***This Module goes from 9/9 – 9/15***

***NOTE 1: THERE IS A LOT OF READING, BUT I HAVE BROKEN THEM INTO CHUNKS TO MAKE IT MORE DIGESTIBLE. THE TASKS ARE MINIMAL; IT’S REALLY ABOUT CHOOSING GOOD RESEARCH FOR YOUR TOPIC***

***NOTE 2: THERE ARE FOUR (4) PARTS TO THIS MODULE. I HAVE BROKEN THEM INTO TOPICS AND TASKS TO GET US GOING.***

|  |
| --- |
| **MODULE 2**  **QUANTITATIVE RESEARCH** |

**TOPIC 1**

|  |
| --- |
| **\*\*Essential Questions\*\*  What is a hypothesis?  What are the types of hypotheses and what do they mean?  What is a research question and how does it differ from a hypothesis?  How does a researcher know which to use?** |

In preparing to conduct a study, researchers typically begin by choosing a topic, based on interest or a question that is already in mind.  They will then conduct a literature review to learn all they can about what is already known on the topic, what is yet unknown, or what is controversial.  If they have a specific question, they will search to see if the answer has already been found. They will look for gaps in knowledge, next steps in current research, or areas that have never been explored, in order to choose a study that will add to the body of knowledge in their field.

At some point, the researcher will decide what is to be studied and must then decide how to design the study to serve the chosen purpose. At this point, the researcher will construct one or more***hypotheses***or ***research questions***, and the type of research - quantitative or qualitative - will naturally flow from that construction.

*A***hypothesis**is a formal prediction of the results of the research study.  It is the researcher's educated guess about what he or she will find out in the course of the research, written in such as way as to be *testable*via the data collected.

Suppose a researcher is interested in the teaching methods used by college history professors.  There are several avenues to explore in this topic, so let's say the researcher believes, through his initial literature review, that adding multi-media presentations to the traditional lecture format would increase student learning of history.  Before even starting the study, the researcher thinks he knows the answer; the purpose of the study then is to *test*his guess.  He would formulate a formal prediction - a hypothesis - that would allow him to apply the data he collects to see if his prediction has merit.

(It should be noted from the start that a researcher does not *prove* a hypothesis through research studies.  Researchers are very careful to say there is *significance of*or *significant evidence of* or *support for*any conclusions.  *Proof*connotes an absolute, and just one contrary or opposing instance wouldreveal the fallacy of that assertion.)

A good hypothesis is based on *sound reasoning and prior research*; is stated such that the *expected results are clear* by identifying variables and relationships; and is*testable*.  A hypothesis can be stated as a ***directional hypothesis***, a ***non-directional hypothesis***, or a ***null hypothesis***.

A ***directional hypothesis*** is one stating there is a relationship between the variables to be studied, and that the relationship is either positive or negative.  The researcher in our example above could construct a directional hypothesis as follows:

History 101 students who are instructed with both lecture and multimedia presentations earn higher grades than students who are instructed with lecture only.

The researcher has predicted that there is a relationship between lecture with multimedia (called the causal or *independent variable*) and grades (called the outcome, effect, or *dependent variable*), and that the relationship is a positive one, "higher grades."  Thus, it is predicted that the more multimedia presentations are added to the lecture format, the higher will be the student grades.  Or, to put it in researcher terms, as the independent variable increases, so does the dependent variable.

The same study could be conducted with a directional hypothesis written in the negative (although it's a little strained): History 101 students who are instructed with lecture only earn lower grades than students who are instructed with lecture plus multimedia presentations.

Here, the researcher has predicted that there is a relationship between the independent variable (lecture) and the dependent variable (grades), and that the relationship is a negative one, "lower grades."  It is predicted that the more students are taught with lecture only, the lower their grades will be.  In researcher terms, as the independent variable increases, the dependent variable decreases.

Writing this particular hypothesis in the negative may not be a good choice, since it makes "lecture" function as the independent variable, when the research actually intended to make "lecture with multimedia" function as the independent variable.  That is one reason why the wording of hypotheses is so important!

Here is an example of a good hypothesis written with a negative direction:

Students in noisy classrooms will score lower on tests.

In this example, the prediction is that there is a relationship between the independent variable (noisy classrooms) and the dependent variable (test scores), and the relationship is negative (lower scores).  As the independent variable increases (classrooms become noisier), the dependent variable decreases (test scores are lower).

A directional hypothesis is often called a ***one-tailed hypothesis***, with the tail being either positive or negative.  The reason for this terminology will become clear when we look at graphs in the coming weeks.

A ***non-directional hypothesis*** is one stating only that a relationship exists between the variables to be studied.  This type of hypothesis is used when the researcher believes there is a relationship, but is not sure whether the relationship is positive or negative.

The researcher in our example above could construct a non-directional hypothesis, as follows:

There is a significant difference between the grades of History 101 students who are instructed with multimedia presentations in addition to lectures and those students who are instructed with lecture only.

This hypothesis would be appropriate if the researcher believes there is a difference between the two teaching methods, but doesn't know if one is better than the other and wants to leave open the possibility that either lecture or lecture with multimedia may lead to higher grades.  Another example of a non-directional hypotheses would be:

There is a significant difference in attitude toward sports between eighth grade boys whose mothers were high school athletes and those whose mothers were not.

That would be an interesting study, wouldn't it?  If there is a difference in attitude, it is unknown whether it would be positive or negative.

Using a non-directional hypothesis allows the researcher to use evidence on either side (positive or negative) to support the hypothesis; because of this, a higher standard is required to show significance.  We will learn more about the specific standards later, but for now, just consider that if the researcher "hedges his bets" with a non-directional hypothesis, he has a higher standard to meet.

A non-directional hypothesis is also called a ***two-tailed hypothesis***, since the results could go in either a positive or negative direction.

A ***null hypothesis*** simply states the prediction that there is no significant relationship between the variables.  A null hypothesis from our history example would be written as follows:

There is no significant difference in grades between History 101 students who are instructed with lecture only and those who are instructed with lecture and multimedia presentations.

A null hypothesis is used when there is little known to support a hypothesis, or when the researcher wants to take advantage of the lower standards required of such a hypothesis.

If you remember the characteristics of quantitative studies, you will realize that all the hypotheses presented so far have been appropriate for quantitative research.  A hypothesis in which the variables and the data can be represented by numbers will serve in quantitative studies.  But is it possible to construct a hypothesis for a qualitative study?

The answer is*yes.* When comparing *groups,* rather than individuals, a guiding hypothesis can be formulated.  But it is more common for the qualitative researcher to start with a ***research question***.  Hypotheses more commonly are generated as a *result* of a qualitative study.

As you have learned, qualitative research addresses issues that cannot be easily counted, measured, or evaluated using quantitative procedures (*statistical analysis*).  Qualitative research attempts to answer questions surrounding *what is happening* and *why it is happening*.  Qualitative studies may begin with general ideas and concepts that gain focus as data is collected; they therefore may include multiple stages.  The concept of predicting a conclusion, as is required in quantitative research, is alien to qualitative research; hypotheses are therefore not constructed.

Instead, a ***research question*** guides the qualitative study.  A typical research question may be formulated with these words:

* How does X affect Y?
* What effects do X have on Y?
* What is the role of X in Y?
* Can Y improve/change with X?

A number of research questions were provided in last week's lesson, and it should be noted that the wording of those examples was exactly what would be used in the research study itself. There is no need to use terminology such as *variable* or *significant difference*or *statistics*in qualitative research questions; common language is fine.  It is more important that the researcher identify the purpose of the study, the phenomenon to explore, and the participant(s) who can provide the knowledge needed.

Two websites may be helpful at this point.  This page provides an explanation of quantitative and qualitative hypotheses/questions and includes practice exercises to help you distinguish among them. The website is: <http://www.bwgriffin.com/gsu/courses/edur8131/content/hypotheses.htm>.

This page contains a list of qualitative research questions asked by teachers and the studies that resulted from them. The website is: <http://curry.edschool.virginia.edu/go/clic/nrrc/ques_r30.html>.

***TOPIC 1 TASKS***

NONE ☺

**TOPIC 2**

|  |
| --- |
| **\*\*Essential Questions\*\*  What are populations and samples?  How are samples chosen?  What are data collection instruments?  What do *validity* and *reliability* mean?** |

When conducting *quantitative* research, it is usually not possible to get data from every person, organization, or object of interest to include in the study.  Instead, researchers collect data from ***samples*** of the ***target population*** in such a way that the results may be fairly *generalized* back to the population.

