Carl Drott (a professor at IST) developed a table which is on p. 4 of the workbook. This table was constructed with a formula that uses a conservative estimate of the split between dichotomous variables, a 50/50 split, which is the least mathematically precise estimate. A 50/50 split in the dichotomy (50% reference questions, 50% directional questions) is the most conservative estimate because it assumes you don't have much knowledge of your population. It assumes that you have no idea, for example, of how many reference versus directional questions the reference desk receives. And, when you do not know much about your population, you need a larger sample size-you need to "work harder" in conducting your study.

Looking at the exercises on p. 4, the first question asks you: How large a sample would you need to estimate a true population value within half a percent either way at the 95 percent confidence level? By looking at the table, the answer is 38,416. The second question asks: How large would the sample have to be at the same confidence level, but with a tolerance of five percent error either way? The answer is 384. You will notice through these examples that **if you are willing to tolerate more error, your sample size is drastically reduced.**

In general, the less error you are willing to tolerate combined with a higher percentage confidence, the larger your sample size must be. If you are willing to tolerate more error, your sample size will be smaller (you don't have to work as hard in conducting the study). If you are willing to have less confidence in your tolerance limits, again, your sample size will be smaller.

If you can be more mathematically precise about the dichotomous variable in the population, you are rewarded (you do not have to work as hard in conducting the study!) by being able to have a smaller sample size. A correction factor is applied to the split in the dichotomy to reduce your sample size. That correction factor is simply the number...
4. Page 5 in the workbook includes exercises which apply the correction factor. The first exercise assumes knowledge of the population to be a 33% / 67% split in the dichotomy. Applying the correction factor of 4:

\[ .33 \times .67 \times 4 = .884 \]

\[ .884 \times 384 = 340 \]

The sample size is reduced from 384 because you have more precise knowledge about your population.

The second exercise on p. 5 asks you to apply the correction factor to the 50/50 split:

\[ .5 \times .5 = .25 \]

\[ .25 \times 4 = 1 \]

\[ 1 \times 384 = 384 \]

The correction factor has no impact on the 50/50 split!

Sample size is not dependent on the population size, because even small samples taken at random are fairly accurate; however, sample size may depend on your knowledge of the population size. Drott's table is conservative and assumes very large populations. If you know that your population is smaller, you could use a table, such as Arkin's, which is described on p. 5 of the workbook. The exercise at the bottom of p. 5 asks you to use Arkin's table (this is only one page of this table, showing a 30/70 split) for 5% tolerance and 95% confidence. Looking at the table, you should find the answer 314. Once again, this is a reduction in the sample size of 384 at the same tolerance and confidence, but with knowledge of the population size.

It was mentioned earlier that Carl Drott used a formula to construct the table on page 4. You could use this formula as well to directly calculate the sample size. The formula and an example follow:

\[ n = \frac{(z^2)(pq)}{se_p^2} \]

Where-
- \( n \) = size of sample
- \( z \) = critical \( z \) based on confidence level (e.g., 1.96 for 95%)
- \( pq \) = split in the dichotomy (such as 50/50)
- \( se_p \) = margin for error you are willing to tolerate (such as 5%)

Example-

\[ n = \frac{(1.96^2)(.5 \times .5)}{(.05)^2} = \frac{(3.84)(.25)}{.0025} = \frac{.96}{.0025} = 384 \]
Where $1.96$ is the confidence level

$.5$ (50%) is the split in the dichotomy

$.05$ (5%) is the error

Sample size is actually quite a bit more complicated when the variable is other than a dichotomy. If you want to know a bit more about the complexities of sample size, consult the two references below:

The author is Hinkle (et al) and both articles are in Educational and Psychological Measurement- at the following volumes, years, and page numbers:

- 43(1983): 1051 - 1060

Also, Carl Drott's web site (http://drott.cis.drexel.edu/) is also a good source for information on sample size.

**Other Factors Important for Sample Size:**

We have seen from discussion of z and t that generally, samples with more than 30 are considered large samples, and samples with less than or equal to 30 are considered small samples. For some statistical tests, that holds true. For other statistical tests, where you have a large number of variables, 30 or a bit more could be too small of a sample. Generally, sample size is a compromise of a number of factors:

1. The type of study you are conducting. A survey sent through the mail would probably require a larger sample size than an experiment with two conditions (such as treatment and control).
2. The care in which you select the sample. This factor refers to a good definition of the population and a sample selected at random. Even small samples taken at random are an accurate representation of the defined population.
3. The time and money to handle the sample. This is a practical factor which may preclude taking a large sample.
4. Availability of subjects.
5. The type of statistical test you want to use. Some advanced techniques require a larger sample size.

On page 7 through 9, the workbook discusses taking a sample at random. I am omitting this part of the workbook since we previously discussed this in session 3. Feel free to read through the section in the workbook as a review.
Coding and Recoding:

In the face-to-face Action Research class, we do not spend much time discussing coding and recoding, but I do provide them with the answers to the exercises on pages 10 through 12. Try to do the exercises and then check your answers with those provided below.

