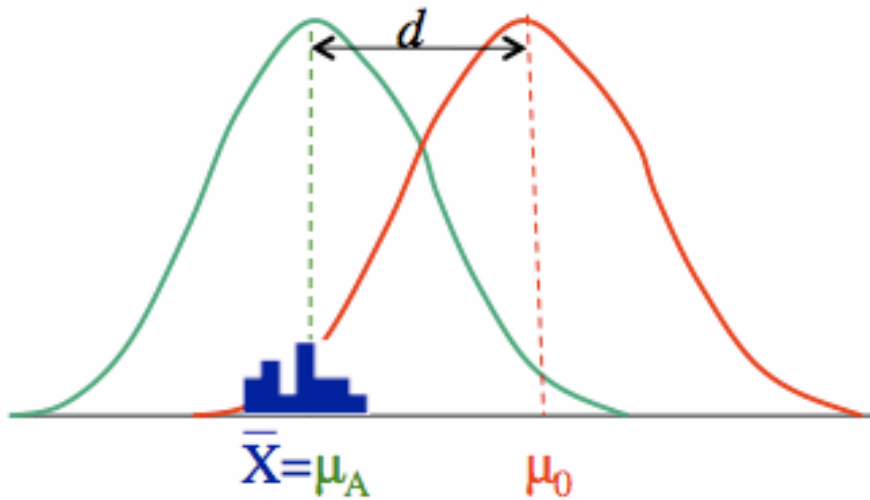


PSY 138
Reasoning in Psychology Using Statistics



Reading and Assignment Packet

Revised for Spring 2017

Dr. J. Cooper Cutting

To the Student

You are about to begin a learning experience especially created for this course. This text is just one component. It provides an introduction to topics covered day-by-day in the course. I will take advantage of PowerPoint technology in the lecture class to lead you through complex sequences of ideas and mathematical formulas. I assume that you have read each chapter before the lecture and lab classes, so I review the material quickly.

In the computer laboratory you will learn to use Excel and SPSS, software that calculates statistics, runs statistical tests, and produces graphs. Available on ReggieNet are Lab Texts to guide you and, during lab time, Lab Exercises for you to complete. Your laboratory instructor is a graduate student with expertise in statistical software. If you find it difficult to complete labs in the hour scheduled, you should read beforehand the Lab Texts, which are always available on the course ReggieNet site. *It is especially important during the early labs to learn and practice Excel and SPSS.* For many of you, these spreadsheet programs are new, while the class material during the first part of the course is review.

All written information in the course is available in this text and on ReggieNet. Included there, in addition to the lab materials noted above, are the course syllabus, PowerPoint slides posted after each lecture, quizzes available when scheduled, a running tally of your grades, and overall class statistics.

We will go through formulas in detail so that you will understand them. In practice, you will use a calculator, Excel and/or SPSS to do your calculations. Most formulas are provided on tests, but you need to know when and how to use them. They are summarized at the end of this text.

Mathematical material is best learned day-by-day. You need to consolidate one set of ideas to be ready to learn the next set. Follow this plan and you will be successful. I know from experience that students who think they don't need to are often not successful.

Dr. J. Cooper Cutting

Table of Contents

1: Introduction	4
I. Producing Data	5
2: Data Basics & Measurement	6
Practice Set 1	12
3: Experiments	14
Practice Set 2	19
4: Sampling & Basic Probability	20
Practice Set 3	24
II. Describing Data	25
5: Displaying Distributions	26
Practice Set 4	36
6: Central Tendency	37
Practice Set 5	41
7: <i>Study Guide for Exam 1</i>	42
8: Variability	44
Practice Set 6	49
9: z-scores	50
10: Normal Distribution	55
Practice Set 7	60
11: Scatterplots & Correlations	61
12: Correlations, continued	67
Practice Set 8	72
13: <i>Study Guide for Exam 2</i>	73
III. Drawing Conclusions about Group Differences	76
Decision Tree for Hypothesis Testing	77
14: Hypothesis Testing in General	78
15: Hypothesis Testing with z-tests	87
Practice Set 9	93
16: One-Sample t-test	94
Practice Set 10	98
17: Related-Samples t-test	99
Practice Set 11	102
18: Independent-Samples t-test	103
Practice Set 12	107
19: <i>Study Guide for Exam 3</i>	108
IV. Drawing Conclusions about Relationships between Variables and about Population Parameters	115
20: Hypothesis Testing with Correlation	116
21: Regression	119

Practice Set 13	123
22: Chi-Square	124
Practice Set 14	128
23: Estimation of Population Means	129
24: Estimation Combined with Hypothesis Testing	132
Practice Set 15	137
25: <i>Study Guide for Exam 4 and Final Exam</i>	138
V. Statistical Tables (z, t, r, Chi-square)	141
The Unit Normal Table	141
The t Distribution	143
Critical Pearson r Values	144
Critical Values of the Chi Square Distribution	145
VI. Solutions to Practice Problem Sets	146
1	146
2	148
3	149
4	150
5	151
Exam 1	152
6	153
7	154
8	155
Exam 2	157
9	159
10	160
11	162
12	164
Exam 3	166
13	172
14	173
15	174
Exam 4 and Final Exam	175
VII. Summary of Formulas	178

Chapter 1: Introduction



Statistical methods are critical tools used in almost all scientific research. As such, gaining a basic understanding of statistical methods and reasoning is essential to both conducting and understanding research findings. However, a good understanding of basic statistical procedures isn't restricted to scientists. You may not realize it, but a good understanding of basic statistics is also extremely useful in one's everyday life as well. As an exercise, go online for a newspaper (or pick one up) or watch the local or national news on television. You'll find statistical reports throughout; you will be better at recognizing and evaluating them after this course.

Statistics are procedures that are used to summarize and sets of data. Data are numbers within a **context**. For example, consider the number 7. By itself it is an abstraction. However, when considered within a context it takes on specific meaning. It could represent the number of days that you study for an exam, or the score on an exam, or the number of questions missed on an exam, or your rank order placement on the exam. So it is the context associated with the number that gives the number an interpretable meaning. So, while this course involves abstract manipulation of numbers, it is also concerned with the context associated with the numbers.

This course is broken into four basic parts:

Producing Data covers basic research methods that are used to collect data. That is, we will discuss the context of the numbers, where the numbers come from, what they are tied to, and how we got them. This section is necessary because it is critical to understand what the data are (what the numbers mean) and how they were collected in order to correctly analyze and interpret what they mean.

Describing Data covers basic methods that are used to summarize and simplify sets of data. These include both graphic and numerical summaries. Statistics that are covered describe central tendency, variability, and correlation of two variables.

Drawing Conclusions about Group Differences describes procedures used for hypothesis testing. We use data drawn from samples and make conclusions about entire populations from which they are drawn. This section focuses on z-tests and t-tests for deciding which of two populations a sample most likely belongs to.

Drawing Conclusions about Relationships between Variables and about Population Parameters continues to describe procedures used for hypothesis testing. This section focuses on correlation, regression, and chi-square tests, which enable us to decide whether two variables co-vary. In addition, the section covers estimation of population means.

Section I: Producing Data



In order to evaluate conclusions made from data, we must first consider the context associated with the numbers. An important part of this context is how the data were produced. A complete understanding of where the numbers came from will help guide our interpretations of the data (i.e., what they mean). Therefore, we must begin our discussion with some basic concepts concerning how data are produced.

Stop for a moment and consider some facts that you know. Now consider how you know what you know. Was this piece of knowledge something that you were told or that you read in a book? Is it based on something that you just believe in your gut? Or is it something that you yourself have observed? Have you observed it just once or have you seen it happen over and over? Does it happen the same way each time, or somewhat differently each time? How systematic were your observations?

We'll start by considering the scientific method, the method used to produce data in the sciences and social sciences. The scientific method is a set of procedures that outlines different methods for making systematic observations of different situations (be they asteroids impacting planets, atoms colliding together, or people interacting in a room). We'll see that different methods allow for different kinds of conclusions to be drawn.

In this section we'll also discuss how we measure characteristics (which we'll call **variables**) of these different situations. Within a particular situation, what we're really interested in doing is describing the potential relationships between the different variables present.

We'll also discuss where the observations are made. That is, if we're interested in examining people interacting in a room, how do we decide which people to examine? Do we wander around campus with a video camera? Do we put an ad in the local paper? What sampling methods are considered "best," and why?

Another thing that we will consider in this section of the course is some basics of probability theory. The nature of knowledge is one of uncertainty. Conclusions are made as our best guesses based on the data that we have, with a known chance that the conclusions are wrong. We'll focus on why this is and on how we know what the level of chance is.