To make sure the sample is *valid*, it must be *representative* of the population; so the researcher must choose the sample appropriately.  So how are samples chosen?

Let's say we want to find out how satisfied the parents of a school district are with the education their children are getting.  It would be costly and tedious to send and analyze surveys of the thousands of parents in this *population.*  Fortunately, there are ways of selecting a *sample*of the parents that would provide us enough good data from which we can draw some valid conclusions.

The simplest way of choosing our sample is by ***random sampling***.  To choose a random sample, we would simply put the name of each parent in the school district into a hat and draw out the number we want to survey.  (A more high-tech way of doing the same thing is to list all the parent names in an Excel spreadsheet and sort them using the function =RAND().)

Random sampling is the easiest way to choose a sample.  It is fair, so we may reasonably generalize the results of our sample back to the population of "parents in the school district."

How would we select a random sample if the total number of parents was huge - say phone book size - and we didn't want to put all the names in a hat or type them into an Excel spreadsheet?  Well, if the names on the list are already random, that is, not segmented by school or ethnicity or any other characteristic we intend to study, we can use a method called ***systematic******random sampling***.  (Note that even if the names are alphabetical, they are still random, unless alphabetizing is somehow related to our study.)

To achieve this sample, we simply choose every 9th name or 15th name or Xth name that will give us the number of names we want.  If we have a population of 1000 names and we want to sample 100 people, we would choose every 10th name.   But there's a little twist.   To make the selection truly random, we don't start with the first name.  We start with a name somewhere between 1 and the interval number (in this example, 10, since we will select every 10th name).  And how do we choose that number?  Randomly!  We just pick a number.

So, for our example, we need a number between 1 and 10; let's just pick the number 7.  That means that the first name in our sample with be #7 on the list.  The second name will be ten names later, or #17.  And, we'll select each 10th name thereafter, until we have selected 100 names.

We have been talking about sampling the parents in the school district to find out their views about their children's education.  But what if we are also interested in learning about the satisfaction of parents of a particular school, or a particular grade level, or a particular neighborhood?  We won't be able to tell with any confidence that the random sample included enough parents from each of those sub-groups to be able to generalize the data in regard to them.  So, we may want to use ***stratified random sampling****.*  This is a two-step process:  first, we will separate the names of the parents in the school district into sub-groups, say for purposes of our example by school.  Second, we will select a random sample from each of the school parent groups.  When we put all the random samples together, we should have a sample that represents all the schools in the district.

In our example, let's say there are 1000 parents total in four schools - School A, School B, School C, and School D.  We decide to survey 100 parents.  Our procedure would be to make a list of all the parents in School A and choose a random sample of  25 parents.  We would do the same for Schools B, C, and D.  Our total sample size would be 25 + 25 + 25 + 25 = 100 parents.  This is a stratified random sample.

But what is School A is very small and only has 50 parents?  If we select a random sample of 25 parents from that school, it will represent a huge percentage (50%) of the total parents to be surveyed and our results will more likely reflect the views of those parents than any others.  How will that be fair?  What if School B has 800 parents and we select a random sample of 25 parents - a little more than 3% of the total number of parents to be surveyed?  How will that be fair?

The answer is to use percentages to decide how many parents from each sub-group should be included in the sample.  We may decide that if School A only has 5% of the parents, only 5% of the random sample should come from that school.  Or, we may decide to increase the representation of School A, even if it is disproportionate, to make sure we get enough data about the views of those parents (particularly if they are different from parents of larger schools, say from a rural environment or of a different ethnicity).  If we use the same percentage for each sub-group we are conducting *proportionate stratified random sampling*.  If we use different percentages for each sub-group, we are using *disproportionate stratified random sampling*.    Both types are acceptable; we just need to be sure to explain what we did and why in our research analysis.

What if the school district we want to survey is very large, such as the Clark County School District in Las Vegas, which in 2007-08 had 308,000 students in 341 schools?  How could we select a representational sample of parents that would make any sense without choosing parents from every school - which could mean surveying thousands of parents?

The solution may be***cluster random sampling***.  The process is to divide the population into clusters, usually along geographic boundaries such as states, counties, cities, or neighborhoods.  In our example, we could divide the school district by high school regions, with each region containing one high school and all its feeder schools.  Each of these regions would be called a cluster. Next, we randomly select an appropriate number of clusters - maybe three, four, five (we'll talk about how to decide the "appropriate number" later)  - which becomes our cluster sample.  Once the cluster sample is selected, we survey every parent in the sample, or...

We can further reduce the number of participants by ***multi-stage sampling****.*  In most applied social research, sampling methods are combined to achieve the most efficient and effective sample possible.  In our example of surveying parents of a school district as large as Clark County, we could start with cluster random sampling.  We could then further choose a random sample of parents of fourth graders in the cluster.  We could further randomly sample the parents of boys in the fourth grade.  We may have to use a stratifying process at any of these levels to make sure all different kinds of parents are represented.

The objective, remember, is to make sure we have a fair sample of the population that is representative of the population, so we can generalize any information we learn back to the population.

Researchers cannot always choose random samples.  For example, a teacher may survey all the students in his classroom, because they are the ***available****or****accessible*** representatives of a particular population.  A research organization may advertise for *volunteers* to participate in a study.  "*Man on the street"* surveys are often conducted to get quick information about public opinion.

These methods of sampling are examples of ***purposive non-random sampling***; that is, they have one or more specific predefined groups in mind from whom they are seeking information. Purposive sampling is useful when we need to reach a targeted sample quickly and easily, or when we need specific information from those knowledgeable about the topic, but it must be noted that it will be difficult to identify and describe the population from which the sample was taken or to whom the results may be generalized.

Purposive sampling is often used in *qualitative* research.  Qualitative research involves understanding a given phenomenon, and the participants may be specifically selected because of their knowledge or experience related to the phenomenon, not by random selection.

There are several kinds of purposive sampling:

* *Convenience sampling* uses whoever happens to be available; the "man on the street" survey and the classroom survey are convenience samples
* *Quota sampling* is surveying only people with specific characteristics; a volunteer participant study or a survey in which only certain people who walk by the surveyor are asked to participate may be quota sampling
* *Intensity sampling* includes participants at different levels of knowledge, success, or experience for a comparison of the differences; studying good readers vs. poor readers is an example of intensity sampling
* *Expert sampling* involves using a "panel of experts" to give views on a specific topic
* *Model instance sampling* refers to surveying the "typical cases" in a study; opinion polls of typical voters or typical home-schooling parents or typical bike riders would be examples of model instance sampling. Note that it may be difficult to define what "typical" is in this sampling method
* *Homogeneous sampling* includes participants who share a similar perspective or experience
* *Criterion sampling* includes participants or cases that share some similar characteristic or meet some set of specified criteria
* *Heterogeneity sampling* is also called "sampling for diversity."  This is the process of obtaining all possible relevant ideas or opinions on a topic, and generally includes a wide range of participants
* *Snowball sampling* is the process of identifying participants by recommendation from the first participant.  Snowball sampling may be necessary when trying to reach inaccessible participants such as the homeless or devotees of a particular sub-culture or members of a secret society.

***How large should a sample be?***

In *quantitative* research, the sample should be as large as possible so the researcher can generalize the results to the population.  Krejcie and Morgan developed a formula in 1970 to determined the appropriate sample size for a given population.  Generally, for populations of 100 or fewer, the researcher should sample the entire population.  If the population is 400-600, the researcher should sample 50%.  If the population is approximately 1,500, the researcher should sample 20%.  And for a population of 5,000 or more, 400 participants is sufficient.  Note that these numbers are *minimums*, and more rather than fewer participants are recommended.

The table is available online:  Krejcie, R.V. & Morgan, D.W. (1970)  Determining sample size for research activities.  *Educational and Psychological Measurement, 30*, 608.  Sage Publications.  Retrieved March 8, 2008 at <http://www.fns.usda.gov/fdd/processing/info/SalesVerificationTable.doc>.