Using the variables and the codes listed on p. 10, you are asked to "content analyze" the short narratives on p. 11 and code the appropriate category for each variable. The answers for each example follow (the code is provided in parentheses):

Example One:

Time: 10:05 a.m. (2)

Day: Monday (1)

Answerer: Professional (2)

Customer: Faculty (4)

Inquiry: Person (1)

Duration: 3 seconds (1)

Question Type: Directional (1)

Action Taken: Directions (1)

Sources: None (1)

Novelty: Not applicable (8)

Example Two:

Time: 9:30 p.m. (7)

Day: Wednesday (3)

Answerer: Nonprofessional (1)

Customer: Graduate student (2)

Inquiry: Telephone (2)
On p. 12, you are asked to do some exercises in recoding. I mentioned recoding briefly last session as a way of combining or redefining categories of a variable. Do you remember why we might want to do this? Two reasons:

1. When you have zero cells or low expected frequencies in a cross tabulation and you want to combine categories.

2. When interval or ratio level data do not meet the assumptions for parametric statistics. You could redefine (recode) the variables to an ordinal or a nominal level and then use nonparametric tests.
The first exercise on p. 12 asks you to recode the variable "action taken" to the dichotomy of "directional transactions" and "reference transactions."

**Answer:** Direction transactions consist of categories 1 through 3 (directions, information on policy, other help) and reference transactions consist of categories 4 through 8 (recommendation to in-house source, use of in-house source, interpretation, instruction, referral).

The second exercise asks you to recode the variable "sources used" to the dichotomy of "bibliographic" or "reference" sources.

**Answer:** Reference sources consist of categories 6 and 7 (general reference, specialized reference) and bibliographic sources consist of categories 2 through 5 (card catalog, abstracts/indexes, national or trade bibliography, computerized bibliography). Categories 1, 8, and 9 (none, circulating book, non-librarian) do not map to reference or bibliographic sources.

**Cross tabulations Continued and Yule's Q:**

On p. 13 of the workbook, you see a 2 X 2, also know as a fourfold table. Using the transactions listed, you are to place the appropriate frequencies in the cells of the table. The answers are listed as follows:

- Nonprofessional/Directional cell-4
- Nonprofessional/Reference cell-0
- Professional/Directional cell-3
- Professional/Reference cell-3

A review of the 2 X 2 table is an appropriate introduction to Yule's Q. Yule's Q is a measure of association for a 2 X 2 table. It is a nonparametric measure of association for ordinal level of measurement. The result of Q is a decimal value between +1 and -1. The closer the decimal value reaches either positive 1 or negative 1, the stronger the association. As the Yule's Q approaches zero, there is little or no association. The sign of Q indicates direction. Because Yule's Q looks at the strength and direction of an association, it is similar to Pearson's r; however, Pearson's r is a parametric statistic used with higher levels of measurement.

Page 14 of the workbook shows you cross tabulations for 3 different associations: perfect positive, perfect negative, and no association. It may be helpful for you to substitute high and low with categories of a variable. In the columns of the table, low could refer to nonprofessionals and high could refer to professionals. Remember, this is an ordinal variable where professionals are considered to have greater skill than nonprofessionals. In the rows of the table, low could refer to directional questions and...
high could refer to reference questions. Reference questions require greater skill to answer than directional questions.

You can see, for example, that the perfect positive association on page 14 shows the frequencies of 5 in the low/low cell and the high/high cell. This is consistent with what we know about a positive relationship. In the other cells, low/high and high/low, we have 0 frequencies.

Page 16, shows the formula and the calculation of Yule's Q. You are asked to calculate Yule’s Q for the four tables on page 17 and 18. I will show you the calculation of Q for the first table and then provide you with the answers for the other three tables. Take note of the management goal at the bottom of p. 16. This goal will give you a clue to which Yule's Q shows "success!"

Calculation of Yule's Q for the first table on p. 17:

\[ Q = \frac{ad - bc}{ad + bc} \]

Note: For Chi Square, it did not matter how the letters in the cells were placed. For Yule’s Q, the cells must be lettered as shown. This is because the cells represent the positive and negative relationships.

\[ Q = \frac{(150 \times 100) - (34 \times 100)}{(150 \times 100) + (34 \times 100)} = \frac{15000 - 3400}{15000 + 3400} = \frac{11600}{18400} = .63 \]

To interpret this result, you can see that it is positive because lots of nonprofessionals are answering directional questions (cell a.), and lots of professionals are answering reference questions (cell d.). But, in cell c., we still have the nonprofessionals answering reference questions. Cell c. keeps the association from being stronger.

**Answers to the other three tables:**

Bottom table on p. 17, \( Q = -.63 \)

Top table on p. 18, \( Q = .02 \)

Bottom table on p. 18, \( Q = 1 \)

Consider the interpretation of the other three tables based on the Yule's Q values shown above.

1. Why do we have a negative Q for the bottom table of p. 17?
2. Why is Q .02 for the top table on p. 18?
3. Why does the Q of 1 indicate success for the bottom table on p. 18?