Chapter 2: Data Basics & Measurement

Scientific Method

The **scientific method** is a way of knowing facts or knowledge. It is one of the ways of knowing that we have available to us. Some of the other ways of knowing things include: tenacity (i.e., knowing something is true because the fact is stored in our memory), common sense (i.e., we know something is true because our understanding of the way things work in the world is consistent with that knowledge), reason or logic (i.e., we know something is true because logically it follows from other facts that we know are true), and authority (i.e., we know something is true because someone with authority or experience told us it was true). Many of the things you are learning as a student, you are learning through authority (i.e., your professors tell you it's true), but we also hope that you are using logic and/or the scientific method to question what you are told by us and other sources of authority (like the media). That's one thing we're hoping you'll get out of this class: methods for evaluating facts given to you by persons of authority (like your professors and the media).

In fact, early scientists did rely on authority for their scientific knowledge. For example, if someone asked an early scientist if a feather or a stone would reach the ground first when thrown off the roof of a building, the scientist would refer to Aristotle's theories on matter to answer the question. When Galileo became a scientist, however, he questioned the reliance on authority for scientific knowledge. Galileo's method of answering the question about the feather and the stone would be to observe the answer by throwing both objects off the roof and observing which one hit the ground first. Galileo changed the way scientists gained knowledge. Since that time scientists have used observations to answer questions.

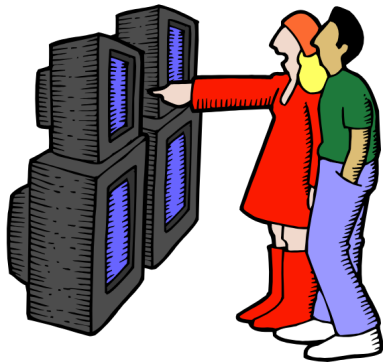


Variables

Our starting place for using the scientific method and making observations is a question we want to answer. The scientific method and statistics are just tools for helping us answer questions and learning new knowledge. Once we decide what we're trying to learn about, we must consider our **variables** of interest. A **variable** is a characteristic or condition that may change from one person to the next. In other words, variables describe different information about **individuals**. Variables can be defined according to the values they take. A **continuous variable** is a variable that can take any number and can infinitely be broken down into smaller and smaller measurements. Time is a continuous variable, because it involves a numerical measurement with an infinite number of smaller categories. It can be measured as a number in years, days, hours, minutes, seconds, etc. A **discrete variable**, on the other hand, involves a finite number of categories. Mood measured by the categories happy, sad, angry, and excited is a discrete variable because there are only four possible responses that can be made. There are no

other responses allowed. Number of people taking PSY 138 this semester is also a discrete variable, the unit of one person cannot be further subdivided. You cannot count the variable in any unit smaller than a whole person.

One important variable in making our observations is the one that we are observing or measuring, called the **response variable** or **dependent variable**. This dependent variable will give us the observations of **data** that we will use to answer the question we're interested in.



Suppose for example that we are interested in the question: Do students with a high GPA watch less TV than students with low GPAs? There are two variables defined in this question: (1) GPA (high or low) and (2) amount of TV watched. In this case we will observe both of these variables from students so they will both be dependent variables in the study we will conduct to answer the question.

Sometimes the variables we're interested in are fairly abstract and we need to decide how to define them for our study. For example, suppose that our question above had been: Do students who are high academic achievers watch less TV than students who are low academic achievers? Academic achievement is an abstract variable that can be defined in several ways: GPA, ACT score, score on final exams, etc. In order to study academic achievement, we must first **operationally define** it so that we can measure it in our study. If we choose GPA as our measure that means that GPA is our operational definition of academic achievement.

Research Design Types

We have a couple of research design types that we can use when making our observations. Our choice of research design is dependent on the kind of question we are asking and helps us decide what kind of variables we will have in our study.

One type of design is the **observational** method or **correlational** study. The goal of an observational study is to look for a relationship between two or more dependent variables that we've measured from a group of individuals. The observational method would be appropriate for our example question in Part B, because we are interested in how GPA that we are measuring is related to amount of TV watched, which we are also measuring. To answer the question, we need to see if GPA and amount of TV watched are related in a particular way, specifically, if students with high GPAs tend to have low scores on the amount of TV watched measure. *Note that we observe only; we do not manipulate anything.*

Another design we can choose to use is an **experiment**. The goal of an experiment is to determine if one or more variables **causes** people to have high or low scores on a dependent variable. In our example above, we would have chosen an experiment if our

question had been: Does watching a lot of TV **cause** a student to have a low GPA? To answer this question with an experiment, we'd still have the same two variables: (1) GPA (high or low) and (2) amount of TV watched. GPA would still be a dependent variable we'd measure from the students. But we're also trying to determine causation so we have to **manipulate** the amount of TV the students watch, not measure it. This means that we'll randomly assign the students to groups, where each group watches a different amount of TV (e.g., Group 1 = no TV, Group 2 = 10 hours per week, etc.). In this way, we control how much TV each group watches and then we can compare the GPA of each group at the end of the study to see if amount of TV watched causes students to have high or low GPAs. Amount of TV is an **independent** or **explanatory variable** in our experiment, because it is the variable that is **manipulated** or **controlled** by the researcher. An experiment **must** have at least one independent variable or it is not an experiment. An independent variable must be used in order to determine **causation**.

We could also choose to conduct a **quasi-experiment** to answer a question we're interested in. A quasi-experiment is similar to an experiment, except that the **explanatory variable we're interested in is one that is not manipulated**, but instead is measured and used to **classify** people into groups based on their scores. So if we measured the amount of TV watched by the students and then grouped them according to their scores (e.g., Group 1 = all students who report watching no TV, Group 2 = all students who report watching between 1 and 10 hours per week, etc.). In this case, amount of TV watched is not manipulated by the researcher, because students haven't been randomly assigned to the groups and told how much TV to watch. Instead, the students were assigned to the groups based on something measured from them. Amount of TV is a **subject variable** in the quasi-experiment, because it is **measured** from the students and then used to **classify** them into groups. Notice that this is different from the way amount of TV watched is used as a variable in the observational study described above. In an observational study, the variable is measured and then the score on that variable is matched with the score on another measured variable for each individual. In an observational study, individuals are not classified into groups.

Scales of Measurement



There are four scales of measurement that can be used to measure dependent variables. The researcher must choose the scale that best fits the response variable he or she wants to measure. There are **nominal**, **ordinal**, **interval**, and **ratio scales**.

Nominal scales involve response **categories** that do not fall in any particular order. For example, if you were asked to describe the weather today to indicate if it is sunny, overcast, or rainy, you would be responding on a nominal scale. There are no possible responses between the categories on the scale and the categories don't follow a particular high to low ordering.

An **ordinal scale**, however, involves **categories with an implicit order** from high to low. For example, if the possible categories for the weather today were good, medium, or bad, this would constitute an ordinal measurement scale. Note that we haven't quantified anything or said how much of a difference there is between good and medium or medium and bad. Since letters imply order, they could represent these categories, as the weather being graded A, B, or C. Numbers should not be used for response categories in ordinal scales, since they imply more than the characteristic of order. Unfortunately, numbers commonly are used for ordinal scales, as when good, medium, and bad are represented by 1, 2, and 3, or by 3, 2, and 1. (Do you see that it is arbitrary which direction you adopt?) Note that this usage does not mean that the difference between good and medium is the same as the difference between medium and bad.

Interval and **ratio scales** always involve **numerical responses**. One of the important characteristics of the number system is that the intervals between numbers are the same, that is, the interval between 1 and 2 is the same as between 2 and 3 or the 10-20 interval is the same size as the 50-60 interval. However, **interval scales** have a limitation: they no true zero point. Temperature scales (F or C) used to indicate weather are interval scales, because a value of 0 does not mean absence of temperature. Likewise, an intelligence quotient (IQ) is an interval scale because it has no true value of 0. Numerical measures that do have a true zero point are called **ratio scales**. Familiar examples are time, height, and weight. Any scale that involves counting is a ratio scale: number of items correct on a test, number of days that you study per week, number of friends that you have, etc.

Variability and Bias

There are two issues of concern when collecting or evaluating observations in a study. They are the variability of the measurements and bias in collecting them. These two topics are often covered in courses on research methods under the labels *reliability* and *validity*.

Variability: How can measurement be reliable?

Later, in Chapter 9, we will discuss the variability in a set of scores. An example is the normal temperature for a time of year. The temperature on any given day is usually not that exact temperature. Departures from the norm are variability in scores. We are accustomed to such variability; however, we expect a single measure to be accurate and without variability. But is it?

You weigh yourself at home, in the gym, and at the health center on the same day and find that you have different weights. Are the scales set (calibrated) differently? Does your weight vary at different times of day? What is your "true" weight? How do we get a reliable measure of your weight?