*Qualitative* sample sizes are generally smaller and less representative because of the nature of the research.  A qualitative study may involve only one participant, and studies with more than 20 or so participants are rare. Sample size in qualitative studies is sufficient when the participants represent the entire range of potential participants.  For example, when studying a problem related to all the students in a particular school, the researcher will include representatives from each grade level, each gender, each characteristic to be studied, etc.  The sample size may have reached a legitimate maximum when the researcher begins hearing the same information from successive participants; this redundancy is called *data saturation.*

There are many ways to collect data for research studies.  A test may be administered, such as an achievement or aptitude test.  Interest or personality inventories may be used.  Attitude scales such as the Likert scale, a semantic differential scale or a rating scale may be used to determine what a person feels or believes.  Many of these instruments are standardized and/or published for use by researchers.  Surveys and questionnaires may be developed by the researcher for a particular study.  Quantitative research generally uses paper-and-pencil type methods of data collection, while qualitative research often uses observation and interview as the method of data collection.

There are several sources for published instruments.  The Mental Measurements Yearbooks, Tests in Print, and Tests: A Comprehensive Reference are the most prominent references for published instruments.  The Pacific University library has links and resources.

Whatever instrument, test, or assessment is used, it must be both *valid* and *reliable*.   Validity relates to the appropriateness of an instrument, while reliability relates to the consistency of the instrument.

***Validity***is the degree to which the instrument *measures what it is supposed to measure*, for the people using the instrument at a specific time. There are four types of validity that are interrelated:

* *Content validity* - this is the degree to which all test items are relevant to the content area tested (the item validity) and sample the full content (the sampling validity).  Achievement tests must have high content validity.  A test that includes topics the participants have not studied would have poor item validity.  A test in which some part of the content is not tested at all would have poor sampling validity.  Content validity is judged by experts who review the test as well as the process used to develop the test and make a determination about how well the test items represent the content it was designed to test.
* *Criterion-related validity (concurrent)*- this is the degree to which performance on a particular test relates to performance on a similar, pre-existing test (concurrent validity) .The second test is the criterion against which the first test is measured.  For example, if a new test claims to be easier to administer than an existing test, the validity of the new test can be measured by giving both tests to a group and comparing the results.
* *Criterion-related validity (predictive)* - this is another form of criterion-related validity which is the degree to which a given test predicts how well a person will do in the future.  Aptitude tests are intended to predict how well a person will do in a particular subject, career, or program of study, so they must have a high degree of criterion-related validity.  It may take a while to determine the validity of a predictive test, because the test must first be administered, then the test administrator must wait for the predicted behavior to occur.  Another measure of the criterion must be administered and the two sets of data correlated to see if the initial test was indeed a valid predictor of the behavior.
* *Construct validity* - this is the degree to which a test measures an intended construct.  A valid test will measure the intended construct and not some unexpected or interfering variable.  There is no single validation technique for construct validity; all the above validation measures must be used, as well as any methods for discovering un-validating factors.
* *Consequential validity* - this is the degree to which a test is harmful (usually without intention) or has negative consequences to the user.  A test in English, for example, may have negative consequences for non-English speakers, and the test results may not be fairly or accurately interpreted.  Consequential validity is determined by observation.

***Reliability*** is the degree to which the instrument is *consistent in measuring* what it is measuring.  Reliability is usually expressed as a reliability coefficient, which is a number between 0 and 1.00, with 1.00 being the perfectly reliable instrument.  Perfect reliability, of course, is impossible to achieve. The acceptable level of reliability differs according to the type of instrument, but generally, achievement and aptitude tests should have .90 reliability or higher, while other types may be lower.  New tests, predictive tests, and short tests are often less reliable.  So are tests used on homogenous groups (because the researcher cannot know if the test would yield the same results for a different group).

Reliability can also be stated as the*standard error of measurement*.  This is an *estimate* of how often there are errors in a given test score.  If a test were given to the same person over and over, there would be some variability of answers.  If the variability is small, the test is reliable.   The standard error of measurement is determined by a formula (discussed next week).

There are five types of reliability:

* *Stability* - the degree of reliability over time
* *Equivalence* - the degree to which two forms of the same test produce similar results
* *Equivalence and stability* - the degree to which two forms of the same test, administered at different times, produce similar results
* *Internal consistency* - the degree to which test items in the same test are consistent among themselves and with the whole test
* *Scorer-rater* - the degree to which different scorers of tests are consistent or a single scorer is consistent over time

Note that a valid test is always reliable, but a reliable test is not always valid.  In other words, a test that measures what it is supposed to be measuring (valid), it will be consistent (reliable).  But a consistent test (reliable) may be consistently measuring the wrong thing (invalid)!

***TOPIC 2 TASKS***

NONE ☺

**TOPIC 3**

|  |
| --- |
| **\*\*Essential Questions\*\*  How are quantitative data counted and categorized?  What are measures of central tendency?  What are measures of variability?  What are measures of relative position?  What are measures of relationship?** |

Working with quantitative data requires being able to translate concepts into numbers.  The researcher then crunches the numbers, records the results, and translates the information back into text.  In order to follow the researcher through the process, one first must understand the system used.

In quantitative research, data are counted and categorized, so that differences, correlations, or cause-effect relationships can be measured.  The words *counted, categorized,*and *measured* imply that *numbers*are used.  Some data, such as height, weight, age, or how many or how much of a characteristic or phenomenon exists are naturally expressed as numbers.  But some concepts, such as attitude, intelligence, personality, motivation, and achievement, must be *operationally defined* in such as way that they can be counted. These concepts are called *constructs*.  

Attitude, for example, is a construct because it is not naturally expressed in numbers.  But attitude could be defined for the purposes of a research study as having certain numeric values, say, for example, five, defined as *strongly like, like, no opinion, dislike,* and *strongly dislike*. Each of the values could then be assigned a score:   1 = strongly dislike, 2 = dislike, 3 = no opinion, 4 = like, and 5 = strongly like.  Attitude could thus be tabulated in numeric form.

Here is how the idea of operationally defining constructs works in the context of a research study.  Let's say we want to study the *relationship between aptitude and attitude toward school*. Attitude, as discussed above, could be defined with five values. Aptitude could also be defined as having a certain number of values; let's say, three:  1 = low aptitude, 2 = average aptitude, and 3 = high aptitude.  If the hypothesis was stated as a *positive hypothesis*, like this, "Students with high aptitude have a better attitude toward school," what would we expect the numeric results to be?

Well, if the hypothesis right, a participant with a high number on the aptitude scale would also have a high number on the attitude scale.  If there is a relationship between aptitude and attitude, the numbers should reveal it.

Here is some terminology to remember:

* *Variable -*a concept that can assume any one of a range of values, scores, or points on a numeric scale. (In the example above, *attitude* is a *variable* with five possible values.)
* *Construct* - a concept that is not naturally expressed as a number, but can be operationally defined in terms of numbers in order to function as a variable.  (Both *attitude* and *aptitude* are constructs.)
* *Statistics* - number crunching in research; that is, the way quantitative data is analyzed and interpreted (discussed below).

The process for all quantitative research studies includes collecting *data*, information used as evidence to support or oppose the hypothesis or research question.  Data are collected through identification, manipulation, and examination of variables***.***

There are four types of variables or, in research terms, *levels of measurement*: ***nominal, ordinal, interval,*** and ***ratio,*** defined as follows:

* *Nominal*or categorical variables simply categorize data and are thus *qualitative variables*.  Gender, for example, is a nominal variable with two categories, male and female, that can be expressed as numbers (1=male and 2=female, or if you prefer, 1=female and 2=male).  Students can be classified in nominal categories (1=first grade, 2=second grade, 3=third grade, etc.)  Schools can be classified as nominal categories (1=public, 2=private, 3=charter, 4=homeschool, etc.).  Colors, clothing, book titles, college majors can be classified in nominal categories.  Note that the categories are just arbitrary labels or codes, not expressions of value (1 is not *better* than 2, or vice versa).
* *Ordinal* variables both classify and rank data, and are *quantitative variables* because they are measured on a mathematic or numeric scale.  The categories are listed from highest to lowest (or lowest to highest).  Students may be ranked according to GPA, with the highest GPA in the first position and the lowest GPA in the last position.  Salaries may be ranked from highest paid to lowest paid employees.  Hotels may be ranked from five stars to one.  Each ordinal variable can be placed in a particular position in relation to the other ordinal variables, so the variables all fall into a line.  Note that the distance between the variables are of no standard size.  Comparisons of *greater than*and*less than* can be made, but how much more (or less) is unanswerable.
* *Interval*variables classify data, rank data, and also have equal intervals between each variable.  They are *quantitative variables*.  Dates, temperature, test scores may be interval variables, as are most units, if each unit is the same.  There may be an arbitrary zero point in the scale, as well as negative numbers.  Some ratios of differences may occur, such as one number being half as large as another. *Averages* may be calculated among interval variables.
* *Ratio* variables classify, rank, have equal intervals, and have a true *zero point*. They are *quantitative variables*.  In mathematical terms, one ratio variable can be described as a fraction or percentage of another ratio variable. Ratio variables are used for physical measurements, such as height, speed, mass, distance.  Age and time (if a starting time of zero is included) may also be ratio variables.