The questions above are not part of the written homework assignment, but you should certainly think about the answers!
On p. 19, you are asked to interpret the three Yule’s Q values that are shown at the top of the page, by answering the questions at the bottom. Remember, a negative value of Q would indicate a "negative" outcome. The answers to the questions in order are as follows:

Long
Reference
Directional
Long
Short
Strong
D!!!

On p. 20, I believe we can say that the Yule's Q values shown for June are better (because they are stronger positive) than the January values; however, I do not think the library is doing well based on the management goal!

**Percentages in Cross tabulations:**

Page 21 discusses three types of percentages in cross tabulations: vertical, horizontal and total. It should be noted that vertical percentages are the most common because you are taking the percentage to the column total, which is the base of the independent variable-this was discussed in the last session.

You were asked to calculate vertical percentages for the four cross tabs shown on p. 22 and 23, and calculate horizontal and total percentages for the top cross tab on p. 22. *(Note: these are the same cross tabs used to calculate Yule's Q).* The percentages should be placed in the cells of the cross tabs on p. 24. Below, I have provided the percentages for each cross tab and also the answers to the questions that correspond to each cross tab:

**Table 1: Vertical**

<table>
<thead>
<tr>
<th></th>
<th>Non-pro</th>
<th>Pro</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dir.</td>
<td>60%</td>
<td>25%</td>
</tr>
<tr>
<td>Ref.</td>
<td>40%</td>
<td>75%</td>
</tr>
</tbody>
</table>
**Answers to Table 1 questions:**

40% of non-pros' transactions are reference questions.

75% of pros' transactions are reference questions.

There is a 35% difference between non-pros and pros in reference questions. Pros answer more reference questions.

**Table 2: Vertical**

<table>
<thead>
<tr>
<th></th>
<th>Non-pro</th>
<th>Pro</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dir.</td>
<td>25%</td>
<td>60%</td>
</tr>
<tr>
<td>Ref.</td>
<td>75%</td>
<td>40%</td>
</tr>
</tbody>
</table>

**Answers to Table 2 questions:**

Non-pros answered more reference questions in this table.

The "bad" outcome in Table 2 is indicated by the negative value of the Yule's Q.

**Table 3: Vertical**

<table>
<thead>
<tr>
<th></th>
<th>Non-pro</th>
<th>Pro</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dir.</td>
<td>48%</td>
<td>47%</td>
</tr>
<tr>
<td>Ref.</td>
<td>52%</td>
<td>53%</td>
</tr>
</tbody>
</table>

**Answers to Table 3 questions:**

1% difference separates pros and non-pros in reference questions.

The status of the answerer has a larger effect in Table 1 when compared to Table 3.

The Yule's Q for Table 1 was .63 and for Table 3, it was .02.

So, a large percentage difference across columns indicates a large degree of association, while a small percentage difference across columns indicates a small degree of association.
Table 4: Vertical

<table>
<thead>
<tr>
<th></th>
<th>Non-pro</th>
<th>Pro</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dir.</td>
<td>100%</td>
<td>0</td>
</tr>
<tr>
<td>Ref.</td>
<td>0</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 5: Horizontal

<table>
<thead>
<tr>
<th></th>
<th>Non-pro</th>
<th>Pro</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dir.</td>
<td>82%</td>
<td>18%</td>
</tr>
<tr>
<td>Ref.</td>
<td>50%</td>
<td>50%</td>
</tr>
</tbody>
</table>

Answer to Table 5 question:
The proportion of directional questions answered by non-pros is different than the proportion of reference questions (comparing down rows).

Table 6: Total

<table>
<thead>
<tr>
<th></th>
<th>Non-pro</th>
<th>Pro</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dir.</td>
<td>39%</td>
<td>9%</td>
</tr>
<tr>
<td>Ref.</td>
<td>26%</td>
<td>26%</td>
</tr>
</tbody>
</table>

Answer to Table 6 question:
Among all transactions 26% of reference questions are answered by non-pros.
Week 10 - REVIEW

As a summary look over the PowerPoint slides below, which list the statistics we have covered, and also put significance and association in perspective.

[week8pp2.ppt]

Significance vs Association
- Both are examples of inferential statistics
- Significance: establishes that chance can be ruled out as the most likely explanation of differences.
- Association: shows the nature, strength, and direction of relationship between 2 or more variables
- Common Tests of Significance:
  - t test (small groups),
  - F test (ANOVA),
  - Pearson chi-square (nominal levels of data)
- Common Measures of Association:
  - Pearson r (interval or ratio level),
  - Yules Q (ordinal level data in a fourfold table),
  - Gamma (ordinal--larger than fourfold table),
  - Phi (nominal level data in a fourfold table),
  - Contingency Coefficient (nominal--larger than fourfold)
  - Cramer’s V (nominal--larger than fourfold)
- Tests of significance and measures of association work in tandem with one another.
  - For example -- you might run a chi-square to determine statistical significance in the differences between two variables, and then run a Yule’s Q to show the relationship between the variables.
  - You can have statistical significance without having association