To prepare for the SAT test, you buy a book that has a number of sample tests. You take the test three times and get a different score each time. Were you varying how hard you

tried, or did your attention vary? Maybe the questions varied in difficulty on each test. What is your “true” scholastic aptitude? How do we get a reliable measure of it?

These are examples of variability in measurement. In everyday practice, we often neglect such variability and accept the value of a single measurement as accurate. To help reorient your thinking on this issue, consider this example. Imagine your “true” score as a bull’s-eye on a target and think of measurement as throwing a dart at it. An unskilled dart thrower will have a lot of variability; the darts will be all over the target. A skilled thrower will have low variability; the darts will be near the bull’s-eye, although getting a bull’s eye may not happen often. Good measurement is like skilled dart throwing; it has low variability. Note that, like dart-throwing, measurement is not expected to be exact; there will always be some measurement error.

Because of variability in measurement, scientists do not have high confidence that a single measurement is accurate. For a new measure, scientists take more than one reading, or administer a test more than once, and average the results. A goal of scientific research is to improve the accuracy of measurement, and a lot of research is conducted on the measures themselves. Scientists acknowledge measurement error and may include an estimate of it in their research reports. In courses on research methods and measurement, you will learn about different kinds of reliability, such as test-retest reliability and split-half reliability. The former involves giving a test twice and the latter involves comparing answers on odd question to those on even questions. The closer the two scores are, the more reliable the measure is.

Bias: How can measurement be valid?

What if you find out that you always weight two pounds less on your scale at home than the one in the gym. The latter is one of the expensive balance models, and the gym attendants assert that it is accurate. Is your scale at home biased? It’s easy to imagine that it could be so without your suspecting it. Probably most home scales are biased in this direction!

What if you find that despite getting As in all your math courses in high school, you score below average on the SAT quantitative section. The test company provides a lot of evidence that its tests are accurate. Are your school grades biased measure of “true” quantitative ability? Perhaps they are influenced by a grading system that rewards doing homework regularly and includes a lot of extra credit. If grades are serving that purpose, which is common, then they provide only a biased measure of quantitative ability.

These examples demonstrate that in everyday practice we often are unaware of bias and instead accept measures as valid. The dart-throwing example should help to sharpen your thinking on this issue. Again, imagine your “true” score as a bull’s-eye on a target and think of measurement as throwing a dart at it. Consider an unskilled dart thrower who throws the darts too low. He or she tries to adjust on the second set of throws, but they again are all too low, although less so than before. We would say the person has a bias to throw low and advise him or her to aim higher. Unlike variability, which is random or in

all directions, bias is a systematic error in one direction. Good measurement seeks to identify and overcome such biases. Note that even for unbiased measures there will still be measurement error, which is less harmful because it averages out.

Scientific measurement seeks to be valid, that is, to measure the construct of interest without being biased by some other factor. Research to achieve this is called construct validity. As in the example above, it seeks to sort out measurement of one construct (quantitative ability) from others (motivation, good study habits). Such research would support the SAT quantitative score rather than grades in math courses as a valid measure of quantitative ability.

Scientific research also seeks to be valid. Reducing bias in how research is conducted is referred to as internal validity. There has been a lot of research on experimenter bias, which is why drugs are tested with double-blind procedures: Neither the dispenser or recipient knows whether it is the drug or the placebo. Reducing bias in how the research method represents the process of interest in the natural environment is referred to as external validity. The problem is that control exercised in research can end up producing a situation that is too artificial. The results may be internally valid, but they cannot be generalized to what the entire population would do in realistic situations. Courses on research methods cover ways of improving construct validity, internal validity, and external validity.

Practice Questions: Set 1

- (1) Here is a portion of a data set that describes major league baseball players as of opening day of the 1999 season:

Player	Team	Position	Age	Salary
Dunwoody	Marlins	Outfield	24	222
Osuna	Dodgers	Pitcher	26	1050
Pettitte	Yankees	Pitcher	26	5950
Sosa	Cubs	Outfield	30	9000

- (a) What are the individuals in this data set?
- (b) In addition to the player's name, how many variables does the data set contain? Which of these variables take numerical values?
- (c) What do you think are the units in which each of the numerical values is expressed? For example, what does it mean when Sammy Sosa's salary is listed as 9000?
- (2) Does regular exercise reduce the risk of a heart attack? Here are two ways to answer this question:

Study 1: A researcher finds 2000 men over age 40 who exercise regularly and have not had heart attacks. She matches each with a similar man who does not exercise regularly, and she follows both groups for 5 years.

Study 2: Another researcher finds 4000 men over age 40 who have not had heart attacks and are willing to participate in a study. He assigns 2000 of the men to a regular program of supervised exercise. The other 2000 continue their usual habits. The researcher follows both groups for 5 years.

- (a) Explain why the first is an observational study and the second is an experiment.
- (b) Why does the experiment give more useful information about whether exercise reduces the risk of heart attacks?

(Continues on next page)

- (3) A researcher evaluates a new growth hormone. One sample of rats is raised with the hormone in their diet and a second sample is raised without the hormone. After six months, the researcher weighs each rat to determine whether the rats in one group are significantly larger than the rats in the other group.

A second researcher measures femininity for each individual in a group of 10-yr old girls who are all daughters of mothers who work outside of the home. These scores are then compared with corresponding measurements obtained from girls who are all daughters of mothers who work at home. The researcher hopes to show that one group is significantly more feminine than the other.

Explain why the first researcher is probably not concerned about the validity of measurement, whereas the second researcher probably is (hint: think about what is being measured).

- (4) Identify the scale of measurement that allows each of the following conclusions:
- (a) Peter's score is larger than Phil's, but we cannot say how much larger.
 - (b) Peter's score is three times larger than Phil's.
 - (c) Peter and Phil have different scores, but we cannot say which one is larger, and we cannot determine how much difference there is.

ANSWERS ON P. 146

Chapter 3: Experiments

Basic Facts about Experiments



So far we've talked mainly about variables as something we measure. But, variables can also be something that we manipulate as experimenters, rather than just observe or measure. In fact, that is one defining characteristic of an experiment, that the experimenter manipulates a variable in the study.

An **experiment** contains both an **independent variable** and a **dependent variable**. An independent variable is **manipulated** by the experimenter, while the dependent variable is **measured** by the experimenter.

When using observational or correlational designs, independent variables can also be called **explanatory variables** because they explain the changes that occur in the dependent variable. Dependent variables are sometimes called **response variables** because they come from the responses collected from our experiment participants (they used to be called “subjects,” a term now in disfavor).

All research requires **operational definitions** of variables. The conceptual variables we are interested in are often abstract theoretical entities. To conduct research on them, we need to define them in a concrete way so that they can be measured or manipulated. Our findings will be about this the operations selected for research and only by inference about abstractions.

An important aspect of experiments is that they allow us to determine **cause and effect** (or **causal**) **relationships** between the independent and dependent variables. In fact, when we conduct an experiment, we make a prediction or **hypothesis** about the way in which the independent variable will affect or change the values we get for our dependent variable.

Independent variables can involve **treatments** that we apply to one group of our subjects and not another group (called the **control group**) or they can just involve different treatments that we set up in our study.

For example, if we were to conduct an experiment to answer a research question about how often people faint at the dentist, we could set up a scenario where some people believe they are at a dentist's office and others do not and then count the number of people in each group who pass out.

The IV = scenario received. It has two levels:
Group 1 - dentist scenario
Group 2 - no dentist scenario



Another type of experiment we could do would be to give different kinds of therapy treatment to people who claim they usually pass out at the dentist's office and then compare the incidence of passing out across the different therapy groups to see which type of therapy works best.



IV = type of therapy received, 3 levels

Group 1 - therapy A

Group 2 - therapy B

Group 3 - therapy C



Experiments can determine cause and effect relationships because:

- We **randomly assign** subjects to groups to make sure the groups are similar before we give each group a treatment. We also make sure that we have enough subjects in each group to have similar groups. If we don't have enough subjects in each group, one subject who is different from the others could cause chance variation in the results.
- We **control variables** other than the manipulated variables that influence the dependent variable. By holding things other than the manipulated variable constant, we reduce any differences in treatment of our groups and therefore reduce variability.
- We reduce **confounding or “lurking” variables** that are not part of the independent variable, but could affect the dependent variable. If we have a lurking variable in our experiment, we can't know if changes in the dependent variable are caused by the independent variable or the lurking variable. Therefore, we must **control** these lurking variables as much as possible.

In any experiment, there will always be some amount of error, even if it is very small. Therefore, we must use statistics to determine if our data support or do not support our hypothesis. We rely on **statistical significance** to decide if our data support our hypothesis. Most of this course is learning how to make such decisions.

We obtain **statistical significance** when an observed difference (or effect) is so large that it would rarely occur by chance.