There may be confusion about the difference between interval and ratio variables.  Test scores, for example, may be either, if the answers are categorized (1=right, 2=wrong, for example), they are ranked (1=one correct answer, 2=two correct answers, 3=three correct answers, etc.), they have equal intervals (the difference between one correct answer and two correct answers is 1, and the difference between two correct answers and three correct answers is 1, etc.); they *may* also have a true zero point (so one score may be 1/3 or five times or 45% of another score).  The real difference is that all statistical measurements can be used with ratio variables (discussed below).

Note that variables may fall into different categories according to how they are used in a particular study.  Height may be a nominal variable if the study only addresses tall and short; but it may be a ratio variable if the study addresses one person's height as a fraction of another person's height (or the same person's height at different times).  It is important to identify the intent of the researcher as well as the way variables are used to determine the type of variable.

Variables are also classified as ***independent*** or ***dependent.***  As discussed in Week 3, an independent variable is the *strategy*, the *treatment*, or the *cause* introduced or examined by the researcher.  The dependent variable is the outcome or effect; it may also be called the post-test.

Finally, variables are also classified as ***qualitative*** or ***quantitative***.   As listed above, *nominal* variables are qualitative variables.  *Ordinal, interval,* and*ratio* variables are quantitative variables.

Researchers need to be able to describe how all the participant responses look as a group - is it a normal distribution of scores, or is it skewed in some way - and also where any one participant's response is situated in relation to all the others.  Four different measures are used to find this information:  *measures of central tendency, measures of variability, measures of relative position,*and*measures of relationship*.  We will practice each in turn.

***Measures of central tendency*** locate the *typical* or *average* score among a group of scores.  A different measure of central tendency is appropriate for each level of measurement (nominal, ordinal, interval, and ratio variables).  We will use the following data set (set of scores) for all the examples:

|  |
| --- |
| Scores:  86   88   89   92   94   96   97   98  100  100 |

* The ***mode*** simply identifies the score that occurs most frequently.   In the data set above, the mode is easily observable to be 100, because 100 occurs more frequently than any other score.
* The ***median*** is the midpoint of the series of scores.  To find the midpoint, simply line up the scores from low to high (or from high to low) and find the middle score.  If there are an *odd* number of scores, the median is the score in the middle of the list.  If there is an *even* number of scores, the median is the point halfway between the two middle numbers.  The data set above has an even number of scores, and the two middle scores are 94 and 96.  The halfway point between 94 and 96 is 95, so the median is 95.
* The ***mean***is the arithmetic average of the scores.  The mean is found by adding the scores together then dividing the total by the number of scores.  In the data set above, the scores added together (86+88+89+92+94+96+97+98+100+100) equal 940.  That total (940), divided by the number of scores (10), equals 94.  The mean is 94.
* The fourth measure of central tendency, ***geometric mean***, will not be discussed in this course.

Which measure of central tendency is the best to use?  The answer, as in most aspects of research, depends upon the purpose.  Let's say, for example, that the data set above represents salaries in thousands of dollars. The *typical*salary could be reported as three different numbers:  $100,000 (if the mode were used), $95,000 (if the median were used), or $94,000 (if the mean were used).  Consider for yourself:  Which is the most *accurate* measure of a typical salary in this case?  Which figure would *management* want to use in a salary negotiation?  Which figure would *employees* want to use in a salary negotiation?   Hmmm. do you see why you need to know about the different measures of central tendency?

A researcher must be careful to use the most appropriate measure of central tendency for the data collected and the type of variable used.  As it happens, the *median* is the most frequently used measure of central tendency.  Here is the complete list of possibilities from which to choose in a given research study:

* Nominal variables - *mode* only
* Ordinal variables - *mode* and *median*
* Interval variables - *mode, median, mean*
* Ratio variables - *mode, median, mean, geometrical mean*

***Measures of variability*** show the spread of a group of scores.  Scores that are close together have little variability; scores that are widely spread out have greater variability.  We will use the same data set as above to determine variability using four different methods.

|  |
| --- |
| Scores:  86   88   89   92   94   96   97   98  100  100 |

* The***range*** is the difference between the highest score and the lowest score.  In our data set, the highest score is 100 and the lowest score is 86.  The difference between the two scores (100 - 86) is 14.   The range of scores, then, is 14.  The range is useful to get a quick estimate of how spread out a group of scores are.
* The ***quartile deviation*** shows the spread of scores around the *median*.  It is found by dividing the scores into quartiles (four equal sections), finding the difference between the upper quartile (the 75th percentile, the point below which are 75% of the scores) and the lower quartile (the 25th percentile, the point below which are 25% of the scores) and dividing the difference in half.  This is not difficult, taken step by step.

First, find the median (we already know it is 95).  Next, find the median of the scores above 95 (98); this is the upper quartile.  Then, find the median of the scores below 95 (89); this is the lower quartile.  The difference between the upper quartile (98) and the lower quartile (89) is 9, and half of that is 4.5.  The quartile deviation, then, is 4.5.  From these calculations, we know that half the distribution of scores, centered on the median, is between 89 and 98.  A small quartile deviation means the scores are relatively close together; a large quartile deviation means the scores are more spread out.  The quartile deviation is useful because it ignores extremely high or extremely low scores that may unfairly affect true variability and looks at only the middle scores to see how spread out they are.  Quartile deviation is an *ordinal* statistic and is most often used in conjunction with the *median*.

* The ***variance*** calculates the amount of spread among scores in a data set, when the *mean* is used as the measurement of central tendency.  Finding the variance requires a series of mathematical steps:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Find the mean of the scores | Find the difference from the mean (94) for each score | Square each difference | Find the sum of the squares | Divide by the number of scores |
| 86             88             89             92             94             96             97             98           100        + 100         -------           940     ÷10 = 94 | 86 - 94 = -8       88 - 94 = -6       89 - 94 = -5       92 - 94 = -2       94 - 94 =  0       96 - 94 =  2       97 - 94 =  3       98 - 94 =  4     100 - 94 =  6     100 - 94 =  6 | -8 x -8 = 64      -6 x -6 = 36      -5 x -5 = 25      -2 x -2 =   4       0 x  0 =   0       2 x  2 =   4       3 x  3 =   9       4 x  4 = 16       6 x  6 = 36       6 x  6 = 36 | 64           36           25             4             0             4             6           16              36        + 36        ------         223 | 223 ÷ 10 = 22.3    The variance is 22.3 |

The variance is not generally used by itself.  It is used to find the *standard deviation*.

* The ***standard deviation*** is the most common measure of variability, used with*interval*and*ratio* variables in conjunction with the *mean*.  The standard deviation is found by taking the square root of the variance.  In our data set, the variance is 22.3, so the standard deviation is 4.72.  A small standard deviation means the scores are relatively close together; a large standard deviation means the scores are relatively spread out.

How are measures of variability used to describe data?  Let's again suppose that the scores in our data set represent salaries in thousands of dollars.  We could then describe the salaries this way:  *there is a range of $14,000 between the highest paid and the lowest paid employee; the middle 50% of the salaries are between $89,000 and $98,000; and the standard deviation shows that.*..wait, we have a problem....  
  
In order for standard deviation to work, the scores have to be normally distributed, forming a symmetrical bell-shaped curve, like this:  
  
***Normal distribution***  
 There are four characteristics of the normal, bell-shaped curve:

* The mode, the median, and the mean are the same.
* Half the scores are above the mean and half are below the mean.
* Most scores are close to the mean, and the farther from the mean, the fewer scores there are.
* The same number or percentage of scores fall between the mean and plus one standard deviation as fall between the mean and minus one standard deviation.

We know right away that our salaries will not form a bell-shaped curve, because the mode, the median, and the mean are not the same.  (Remember, the mode is $100,000, the median is $95,000, and the mean is $94,000.)  So what will the distribution of our salaries look like?