We have some specialized experimental designs for controlling for internal validity problems. Using these methods will not guarantee control of all bias, but when they are appropriate, these methods can help us control for some forms of bias. They are mentioned briefly here but covered extensively in a course on research methods.

- **Matching design:** Sometimes random assignment is not enough to control for subject differences across groups, especially if we have a low number of subjects in each group. If we are concerned about a particular difference between subjects, we can match a pair of subjects in the different groups on this characteristic. For example, if we think there may be differences on our measure based on gender alone, we might match subjects on gender and then randomly assign each member of the matched pair to a different group.
- **Double-blind designs:** In some studies (for example, drug studies) we may be concerned with experimenter bias. This might occur if the experimenter knows the hypothesis for the experiment and then administers the treatment (e.g., drug vs. placebo) to the groups. In this case, the experimenter may inadvertently treat the groups differently, which could affect the results of the study. Using a double-blind design controls for experimenter bias such that neither the subject nor the experimenter knows which treatment (e.g., drug or placebo) a subject is receiving.
- **Block designs:** Blocks designs can be conducted in a similar manner to matching designs. The primary difference is that in a block design, a group of similar subjects (e.g., females) constitute a **block** and randomization is done for each block separately, instead of within pairs as in the matching design. The matching design just a special case of the block design, where pairs constitute a block.
- **Completely randomized design:** As experiments get more complex (e.g., more than one independent variable), we may end up with more than just two or three groups of subjects. By assigning all experimental subjects at random to all treatments, we can reduce group difference biases.

Designing an Experiment

At the heart of an experiment is a comparison between two (or more) conditions. In other words, you (the experimenter) will always be comparing at least two things. This may include comparing your sample with a known population, or two (or more) different samples (groups) against each other, or even multiple scores within a single sample of individuals.

Generally the process involves a number of steps:

- Identifying your research questions
- Identifying your variables of interest
- Specifying your hypotheses (how are the variables related to one another)
- Selecting a research design
- Collecting and analyzing your data
- Drawing conclusions from your data about your hypotheses.

Let's consider as an example the steps that one may go through trying to design an experiment to test the following claim: Chocolate-covered peanuts enhance memory.

- Construct a formal hypothesis
 - e.g., Chocolate-covered peanuts improve recall scores of words.
- Identify the independent and dependent variables.
 - IV: consumption of chocolate-covered peanuts
 - DV: a measure of memory performance
- How to manipulate the IV
 - Presence or absence of chocolate-covered peanuts m & m's.
 - (How about manipulating the quantity of m & m's)
- Do we need a control group, a placebo? Any other control variables?
- Identify how we'll measure the DV.
- Are there any subject relevant variables (use randomization and matching)
- Are the effects the same for all sexes? Ages? Majors?
- Situation relevant variables (test conditions, experimenter behavior, timing)
 - e.g., the list of words, how fast presented

Selecting an Experimental Design

As mentioned above, an experiment involves a comparison between at least two groups. But there are a number of different ways to create two (or more) groups. For example, there are two different ways to handle the different levels of an independent variable. There can be **independent samples** as in the various examples above. You manipulate your independent variable across separate groups of people, so each level of your IV is given to a different group or sample of individuals. The alternate is to use **related samples**, in which you match pairs of participants into different groups receiving different levels of the IV (**matched-pair design**) or have one group of participants received each of the IV (**repeated measure design**).

We will spend the rest of the course learning about these different designs and the different statistical procedures they require. As an aid, we will use the decision tree below that lists common experimental designs and their statistical tests.

Here is an example of how to use the decision tree. Suppose that you (a statistics instructor) are interested in how well your lecture on displaying distributions worked. You decide to test your students before and after the lecture. Both tests are designed to measure the students' knowledge of experimental design.

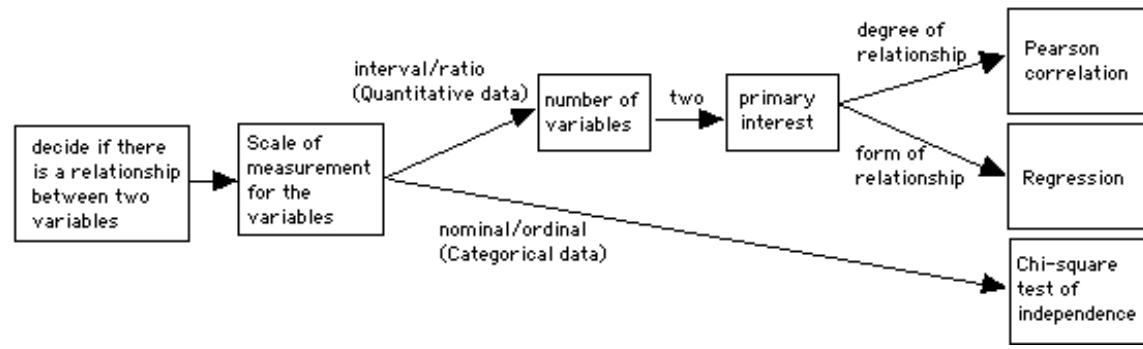
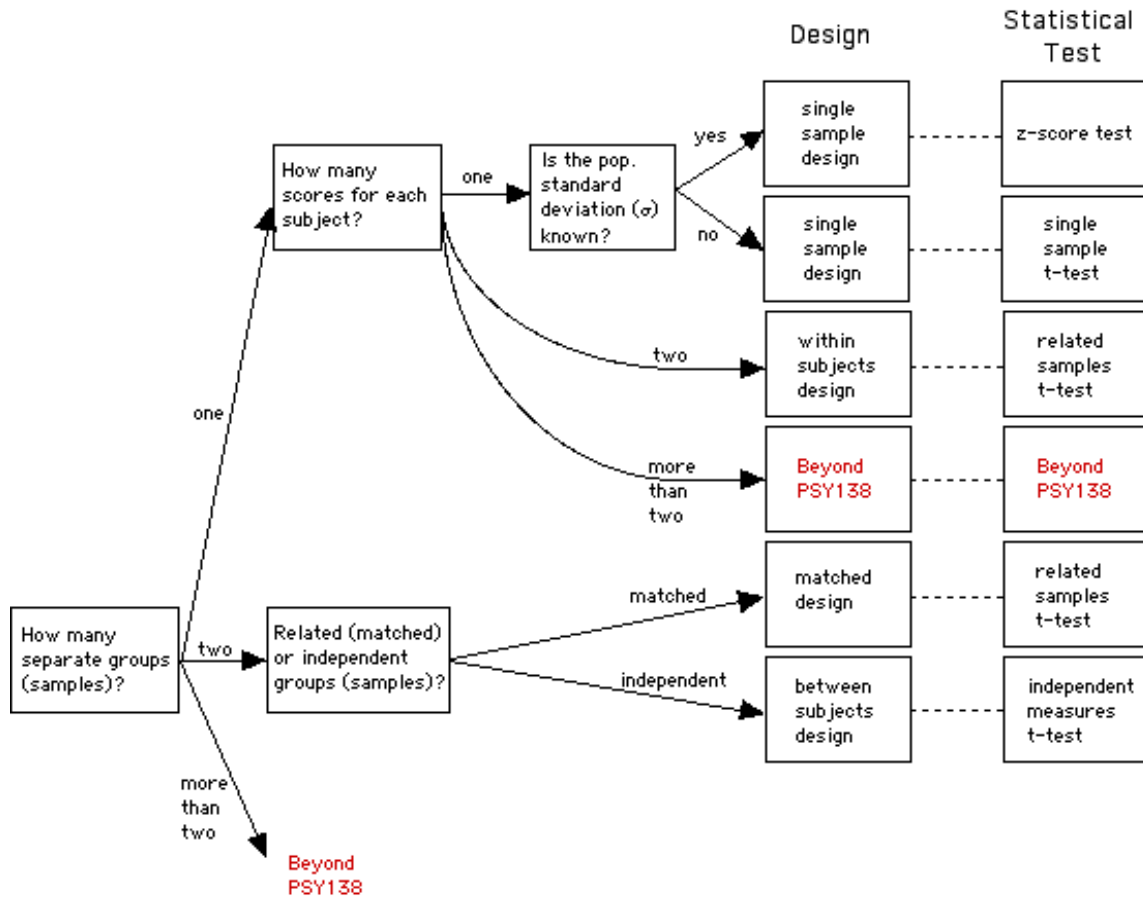
What kind of experimental design is this? Go through the questions in the tree.

- How many groups (samples) of people do you test? 1
- How many scores (pieces of data) do you collect from each person? 2

This leads you to a "within-subjects design." You've got one group, with two scores (pre-test scores and post-test scores) from each person (so the scores are "related" to each other by virtue of being from the same people). The IV here (2 levels: before lecture and

after lecture) is being manipulated as a "within groups" variable. We will return to the decision tree after we have completed our coverage of descriptive statistics and are ready to learn about inferential statistics (statistical tests).