It will be skewed in one direction.

Negatively skewed distribution                                      Positively skewed distribution

* If the *mean* is the smallest of the measurements of central tendency, the *median*is the middle measurement, and the *mode* is the largest measurement, the curve will be skewed *negatively*.  That means more of the scores are located at the higher end of the distribution.
* If the mode is the smallest of the measurements, the *median* is the middle measurement, and the *mean* is the largest measurement, the curve will be skewed *positively*.  That means more of the scores are located at the lower end of the distribution.

In our data set, the mean is the smallest, the median is in the middle, and the mode is the largest ($94,000<$95,000<$100,000), so our distribution is skewed negatively.  So, in our description of the data...

*There is a range of $14,000 between the highest paid and the lowest paid employee; the middle 50% of the salaries are between $89,000 and $98,000; and there are more salaries at the higher end of the scale than at the lower end.*(That makes sense, logically.  There are two incidents of $100,000 salaries, the highest in our data set.)

Are you beginning to see how all these measurements work together to provide useful information?

Note:  Graphics courtesy of CliffsNotes.com. *Measures of Central Tendency*. Retrieved 5/6/08 from <http://www.cliffsnotes.com/WileyCDA/CliffsReviewTopic/topicArticleId-25951,articleId-25905.html>.

***Measures of relative position*** convert scores on different tests or instruments into a *common scale*, so they can be accurately compared (apples to apples instead of apples to oranges). Consider, for example, this scenario:  a student takes a standardized test in reading and scores 60 correct out of 140 total.  The same student scores 60 out of 70 on the standardized math test.  If all the teacher and parent receive are the test scores - 60 in reading and 60 in math - how will they know if the student is doing well?  Will they assume the student is performing equally well in reading and math?  Even if they figure out the student got 54% of the answers correct in reading and 86% of the answers correct in math, how will they figure out where the student stands in comparison with other students his age, or at his grade level?  A common scale will allow those questions to be answered.

The two most common measures of relative position are*percentile ranks* and *standard scores*.

* ***Percentile ranks*** are most commonly used to interpret the scores on standardized tests or scales.  A percentile rank refers to the percentage of scores in a given data set that are *lower*than the one score of interest.  For example, if John is 3'1" tall, we know his height but don't know if he is tall or short for his age.  If however, we learn that 95% of the children John's age are shorter than 3'1" tall - meaning that John is at the 95th percentile in height - we know he is TALL.

Likewise, if Jane scores 80 on an achievement test, does that mean she did well or poorly?  We can't know without some context.)  If we learn that a score of 80 falls at the 30th percentile, we know that 30% of the students taking the assessment scored lower than Jane.  Jane's score, then, is not very high.

We must compare apples with apples, percentile ranks with percentile ranks - *not* raw scores on one test with raw scores on a different test.

There is a simple formula for converting raw scores to percentile ranks: count the number of scores lower than the score of interest, divide that number by the total number of scores, then multiply by 100.

Let's say there were 522 scores, and 447 of them were lower than Jane's score of 80.  At what percentile is Jane's score?  The calculation would be:  447 ÷ 522 =  .8563 x 100 = 85.63. Jane's score, then, was at about the 86th percentile.

Percentile ranks are used for *ordinal* and *interval* data in conjunction with *medians*.  (Remember that the median is the 50th percentile.)

* ***Standard scores*** convert raw scores into a common scale to show the position of a particular score in relation to some reference point, typically the *mean*.  The distance of the particular score from the mean is expressed in units of standard deviations.   Standard scores can be used with data using*interval*or *ratio* scales of measurement, and are accurate when the distribution is *normal* (although there are ways of *normalizing*scores, which is beyond the scope of this course).  The most common standard scores are *z scores*, *t scores*, and *stanines*.

Now that we know the standard deviation for our particular set of scores, we can talk about how far any given score is from the center of the scale.  We can say, for example, that 87 is *one standard deviation below the mean*.   When we refer to scores in that way - that is, in terms of how far they are from the mean in standard deviations - we are using what is called***z scores***

There are some other interesting facts about normal curves:  approximately 68% of scores in a normal distribution will fall between -1 and +1 standard deviation, 95% of scores will fall between -1.96 and +1.96 standard deviations, and 99% of scores will fall between -2.58 and +2.58 standard deviations.  (Other mathematical operations can also be performed on standard scores that are beyond the scope of this course. )  Note that standard deviation is identified by the sigma symbol: σ  when referring to a *population* rather than a *sample*.

Although researchers like to talk in terms of standard deviations, parents and teachers may prefer not to refer to a student's score as "*below the mean*."  So, another translation of standard deviations that is more positively expressed, and more commonly understood, is the *T score*.

The***T score*** (or, just to be confusing, called the*Z score*, with a capital Z) simply converts the mean (the zero *standard deviation*, the zero *z score*) to 50.  Fifty is a common *average* on a scale of 0-100, yes?  All the other scores then become positive: 10, 20, 30, 40, etc.

The formula for converting *z scores* to *T scores* is:   (10 x *z score*) + 50 = *T score.*The conversions would be as follows:

A *z score* of -2 would convert to (10 x -2) + 50 = 30

A *z score*of -1 would convert to (10 x -1) + 50 = 40

And so on.

* ***Stanines*** are another measure of relative position that are often used in schools.  A*stanine* is a standard score that divides the distribution into nine equal parts (stanine actually means *standard nine*).  Stanines are not as exact as other standard scores, but are the common scoring system used for special education placements, standardized tests, and ability grouping.

The formula uses*z scores*, and converts them as follows:  (2 x *z score*) + 5 = *stanine*.  Stanines only use whole numbers, so we must round up or down to the nearest whole number.

* A *z score*of -2 would convert to (2 x -2) + 5 = Stanine 1
* A *z score* of -1.4158 would convert to (2 x -1.4158) + 5 = 2.1684 or Stanine 2
* And so on.

If we did all the math, we would find that half the standard deviations in a normal distribution are in Stanines 2 through 8; the other half are in Stanines 1 and 9.

So, in summary, the score of 85 on the data set we've been using could be referred to in any of the following ways in terms of *relative position*:

* As a score between 1 and 2 *standard deviations* below the mean in our distribution of scores
* As a *z score* of -1.4158 (exactly -1.4158 standard deviations below the mean)
* As a *T score* of 35.842 (with 50 as the average)
* As a score in *Stanine 2*

Why would we care about converting a score of 85 to any of these *measures of relative position*?  Remember the purpose, which is to be able to compare these scores with other scores.  So, for example, if 85 represented a salary in thousands of dollars ($85,000), converting that salary to a *z score* or a *T score* would allow us to compare it with salaries of other companies that were also converted to *z scores* or *T scores*, even if the other companies paid very different salaries.  (Apples to apples, remember?)   Wouldn't that be of value in a salary negotiation?

What other scores would you like to see in apples-to-apples comparisons?  How about baseball batting averages?  How about profits of different businesses?

We also need to talk about ***measures of relationship***, ways in which two variables are related to each other.  The degree of relationship between variables is called the ***correlation coefficient***.  Correlational analysis is essential in *correlational*research studies (that should be obvious), and the type of methods used to calculate the correlation depends upon the the scale of measurement (nominal, ordinal, interval, or ratio).  The types of correlation coefficients are the *Pearson r, Spearman rho, Kendall's tau, biserial, point biserial, tetrachoric, phi coefficient, intraclass*, and *correlation ratio*.   We are not going to discuss the formulas for determining these correlations;  we will only talk about what will be said about them in research reports.

Correlations coefficients are numbers between -1 and +1.  If two variables are high related, the correlation coefficient will be close to -1 or close to +1; if the variables are not related, the coefficient will be close to 0.

Correlations are often illustrated in graphic form, on scatterplots.  This website shows several scatter plots, using the *Pearson r*.  Notice that if two variables are highly correlated, they move together on the plot - both going up (more of one is also more of the other) or both going down (less of one is also less of the other).  Notice the value of*r* (the correlation coefficient) in those plots.  If the variables are not correlated or are less correlated, they are located all over the plot in what appears to be random order.  Here is the website: <http://www.webster.edu/%7Ewoolflm/correlation/correlation.html>

You have made it through the basic explanation of descriptive statistics!  You should now be able to follow along when researchers explain the methods used and what the results mean in terms of variables, constructs, *measures of central tendency, measures of variability, measures of relative position,*and*measures of relationship.* You should supplement your learning through exploration of relevant websites; many professors are posting definitions, graphs, and examples on the web.  Here are a few helpful sites in particular:

Definitions of the symbols used in statistics: <http://www.statistics.com/resources/statsymbols.pdf>

**TOPIC 3 TASKS**

1. Before going any further, we need to use a different data set, one that is normally distributed.  This will give us a chance to practice calculations.   Here is the data set:

|  |
| --- |
| Scores:  85   86   87   92   92   92   97   98   99 |

Do the calculations yourself, to make sure you know how.  Find the *mode*, the *median,* the *mean*, the*variance*, and the *standard deviation*.  When you have all the answers, see what others have to say on the blog! I have started the thread on ***9/9 -9/15.***

**TOPIC 4**

|  |
| --- |
| **\*\*Essential Questions\*\*  What are inferential statistics?  How do researchers generalize samples to populations?  What does "significant difference" mean?  What are tests of significance and what kinds are there?** |

We know that if we want to find out something about a population, we usually are not able to study the entire population.  Instead, we choose one or more samples from the population to study.  If we choose the sample(s) wisely, follow appropriate methodology in conducting the study, and analyze the results correctly and accurately, we should have data that we can relate back to the population.

In this topic we will discuss how well a sample relates to the entire population*.*  We will, in research terms, make *inferences* about the population based on the results of the sample.   This is the key to *generalizability*, and it will tell us if the study results have any *significance.*

In research, the term *significance* is the probability that the treatment, intervention, instruction, test, etc. in a given study made a real and actual difference.  If we do not find *significant*difference, we have to conclude that any differences observed were due to chance.

Let's say, for example, the purpose of a study is to determine if a particular reading program is better than the current one used in our school district.  We choose to pilot the new reading program in one first grade classroom in the district, so we administer a pre-test to the students, try out the strategies in the program for several months, then administer a post-test to the   
same students.  We compare the pre-test scores with the post-test scores and find that most of the post-test scores are higher than the pre-test scores.  We use descriptive statistics to find the means of the two sets of scores, the standard deviations, and even z scores, and we *know*that the post-test scores are better in all measures.

How do we interpret that information?  It could be that the reading program was so good the students progressed far more than they would have with the current reading program.  But, it could be that the students learned just as much as they would have with the current program.  Or, it could be that the students were just having a great day during the post-test.  Or, it could be that the students were simply high achievers who would learn with any program.  How would we know if the differences observed in the post-test scores had anything to do with the reading program? Were the differences *significant* or just *chance*?

Here is another way we could conduct the study.  Let's say we use two classrooms of first grade students, one of which tries out the strategies in the new reading program and the other of which uses the current reading program (a control group) for an equal amount of time.  We pre-test and post-test both groups of students and compare the scores.  We use descriptive statistics and thus *know*the post-test scores of the students using the new program are better than the post-test scores of the students using the current program.

Does that help our interpretation of information?  We know what the results are for our sample first graders, but would the results be the same for the entire population of first graders in our district?  Would the results of first graders all over the country be the same?  How will we know?

The problem is that we can't realistically conduct a study with all first graders to see how our results compare.  Even if we conducted the study with all first graders in our district, we still would not know for sure that the new reading program would be better with *next year's first graders*, or first graders in the *neighboring district*or *anywhere else*.

However, we can answer these questions through mathematical formulas designed to get *estimates* about the population, based on samples.  In fact, we can even get an estimate about the population based on the results of *one single sample*. Those are *inferential statistics*.

Before proceeding to a discussion of the kinds of ***inferential statistics*** used by researchers, it must be understood that any inferences made about a population are only *probabilities*; nothing is proven or absolutely certain.  Researchers therefore make *estimates*when referring to populations.

The larger the sample, or the more samples tested from a given population, the more confident the researcher can be about the accuracy of any inferences or estimates.  Chance variations - called *sampling errors* - occur when, for instance, a particular score is far different than all the others, or when a single participant earns different scores on the same instrument at different times. These sampling errors occur naturally and are not the fault of the researcher, but they have less impact in a large sample than they would in a small sample. So using a large sample size is important to the accuracy of the estimates.

It should also be noted that, when referring to *samples*, the term *statistics* is used.  When referring to *populations*, the term *parameters* is used.

In descriptive statistics, we learned that we must know the *mean* of the scores in order to perform other statistical analyses on the data.  Can we also find the mean of the population?

Yes.  We can assume that each sample would likely have a different mean, but the means taken together will be *normally distributed* around the mean of the population.  (Note:  this is assuming the samples are sufficiently large, of equal size, and all taken from the same population.)  The *mean of the population* can thus be estimated by calculating the average of all the sample means.

The formula is this:  add all the means together and divide by the number of means to find the average.  This will be an estimation of the *mean of the population*.

It may be difficult to estimate the mean of the population with any accuracy.  A researcher will often use only one or two groups and collect perhaps two to four sets of data in a given study.  That would not generate enough *means* of the scores with which to make a good estimate of the mean of the population.   Fortunately, there are formulas that allow the researcher to calculate other measures without first estimating the mean of the population.   These are discussed below.

An estimate of the *standard deviation* of a population can be calculated even if the researcher doesn't know the mean of the population.   In fact, the standard deviation of a population can be estimated from a *single sample*.  This is the formula:   (standard deviation of a sample)  ÷ (square root of [sample size minus one]).  The standard deviation of a population is called the ***standard error of the mean.***

The *standard error*can also be calculated for the measures of central tendency, variability, relationship, and relative position (which are beyond the scope of this course).

The *standard error* can also be found for the *difference between means* as well.  This calculation is important if looking, for example, at a pre-test set of scores and a post-test set of scores in order to determine if the difference between the two samples is *significant* (as discussed below).

As you will remember from the discussion about hypotheses, the hypothesis is the researcher's prediction about what the results of the study will show.  The researcher will often set out to show that a relationship exists between two variables, that one method is better than another, or that a particular behavior was caused by a certain circumstance or characteristic.  The researcher may frame the study in terms of a***null hypothesis***and an ***alternate hypothesis*** (a non-directional or a directional hypothesis).  Why both?

Basically, the answer is that while a researcher cannot *prove* a hypothesis, he can *disprove* a hypothesis.  And a null hypothesis is easier to disprove than other types.

If, in our example above of a study comparing reading programs, we hypothesize there is a difference between the current reading program and the new reading program, and in our study we find that there is a difference in the scores between students using the current reading program and students using the new reading program, what have we *proven*?  Will we know if there is actually a difference in the programs or if we have just found a group of students *by chance* that performed differently with one or the other methods?  We have not *proven*the hypothesis that there is a difference between the two programs, although we may have some indication that such is true.

But if we stated our hypothesis as null - that is, there is no difference in the two programs - and the statistical analysis shows there is a *significant* difference in scores, we have a stronger argument that there is truly a difference in the programs, and we can *reject* the null hypothesis.

Here is another rather silly example:  suppose we hypothesize that "the sun rises in the morning."  When the sun rises tomorrow morning, what have we proven?  We have one instance of the sun rising, but do we know if the sun rises *every*morning? Will the sun rise *the next* morning?

On the other hand, we could propose the null hypothesis that "the sun does not rise in the morning."  Tomorrow morning, the hypothesis will be disproven.  Easily.  And if we need more evidence, we only have to wait until the following morning.

A researcher will often work with both types of hypotheses, a null hypothesis that he hopes to disprove, and an alternate hypothesis stating what he predicts will happen, because together, the two hypotheses make a very good case for whatever the researcher is trying to show.  In addition, both types of hypotheses are required for tests of significance (discussed below).

This is how a null hypothesis is used in a study.  If a null hypothesis is *true*, that means there is *no significant difference* between the two samples, or the two programs, or the independent and dependant variables.  It means that whatever differences may be discovered are either too small to be able to show any relationship or cause-effect or that the differences occurred merely by chance.  If the null hypothesis is *false*, that means the differences discovered in the study *are significant* and did not occur by chance.

There are two kinds of *errors* a researcher could make when determining if a null hypothesis is true or false:

* Type I error - rejecting the null hypothesis when it is true.
* Type II error - failing to reject the null hypothesis when it is false.