Decision Tree



Practice Questions: Set 2

- (1) Dr. Jacobs conducts a research study investigating the effects of a new drug that is intended to reduce the craving for alcohol. A group of alcoholics who are being treated at a clinic is selected for the study. One-half of the participants are given the drug along with their regular treatment, and the other half receives a placebo. Dr. Jacobs records whether or not each individual is still sober after 6 months.
- (a) Identify the independent variable in this study.
 - (b) Identify the number of levels (and what they are) of the IV.
 - (c) Identify the dependent variable of the experiment.
 - (d) Assuming that the study includes participants in age from 18-62 years of age, what kind of variable is age?
 - (e) If the participants in the drug group are noticeably older than those in the placebo group, age may be what kind of variable?
- (2) In an experiment, participants are usually assigned to treatments using a random assignment procedure. Explain why random assignment is used.

ANSWERS ON P. 148

Chapter 4: Sampling & Basic Probability

Sampling Techniques



Once we've chosen our research design and specified our variables, we're ready to begin collecting our observations (i.e., our data). But we need to decide who to collect the observations from or who our subjects are going to be. This group of individuals who will participate as subjects in our study is called the **sample**.

In most cases the sample is a much smaller group of individuals than the group we're trying to learn about. This larger group that we want to learn about is called the **population**.

The population we're interested in is often a very large group like all humans, all Americans, or all children. Although it's the large group we're really interested in, it would be quite difficult in most cases to collect observations from all humans or all Americans. Therefore, we must represent this population with a sample that is as much like the population as possible. For example, if the population includes men and women, then our sample should also include men and women. The larger our sample, the better we will be able to represent the full population. But we must also be practical, because there will be a limit to the number of subjects we will be able to run due to time constraints and availability of the subjects.

In order to ensure a good representation of the population by the sample we choose, we rely on specific **sampling techniques** to choose members of the population for our sample. This is important because a good sampling technique can reduce **variability**. Variability is how much the scores in a data set (or distribution) are different from one another. If the variability is high, then the scores are spread out across the measurement scale and are very different from each other. Reducing the variability in a data set allows us to represent the population better and also allows us to more accurately answer our research question (more on this in Unit III of the course).

The best way to choose a sample that will represent the population (and reduce variability) is to use a **simple random sample**. This means that each member of the population has an equal chance of being chosen. This is the same thing as choosing people at random from the population. However, if we have a large population (e.g., all Americans), it will very difficult to make sure that every member of the population can be chosen. We would need to have a list of all Americans to choose from. Not even the US Census Bureau can collect responses from all Americans, and the responses they do collect require more resources than most researchers have at their disposal. Therefore, although simple random samples are the best type of samples, they are rarely used because they are too difficult to collect.

A more practical type of sample is called a **convenience sample**. A convenience sample is chosen from the population based on who is available and willing to be sampled. A convenience sample can use **volunteers** like the subject pool that is sampled from for many psychology studies. A sample can also be obtained using a technique called **stratified sampling** to ensure that certain characteristics of the population are preserved. In a stratified sample, subjects are chosen in equal proportions to those that exist in the population. For example, if in the population 30% are left-handed, then in a stratified sample, subjects will be chosen such that 30% of them are left-handed.

Probability

Why do we need to know anything about probability? In part, it is because we deal with it just about every day of our lives.



Weather forecasts (50% chance of rain, means under conditions like those that we predict we'll have today, it rains half the time)

Lotto tickets odds (chances of winning \$1,000,000 are 1 in 10,000,000; of course this means chances of winning \$0 is 9,999,999 out of 10,000,000).

The main reason that we're discussing it in this course is because probability is a central to **inferential statistics**. Inferential statistics are techniques that allow us to make decisions about entire populations using only samples. In other words, instead of testing every member of a population, we can use a subset of individuals in coordination with inferential statistics and still draw strong conclusions. However, because we use just a subset of individuals, our conclusions are made within a probabilistic context. Let's briefly consider the logic of why this is.

Suppose that you wanted to test whether all dogs have four legs. You could go out and try to check every dog in existence (do a census of dogs) so that you can count their legs. This will take a lot of resources. An alternative approach is to take a reasonable large sample of dogs and count their legs. If you find a single dog with fewer (or greater) than four legs, then you may reject the claim that all dogs have four legs. So a sample can provide enough evidence to reject a claim. On the other hand, suppose that there is a dog (Spot) in the population with 3 legs; however, that dog does not get into your sample (say of 1,000,000 dogs). Based on your sample you may wish to conclude that the claim is correct, all dogs have four legs.



However, you'd be wrong since Spot has only 3. Because the sample is only a subset of the population, you may end up missing the critical individuals within the population who would lead you to reject the claim.

So how does this all relate to probably? Recall that inferential statistics are interpreted within a probabilistic framework. This means that rather than concluding that a claim is correct, we argue that the evidence from a sample *supports* the claim with a certain level of confidence (later in the course we'll talk about something called confidence intervals). In other words we will make statements like, "based on our sample of ten thousand dogs, we conclude that it is very likely that all dogs have four legs."

The Basic Probability Formula

You should have been introduced to probability in previous courses, such as Finite Mathematics. Here we cover it briefly. In a situation where several different outcomes are possible, we define the **probability** for any particular outcome as a fraction or proportion. If the possible outcomes are identified as A, B, C, D, and so on, then:

$$\text{Probability of A} = \frac{\text{number of outcomes classified as A}}{\text{total number of possible outcomes}}$$



Let's make this more concrete with an example. Imagine that you are playing card wars with your kid sister, and each of you has your own deck of 52 cards. She picks the King of spades from her deck.

What are the odds that you'll pick the King of Spades from your deck?

$$\begin{aligned} \text{Probability of King-spades} &= \frac{\text{picking the King of Spades}}{\text{total number of possible cards picked}} \\ &= 1 / 52 \end{aligned}$$

Another way that we state the same thing is with the following notation using f for frequency of the event:

$$p(K\spadesuit) = f / N$$

This f / N formula will be important to our discussion of frequency distribution tables. This formula will be used to figure out the values in the proportions column. In fact, probabilities are most often given as proportions (but we can also give them as fractions or percentages).

Probability and Random Sampling

For this formula of probability presented above to be accurate, the selection of individuals (sampling) must be obtained by **random sampling**.

A **random sample** must satisfy two requirements:

1. Each individual in the population has an *equal chance* of being selected.

2. If more than one individual is to be selected for the sample, there must be *constant probability* for each and every selection.

Let's reconsider our card game situation. Suppose that you are a card cheat and you stacked the deck so that all of the high cards are on the top, and the low cards are on the bottom. So you turn over the top card, and surprise, it is a high card.

Was this a random sample?

No, because not every card had an equal chance of being selected (because the low cards were not near the top of the deck).

Suppose that you and your sister are playing with one deck of cards. Now she picks the King of Spades. Now you pick from the remaining cards.

Is your chance of picking the King of Spades still 1 in 52?

No, because she already picked the King of Spades, so it isn't available for future selection. To have a truly random sample, you must *replace* the King of Spades into the deck.

Sampling with replacement - a sampling method in which each sample (individual) is replaced into the population before the selection of the next sample (individual).

Practice Questions: Set 3

- (1) Your college wants to gather student opinion about parking for students on campus. It isn't practical to contact all students.
- (a) Design a bad sample. Give an example of a way to choose a sample of students that is poor practice because it depends on voluntary response.
 - (b) Design another bad sample. Give an example of a way to choose a sample of students that is poor practice that doesn't involve voluntary response.
 - (c) Design a good sample. Give an example of a way to choose a sample of students that is good practice.
- (2) Suppose that a University Club has 25 student (S) members and 10 faculty (F) members. Their names are as follows:

Barrett	S	Duncan	S	Hu	S	Lee	S	Reeder	F
Bergner	F	Frazier	S	Jarvis	F	Main	S	Ren	S
Brady	S	Gibellato	S	Jimenez	S	McBride	F	Santos	S
Chen	S	Gulati	S	Kahn	F	Nemeth	S	Sroka	S
Critchfield	F	Han	S	Katsaounis	S	O'Rourke	S	Tobin	F
Desouza	F	Hostetler	S	Kim	S	Paul	S	Tordoff	S
Draper	S	House	F	Kohlschmidt	S	Pryor	F	Wang	S

Assuming that the club may send only one person to an international conference.

- (a) What are the odds of sending Dr. Tobin to the conference?
- (b) What are the odds of sending a student to the conference?
- (c) What are the odds of sending somebody with the last name that begins with the letter K?

ANSWERS ON P. 149

Section II: Describing Data



Typically, datasets involve large sets of numbers. It is often difficult to understand the “story” that these numbers are “telling” us if we attempt to look at them all at once. In the next unit in the course we’ll examine some techniques we use to describe data. In other words, how do we take the raw scores we get (from the methods we discussed in the last unit) and turn them into something that tells a meaningful story?

As we procedure through this section of the course we’ll see that the key concept is ***variability***. The numbers in a dataset are generally measurements of variables. Recall that variables are characteristics of situations. They are called variables because they vary. The goal of the researcher is to understand and describe how (and ultimately why) the characteristic varies.

We will begin with a discussion of different ways of “seeing” how the scores of a variable are distributed across our entire dataset. Then we will discuss how to quantify (describe with numbers) the overall distribution of the variable. We will end with discussions about how we can compare distributions of different variables, which is what lies at the heart of examining how variables are related to one another.



Chapter 5: Displaying Distributions

Frequency Distributions

We can summarize the data with tables listing the frequency of each score. Creating such a table by hand is more time-consuming than stem and leaf displays, but we now have computer programs like SPSS to do so for us.

A **frequency distribution** is an organized tabulation of the number of individuals located in each category on the scale of measurement.

To help you understand the importance of a frequency distribution consider this scenario:

Assume you are interested in people's ability to regulate their actions. You administer the Temporal Discounting Measure (TDM) to 25 introductory psychology students and record their scores. Possible scores range from 10 to 20. The students' raw scores are shown below:

11	12	13	18	12
12	17	10	15	14
17	16	14	14	15
19	11	13	20	12
14	15	12	13	16

Suppose you have just finished collecting this data and are interested in determining whether this sample has scored similarly to other samples you have studied in the past. That is, other samples have usually produced scores from 10 to 20.

Consider these questions:

- Does it look to you as if this sample is scoring similar to those other samples?
- Does it appear that this sample has a majority of high-scoring individuals (closer to 20) or low-scoring individuals (closer to 10)?
- Do you see scores that look abnormal or outside the range of most other scores?

With the data in its current form, it is very difficult to answer these questions and it would be very time-consuming. There must be a more organized way to present the data that will help us answer these questions quickly. One way to organize data is to create a frequency distribution table.

A **frequency distribution** takes a disorganized set of scores and places them in order from highest to lowest, grouping together all individuals who have the same score.

Frequency distributions also show:

- Whether scores are generally high or low.
- Whether they are concentrated in one area or spread out across the entire scale.

- An organized picture of the data.
- The location of any individual score relative to all of the other scores in the set.

Creating a frequency distribution table

We will go through the steps here to create frequency distribution tables. In lab, you will learn how to use SPSS to create them for you.

1. Find the **range of responses** (highest to lowest). Here, our participants scored between 10 and 20, so we create a table with those values (in descending order) in the X column in the table. We'll get to the other columns in the steps below.

X	<i>f</i>	p	%
20			
19			
18			
17			
16			
15			
14			
13			
12			
11			
10			

2. How many of each score did we get? Fill these numbers in the *f* column; this is the **frequency** of each score.

X	<i>f</i>	p	%
20	1		
19	1		
18	1		
17	2		
16	2		
15	3		
14	4		
13	3		
12	5		
11	2		
10	1		

If you wanted to know what the total of all of the Xs were, how would you do it? The easiest way would be to multiply the X and *f* columns and then add (sum) the results: $\Sigma X = \Sigma (X * f)$. (The upper-case Greek letter for s, *sigma*, is our symbol for sum: Σ .)

Notice that if you add up the frequency (f) column, you get the total number of observations: $N = \sum f$.

3. We also typically add a column labeled p for **proportion**. This answers the question of how much of the total group got this value of X ? Using the following formula, we can fill in the p column on the table.

(Recall that N = the total number of observations.)

X	f	p	%
20	1	.04	
19	1	.04	
18	1	.04	
17	2	.08	
16	2	.08	
15	3	.12	
14	4	.16	
13	3	.12	
12	5	.20	
11	2	.08	
10	1	.04	

4. What percentage of the group got this value for X ? To answer this question we need to fill in the % column for **percent representation** in the sample. Note that all numbers entered in this column are percents; the % sign is not needed.

This information is found by multiplying the proportion column and 100: $p * 100$.

X	f	p	%
20	1	.04	4
19	1	.04	4
18	1	.04	4
17	2	.08	8
16	2	.08	8
15	3	.12	12
14	4	.16	16
13	3	.12	12
12	5	.20	20
11	2	.08	8
10	1	.04	4

- Note: Grouped frequency distribution tables. Often ranges or categories, rather than specific values, are used for X. Think of a grading scale, (A = 90-100, B = 80-89, etc...). It is possible to set up frequency distribution tables for these too.

Percentile Ranks: Locating Individuals

So far we've talked about describing an entire set of observations, but we can also use frequency distributions to describe the position of individuals within the set.

The **rank** or **percentile rank** of a particular score is defined as the percentage of individuals in the distribution with scores at or below the particular value. When a score is identified by its percentile rank, the score is called a **percentile**.

Suppose the following table contains the scores of a vocabulary quiz:

X	<i>f</i>	p	%	<i>cf</i>	c%
5	2	.05	5	40	100
4	10	.25	25	38	95
3	16	.40	40	28	70
2	8	.20	20	12	30
1	4	.10	10	4	10

cf = cumulative frequency
 c% = cumulative percentage

Cumulative frequencies (*cf*) and **cumulative percentages** (c%) simply cumulate the data from lowest score to highest score. Starting at the lowest value, add the frequencies from the first and second X values to arrive at the cumulative frequency for the second X value. (We are showing the lowest score at the bottom of the frequency table, but sometimes it is displayed at the top.)

Thus to calculate the **cumulative frequency** of value 2, you add 4 (Frequency of "1") + 8 (Frequency of "2") = 12. So, the cumulative frequency of value 2 is 12. This tells us that 12 people scored a 2 or lower. You can then add the frequencies from the first, second, and third values to arrive at the cumulative frequency of "3", and so on for all values.

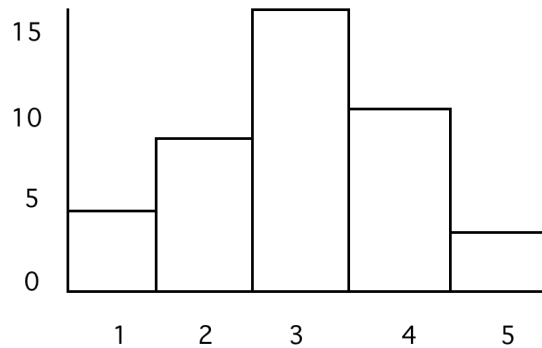
You use a similar procedure to arrive at **cumulative percentages**. Starting with the lowest value, add the percentages from the first and second category to arrive at the cumulative percentages for the second value (10% + 20% = 30%). This tells us that 30% of our sample scored a 2 or lower. Keep using this procedure to find cumulative percentages of the other values.

Probability and Frequency Distributions

How does probability relate to frequency distributions?

Consider the following frequency distribution of scores from a small population.

X	f	p
5	2	.05
4	10	.25
3	16	.40
2	8	.20
1	4	.10



Imagine that this population is made up of numbered tokens in a bag, and that your task is to reach in and pull out one token. The proportion column corresponds to the probability of selecting a token (an individual) with a particular value of X . In the frequency histogram graph, this probability is represented as the area under the bars for those intervals.

- What is the probability of selecting (sampling) a token with a 3?

$$p(3) = f / N = 16 / 40 = .40$$

So there are 40 tokens in the bag, 16 of the tokens have a 3 on them.

- What is the probability of selecting (sampling) a token with a 5?

$$p(5) = f / N = 2 / 40 = .05$$

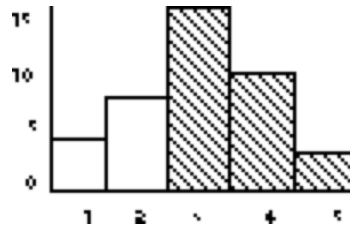
So if somebody asked you what's the likelihood of selecting a token with a 5 on it you should answer .05 (or 5%).

We can also find the answers to more complex questions.

- What is the probability of selecting a token with a value greater than 2?

$$p(X > 2) = ?$$

$$.05 + .25 + .40 = .70$$

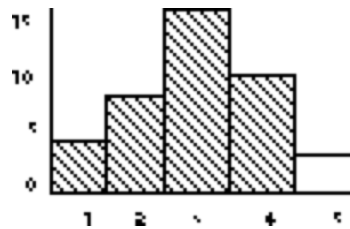


We simply add the probabilities for each value of X that is less than 2.