An example from medicine may be helpful in illustrating these errors.  Let's say a disease-free body is the normal state (the null hypothesis); if a physician draws blood and *incorrectly* concludes there is disease present, the conclusion is a *false positive* (incorrectly rejecting the null hypothesis, a Type I error)*.*  On the other hand,if a physician draws blood and *incorrectly* concludes there is no disease present, the conclusion is a*false negative* (incorrectly failing to reject the null hypothesis, a Type II error).

An example from the justice system may take us further.  In the United States, all defendants are considered innocent (until proven guilty); we could write the null hypothesis as "the defendant is not guilty."  Let's say a jury hears the evidence and incorrectly determines the defendant is guilty.  That would be a Type I error.  If, on the other hand, a jury hears the evidence and incorrectly determines the defendant is not guilty, the jury has committed a Type II error. Here is the website: <http://www.intuitor.com/statistics/T1T2Errors.html>

The reason for offering the judicial example is as follows:  both Type I and Type II errors cause problems, but the *Type I error is usually the worse*of the two.   Incorrectly convicting an innocent person is a nightmare scenario in the American system, giving rise to the maxim:  it is better to let a hundred guilty men go free than to convict one innocent man.  So we need to avoid Type I errors.

Let's continue the analogy another step.

If the null hypothesis is that a defendant is "not guilty," what would the *alternate hypothesis* be?  From the prosecutor's standpoint, it would be "the defendant is guilty," and that is what the prosecutor sets out toshow.  And what is the *standard* used to determine guilt?  In the American system, the standard is not absolute proof, but proof *beyond a reasonable doubt.*  The null hypothesis (not guilty), then, will be rejected upon proof of guilt at the acceptable standard (beyond a reasonable doubt).

How would a researcher avoid making a Type I error?  There is a standard of proof (just as in the judicial example above) for accepting or rejecting a null hypothesis. It is usually set at 5%, meaning 5% is the level of proof, or, in research terms, the maximum acceptable *probability*that differences observed in scores or means are due to chance.  This is called the ***significance level****,* the*probability level*, or *alpha.*  At a standard set at 5%, the *probability level =* *.05 alpha*.

Looking at a normal distribution, the area in which 95% of the scores are found is between - 2 standard deviations and + 2 standard deviations. That is called the *region of chance*or the *region of acceptance of the null hypothesis*.  The other 5% of scores would be found in the area less than -2 standard deviations or greater than +2 standard deviations; in other words, in the *tails* of the curve.  These are called the *regions of rejection,*and are where we would find the scores in a particular study if the null hypothesis is false.  When the null hypothesis is rejected, researchers say the hypothesis has been rejected *at the alpha level of significance.*

Some introductory studies will use a probability level of .10, meaning that the researchers will accept a lower standard of proof that the null hypothesis is false.  A third common standard is a probability level of .01, which requires a very high level of proof before a null hypothesis may be rejected; this standard is used when the researcher absolutely does not want to commit a Type I error.  The probability level can in fact be set at any number between 0 and 1, and should be determined before the study is conducted, so the researcher is not influenced by the data before choosing the standard of proof.

How would a researcher avoid making a Type II error?   The term for the probability of avoiding a Type II error is***power***.  As the *power* of a test increases, the probability of making a Type II error decreases.  The *power*of a hypothesis test is affected by three factors:

* *sample size* - the larger the sample size, the greater the*power* of the test
* *significance level (alpha)* - the higher the significance level, the smaller the region of acceptance, the greater the *power,*and the more likely the hypothesis will be rejected (of course, this increases the likelihood of a Type I error!)
* *the true value of the parameter* -  the greater the difference between the true value of a parameter and the value stated in the null hypothesis, the greater the power of the test.

We will not go into details about *power* in this course.  Just know, when you will encounter this term in your researcher reports, that increasing *power* is a way of avoiding Type II errors.  
It should be noted that the researcher must balance the likelihood of making a Type I error with the likelihood of making a Type II error when choosing the probability level.  The reason is that as the degree of proof increases for a null hypothesis to be rejected, the danger increases of failing to reject a hypothesis that should be rejected.  In research terms, as the likelihood of making a Type I error diminishes, the likelihood of making a Type II error increases.  The researcher must weigh the consequences of each when deciding what an appropriate probability level will be.

Researchers test hypotheses using ***tests of significance***.    A*test of significance* determines if there is a *significant difference*between the means of samples at the selected *probability level*.   (Note:  we are *not* going to perform any calculations for tests of significance in this course.  There are many sources online if you are interested in learning more about any test.)   
Tests may be***two-tailed*** or ***one-tailed***.  For a null hypothesis or a non-directional hypothesis, a *two-tailed test* will be used, which means that there are *regions of rejection* on both *tails* of the distribution curve.  A two-tailed test acknowledges the possibility that the differences between the means of samples could occur in either tail (e.g., the new reading program could be*better*or *worse* than the current one).  For a directional hypothesis, a *one-tailed test* would be used, meaning that a region of rejection is on only one *tail* of the curve. This assumes that any significant differences will be in the tail at the end stated by the hypothesis (e.g., if the new reading program is *better* than the old one, the scores will be in the tail at the *positive end* of the curve).

Every test of significance has a formula for that test's ***degrees of freedom***.  Degrees of freedom are a measure of how precise an estimate may be, and they decrease by one with each sequential estimation.  Basically, in any data set, each estimation costs one degree of freedom for the next estimation.  Therefore, finding the *mean*of a set of scores costs one degree of freedom. Finding the *standard deviation* from the mean costs another degree of freedom.

Here's an example to illustrate degrees of freedom.  Let's say you have complete freedom to choose five scores.  You wisely decide to keep it simple and choose 2, 4, 6, 8, and 10.  Now suppose a restriction is placed on the last score: instead of choosing any score you want, you must choose the score that makes the *mean* of the scores equal 8.  (Remember how to find the mean?)  Now you do not have complete freedom to choose five scores.  You chose 2, 4, 6, and 8, but the last score can no longer be any of your choice; it will have to be 20 in order for the mean of the scores to equal 8.  You lost one *degree of freedom*.

There are two types of tests of significance: *parametric*tests and *nonparametric* tests.   Factors such as sample selection method, the number of groups, the number of independent variables, and the scale of measurement determine which test is appropriate.***Parametric tests***are the appropriate tests of significance for interval or ratio data that are sampled from a population that is *normally distributed*.  Parametric tests are generally preferable to nonparametric tests, because they have more *power.*

There are two commonly used parametric tests:  the *t test* and the *simple analysis of variance (ANOVA).*Both tests compare means at a selected probability level to see if the means are significantly different.

The differences between the tests are as follows:

* The ***t test***compares*two* means
* Because some difference between the means is expected (due to chance), the*t test* determines whether the actual difference is significantly larger than the chance difference
* The*t test*can be calculated for either independent samples (random samples that are selected without any kind of matching with the other sample, other than they both came from the same population) or non-independent samples (two groups that are matched to each other or one group in a pre-test/post-test situation)
* The formula is a ratio in which the numerator is the difference between the means and the denominator is the standard error of the difference between the means.  (This is a complicated formula not practiced in this course.)
* Fortunately, there is computer software to calculate*t tests*.  There is even a t test calculator online here: <http://www.graphpad.com/quickcalcs/ttest1.cfm>
* The ***ANOVA*** compares two means, but can also be used to compare several means
* The ANOVA requires randomly selected groups
* The ANOVA assumes all the groups were approximately the same at the beginning of the study
* The total variance of scores is divided into 1) variance caused by the treatment or variance between groups and 2) variance within groups, and a ratio called the F ratio is formed, with group differences in the numerator and variance within groups in the denominator (another complicated formula)
* Fortunately, there is computer software to calculate ANOVA.  There are many ANOVA calculators online, including [this](http://faculty.vassar.edu/lowry/vsanova.html).

One of these tests would be helpful in our study comparing two reading programs, if we had randomly selected the participants and if we had two groups, one of which tried to new reading program and the other, a control group, used the current reading program.

There are other parametric tests to mention as well.  The analysis of covariance (ANCOVA) is a form of ANOVA that adjusts post-test scores based on initial differences in a variable such as the pre-test score, sort of like handicapping in golf or bowling.  This levels the playing field, as it were, so that the post-test scores are more likely to be accurate reflections of the treatment administered.  Variables that can be controlled using ANCOVA are IQ and aptitude, as well as pre-test scores.  ANCOVA is best used in true experimental studies with randomly selected participants.