- What is the probability of selecting a token with a value less than 5?

$$p(X < 5) = ?$$

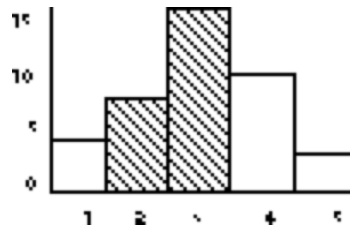
$$.10 + .20 + .40 + .25 = .95$$



- What is the probability of selecting a token with a value greater than 1 & less than 4?

$$p(4 > X > 1) = ?$$

$$.20 + .40 = .60$$



Graphs

We can also summarize the data with pictures, also known as graphs.

Graphs are pictorial representations of data from which particular characteristics of the distribution emerge.



We often use the frequency distribution table as our basis for creating graphs. Graphs can also be used to get an idea of the shape of the data distribution, that is, where are most scores clustered? Graphs and tables help us summarize data in a more efficient way than a listing of all data points. In this section, we will focus on the use of different graphing options. Your choice of how to graph and display your data will have something to do with the level of measurement of your variables, that is, what scale did you use to measure your variables?

Histogram. A histogram is used when the data are measured on an **interval** or a **ratio scale**. For a histogram, vertical bars are drawn above each score so that:

- The height of the bar corresponds to the frequency.
- The width of the bar extends to the real limits of the score (there are no spaces between the bars).

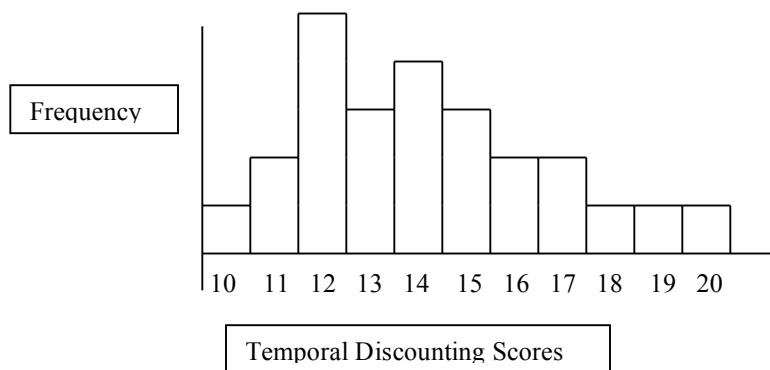
Notes:

- For a continuous variable, each score actually corresponds to an interval on the scale. The boundaries that separate these intervals are called **real limits**. The real limit separating two adjunct scores is located exactly halfway between the scores.
- Each score has two real limits, one at the top of its interval called the **upper real limit**, and one at the bottom of its interval called the **lower real limit**.
- The upper real limit of one interval is also the lower real limit of the next higher interval.

A histogram of the temporal discounting scores we previously looked at:

Horizontal bar (the X axis or abscissa) – the values of X

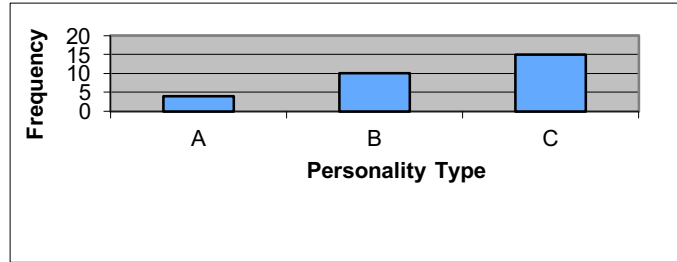
Vertical bar (the Y axis or ordinate) – the frequency value



Bar Graph. A histogram and bar chart are very similar, but a bar graph is used when the data are measured on a *nominal* or an *ordinal scale*. For a bar graph, a vertical bar is drawn above each score (or category) so that:

- The height of the bar corresponds to the frequency
- There is a space separating each bar from the next.

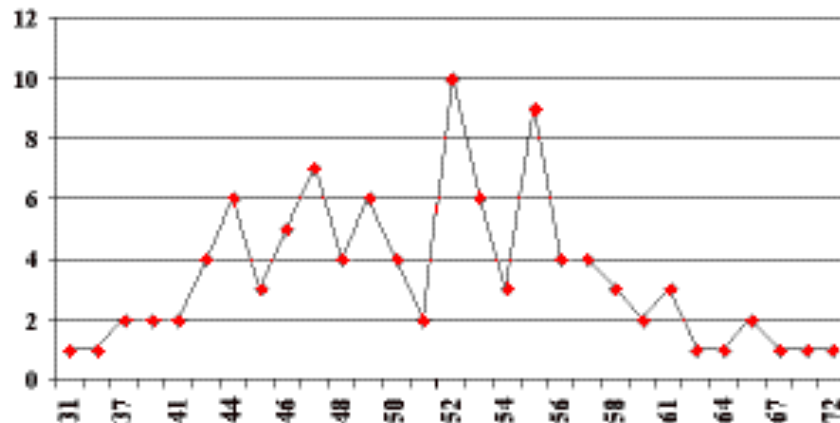
The following bar graph represents the distribution of personality types in a sample of Introduction to Psychology students. Because personality type is a discrete variable measured on a nominal scale, the graph is drawn with space between the bars.



Line Graph. (a frequency distribution polygon). A line graph is used when the data are measured on an *interval* or a *ratio* scale. In a line graph a single dot is drawn above each score so that:

- The dot is centered above the score
- The height of the dot corresponds to the frequency
- A continuous line is then drawn connecting these dots

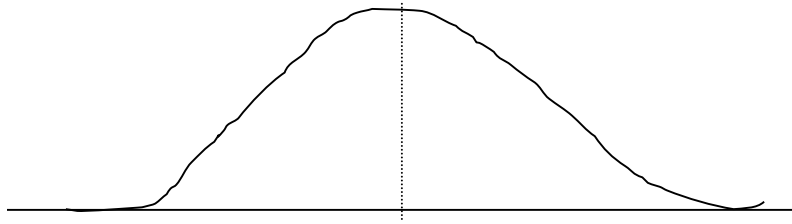
The following line graph represents the scores on an Introduction to Psychology quiz. Because the grading scale is set up on an interval scale, the data can be viewed as a line graph.



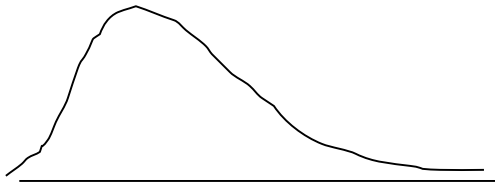
The Shape of a Frequency Distribution

When you look at a frequency distribution graph, you might have questions about the shape of the distribution. That is, is there an equal spread of scores across all possible values or did most people score at the high or low end of the scale? Nearly all distributions can be classified as being either symmetrical or skewed:

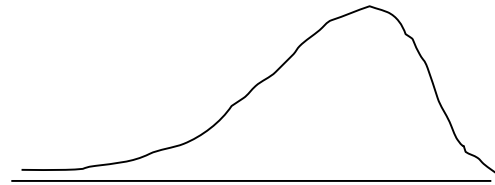
Symmetrical (bell-shaped). In a symmetrical distribution it is possible to draw a vertical line through the middle so that one side of the distribution is an exact mirror image of the other.



Skewed. In a skewed distribution, the scores tend to pile up toward one end of the scale and taper off gradually at the other end. Most scores will fall at one end of the scale and only a few scores at the opposite end of the scale.



Positively Skewed. The tail extends toward the positive or higher end of the scale. Most scores are low, fewer are high.



Negatively Skewed. The tail extends toward the negative or lower end of the scale. Most scores are high, fewer are low.

Let's see if you are able to determine the shape of the distribution of various frequency distribution graphs!

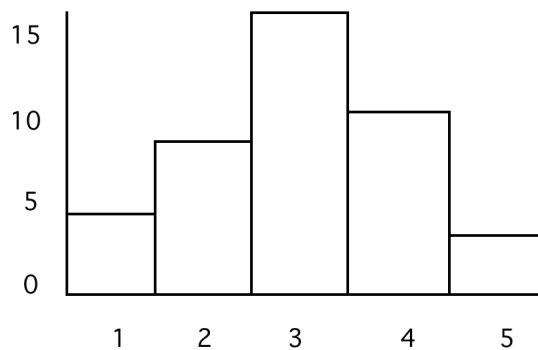
- Looking back at the bar graph of the frequency of personality types in a sample of Introduction to Psychology students, what shape is the distribution? Most scores are toward the high end so the distribution looks like it is negatively skewed.
- Now look at the line graph of the scores on an Introduction to Psychology quiz., what shape is this distribution? More scores are on the low end so this distribution looks like it is positively skewed

Probability and Frequency Distributions

How does probability relate to frequency distributions?

Consider the following frequency distribution of scores from a small population.