Nonparametric tests are used for data that do *not make assumptions about the distribution*. Nonparametric tests simply rank variables from low to high and then analyze the ranks.

The ***chi square*** test is a nonparametric tests appropriate for nominal variables - those that can only be assigned numbers and then counted for frequency (e.g., male, female or yes, no).  A chi square test compares the frequency of variables that actually appear in a study to see if they are significantly different from the expected proportions if the groups were equal (or based on past data).   A chi square test can determine significant differences between such variables as how many college students voted in the last election compared to how many were eligible to vote (and thus expected to vote), how many students take online courses compared to the number of students who live 30 or more miles away from campus (and thus expected to choose online learning), and so on.  These are all called *one-dimensional chi square.*

A *two-dimensional chi square* is used when the variables are categorized in more than one way.  For example, the number of students who take online courses can be categorized by male and female as well as by distance from campus to see if there is a significance difference between the observed frequency and the expected frequency.

Chi squares are fairly easy to calculate, but software and the internet are quicker.  Try this site for more explanations and calculators: <http://home.clara.net/sisa/two2hlp.htm>

One increasingly common predictor test is the ***multiple regression equation***.  This test uses variables that are already known to predict the outcome of a particular study with the new variables explored in the study to result in a more accurate prediction.  For example, multiple regression can predict success in college using SAT scores, high school GPA, high school class rank, and college freshman GPA.  It will not only determine which of the variables were predictors of success in college, but which of the variables were *better* predictors and by how much.   Multiple regression may be used with any scale of measurement and any kind of quantitative study.  It can also determine not only if variables are related but the degree to which they are related.   

Here are summaries of some of the common statistical tests and when they are appropriate to use:

|  |  |  |  |
| --- | --- | --- | --- |
|  | This test determines if... | Requirements | Examples:  When we want to test... |
| one-sample t-test (parametric) | a sample mean significantly differs from a hypothesized value | interval or ratio variable normally distributed | if the average math score (the mean) in our class differs significantly from the Oregon benchmark score of 142. |
| paired t-test (parametric) | the means of two related observations are significantly different | interval or ratio variables normally distributed | if the average math score (the mean) in our class is significantly different from the average reading score (the mean) |
| Wilcoxon signed rank sum test (non-parametric) | same as above | non-parametric version of paired t-test does not assume normally distributed does not assume interval variables | same as above |
| one-sample binomial test (non-parametric) | the proportion of responses on a two-level categorical dependent variable significantly differs from a hypothesized value | categorical data two levels (e.g., male, female; short, tall) | if the proportion of females buying books at Aspen Bookstore in December differs significantly from 50% |
| two independent samples t-test (parametric) | the mean of a dependent variable of one group significantly differs from the same variable of another group | groups independent interval variable normally distributed curve | if the average math score (the mean) in our class is significantly different for males and females |
| Wilcoxon-Mann- Whitney test (non-parametric) | same as above | does not assume the dependent variable is normally distributed | same as above |
| Chi-Square goodness of fit (non-parametric) | the observed proportions for a categorical variable significantly differ from hypothesized proportions | categorical variables | if the racial composition (by percentage of each) of OSU students differs significantly from the racial composition (by percentage of each) of Oregonians |
| Chi-square test (non-parametric) | there is a relationship between two categorical variables | categorical variables must have at least five responses in each category | if there is a relationship between the kind of college attended and gender |
| Fisher's exact test (non-parametric) | same as above | can have a frequency of less than five | same as above |
| One-way ANOVA (parametric) | there are significant differences in the means of a dependent variable and one or more categorical variables | categorical independent variable interval dependent variables normally distributed curve | if the average math score (the mean) in our class is significantly different for TAG, regular ed, and special ed programs |
| One-way MANOVA (parametric) | same as above, but with two or more dependent variables | same as above | if the average math score, reading score, and writing scores are significantly different for TAG, regular ed, and special ed programs |
| ANCOVA (parametric) | there are significant differences in the means of a dependent variable and one or more categorical or continuous variables | normally distributed curve | if the average math score in our class is still significantly different for TAG, regular ed, and special ed students, even after adjusting for reading scores |
| Kruskal Wallis test (non-parametric) | same as above | non-parametric version of ANOVA does not assume normally distributed curve | same as above |
| Friedman test (non-parametric) | there are significant differences in the ranks of two or more independent variables in the same subjects | dependent variable is not interval and/or normally distributed | if the distribution of the ranks of reading, writing, and math scores in our class are significantly different |
| simple linear regression (parametric) | there is a relationship between one independent (predictor) variable and one dependent (outcome) variable | both variables normally distributed | if there is a linear relationship between math scores and reading scores in our class; can we predict math scores from reading scores? |
| Spearman correlation (non-parametric) | there is a correlations between two variables | does not assume variables are normally distributed or interval variables are converted in ranks | if there is a relationship between math and reading scores in our class |
| Simple logistic regression (non-parametric) | independent variable predicts the dependent variable | outcome variable is binary (0 or 1) predictor variable cannot be categorical | if reading scores are a significant predictor of high school graduation |
| Multiple logistic regression (non-parametric) | same as above, with more than one predictor | outcome variable is binary (0 or 1) predictor variables cannot be categorical | if reading scores, math scores, and writing scores are significant preditors of high school graduation |

What does it all mean?

Remember, inferential statistics are the method of making inferences about a population based on a sample.  Two factors must be present:  the sample must be representative of the population and the assumptions of the statistical test must be met.    Other factors must be addressed if a study is to be valid:

* the measurements must be error-free, to the largest extent possible
* the test of significance must avoid a Type II error (have high statistical *power*) through large sample size, high significance level, and significant departure from the null hypothesis
* the number of comparisons should be minimized in order to avoid a Type I error

When a researcher rejects a null hypothesis, it means the alternate hypothesis (what the researcher probably wanted to predict) is supported, but it is not proven.

When, on the other hand, a study does not reject a null hypothesis, this simply means there was no verifiable statistical difference between or among the variables studied.  This is also good information.

It should be noted that there may be a difference between statistical significance and practical significance.   Say, in our reading program study, we do not find a statistical difference in post-test scores between the current reading program and the new reading program.  If that were the case, the null hypothesis would not be rejected and we would not be able to say that one program was better than the other.  But along the way, we may find that the scores of a small group of first grade students who did not know letters or sounds when they entered school made tremendous gains with the new reading program.  Would that be *practically* significant?  It would certainly warrant more study.

And finally, one of the strengths of using hypothesis research is the possibility of *replication*.   Conducting the same study with different participants or using a different approach to the same hypothesis increases the generalizability of the results, when the results of two studies are the same.   Or, the replication could show some fallacy or weakness in the first study, such as a different outcome when a variable is introduced in a natural setting as opposed to a controlled setting.

***TOPIC 4 TASKS***

OK. You just read a bunch of stuff and discussed it with your colleagues, and much of this is still swimming in your head. I know that, and I don’t expect you to become an expert in a couple of weeks! So, let’s try it out and see where we are. There are two primary tasks, but I think we will have a lively discussion on the blog and in class.

***1.*** Your first task is to review the research you are beginning to compile for your research presentation. Find any that use quantitative methods.  (You will need one (1) for this assignment, so if your portfolio is lacking, you will have to conduct a further search of the literature for more studies – or see me for alternate arrangements.)  For each of at least ***one (1)*** research report (you can do more if you want ☺), write a summary ***paragraph***. Don’t copy the abstract. That would be bad! Some things to think about when you are writing your summary:

1. the problem in the study
2. the purpose of the research
3. the research method and why you think that method was appropriate for the kind of information desired
4. the sample used and how it was determined (was it biased?)
5. the types of variables defined and measures used in the study
6. the types of descriptive statistics used (if there aren’t any what do you think about the validity of the article?)
7. which ones use inferential statistics.  Write a sentence of two about the generalizations to the population, the estimates, and the tests of significance
8. Use APA style for each full citation, and place your summary as an indented, single-spaced paragraph under the citation.

***2.*** Second, describe a quantitative research study in your topic area that you might want to conduct.  Be practical.  Be realistic.  Include all the components discussed so far, such as a null hypothesis, an alternate hypothesis, the variables, the data collection method, the kinds of descriptive statistics that might be appropriate, and the method by which the results might be generalized.  Keep it short and to the point!

Email me your article summary(ies) and study description by ***9/15***. My email address is [ttwyman@pacificu.edu](mailto:ttwyman@pacificu.edu).