X	f	p
5	2	.05
4	10	.25
3	16	.40
2	8	.20
1	4	.10



Imagine that this population is made up of numbered tokens in a bag, and that your task is to reach in and pull out one token. The proportion column corresponds to the probability of selecting a token (an individual) with a particular value of X . In the frequency histogram graph, this probability is represented as the area under the bars for those intervals.

- What is the probability of selecting (sampling) a token with a 3?

$$p(3) = f / N = 16 / 40 = .40$$

So there are 40 tokens in the bag, 16 of the tokens have a 3 on them.

- What is the probability of selecting (sampling) a token with a 5?

$$p(5) = f / N = 2 / 40 = .05$$

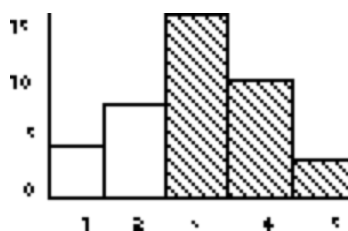
So if somebody asked you what's the likelihood of selecting a token with a 5 on it you should answer .05 (or 5%).

We can also find the answers to more complex questions.

- What is the probability of selecting a token with a value greater than 2?

$$p(X > 2) = ?$$

$$.05 + .25 + .40 = .70$$

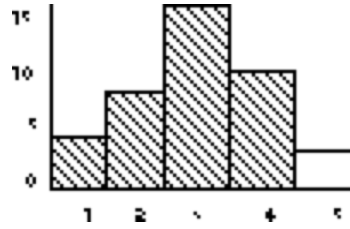


We simply add the probabilities for each value of X that is less than 2.

- What is the probability of selecting a token with a value less than 5?

$$p(X < 5) = ?$$

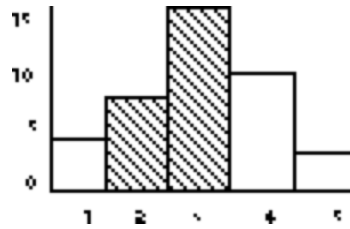
$$.10 + .20 + .40 + .25 = .95$$



- What is the probability of selecting a token with a value greater than 1 & less than 4?

$$p(4 > X > 1) = ?$$

$$.20 + .40 = .60$$



Practice Questions: Set 4

(1) Create a frequency table including the range of responses, frequency, proportion, percentage, cumulative proportion, and cumulative frequency for the following data illustrating the number of correct responses on a quiz:

1, 4, 3, 2, 3, 4, 5, 2, 3, 5, 5, 3, 2, 1, 4, 3, 2, 3, 1, 3, 4, 3, 2, 4

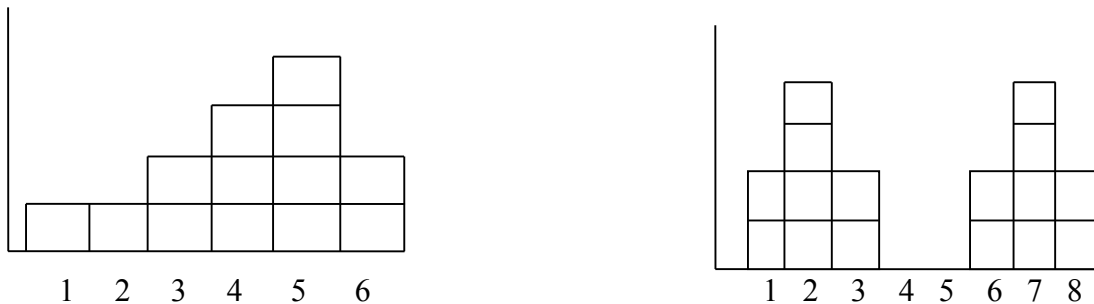
(2) What percentage of students scored a 3 or lower on the quiz in Question 2?

(3) Draw a line graph of the data from Question 2. What is the shape of this distribution?

ANSWERS ON P. 149

Chapter 6: Central Tendency

The goal of descriptive statistics is to summarize a distribution of scores (or categories, but most of this discussion is about numerical scores). The common way to do this is to find a single score that best represents the entire set of scores that you are looking at. **Central Tendency** is a statistical measure that identifies a *single score* as representative of an entire distribution. The goal of central tendency is to find the single score that is most typical or most representative of the entire group. Central tendency is a useful measure because a single score is much easier to understand than a large set of scores. Before you read on about the different types of central tendency, try to identify the single value that is most representative of each distribution below.



As you can see it is hard to locate one single score that represents both of the distributions. This is why there is not just one procedure for determining central tendency. Instead, there are three different methods for determining central tendency: the **mean**, the **median**, and the **mode**. The shape of a distribution and the type of variable you are measuring determine when you use which method of central tendency.

The Mean

The **mean** is the arithmetic average of all the scores in the distribution. It is the most common method of central tendency because it takes every item in the distribution into account and is closely related to measures of variability (which we will talk about later). For a population, it is identified by the Greek letter for m , which is *mu*, written as μ , or it may be identified by an upper-case M . For a sample, it is identified by a lower-case m or \bar{X} (pronounced “X-bar”). (There are different conventions for notations. The most common is Greek letters for population parameters, including N for the number of cases, and the corresponding Roman letter for sample statistics, including n . Also, statistical symbols are often italicized.) Since it is the arithmetic average, the formula is as following:

$$\text{Mean for a population: } \mu = \frac{\sum X}{N} \quad \text{Mean for a sample: } \bar{X} = \frac{\sum X}{n}$$

Note: These are the first of many statistical formulas that you must learn. They will all be put in shaded boxes and summarized at the end of this text.

For example, consider the sample of $n = 5$ scores: 3, 4, 5, 6, 7

$$\bar{X} = \frac{(3 + 4 + 5 + 6 + 7)}{5} = \frac{25}{5} = 5$$

The **mean** has several important properties or characteristics:

- If you change a given score, add an observation, delete an observation, then the mean will change
- If you add (or subtract) a constant to each score, then the mean will change by adding that constant.
- If you multiply (or divide) each score by a constant, then the mean will change by being multiplied by that constant.

The Median

The **median** is the score in the middle of the rest of the score. That is, half of the scores are to the right of the median and half are to the left. Thus, the goal of the median is to find the midpoint of a distribution of scores. There are different ways to find the median depending on whether you have an even number of scores or an odd number of scores.

For example, consider the sample of $n = 5$ scores: 12, 3, 6, 4, and 10.

Since the sample size is odd, you first list the scores in order from lowest to highest:

3 4 6 10 12

Then you find the middle score (6) and this is your median.

If the sample size is even like in the following set of scores: 12, 3, 7, 6, 4, and 10

Again you put the score in order:

3 4 6 7 10 12

Then you find the 2 middle scores (6 and 7) and add them together and divide by 2.

$$\text{Median} = \frac{6 + 7}{2} = 6.5$$

The **median** is best to use in the following situations:

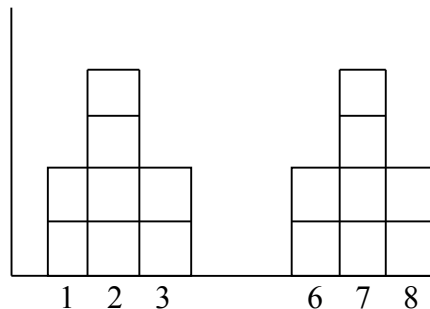
- You have extreme scores in a sample or population
- You have undetermined values for the variable you are measuring
- There is no upper or lower limit for the variable
- The variable is measured on an ordinal scale

The Mode

The **mode** is the most frequently occurring score or category in the distribution. While it can be used for any scale, this is the only measure of central tendency that can be used for nominal data. To find the mode you just need to look for the category or score that occurred most often. For example, 100 Psychology majors were asked to name their favorite Psychology course and the results were as following:

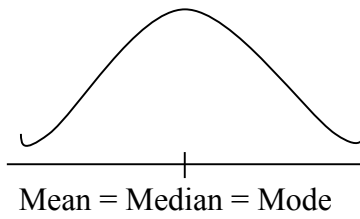
Course	<i>f</i>
Introduction	25
Abnormal	45
Developmental	10
Child	5
Educational	15

The mode of the distribution of favorite psychology courses is Abnormal Psychology since most people picked it as their favorite course. A distribution can have more than one mode. For example, in the distribution below 2 and 7 would both be modes of distribution. This type of distribution is then termed **bimodal**.

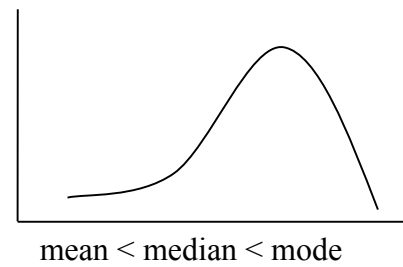
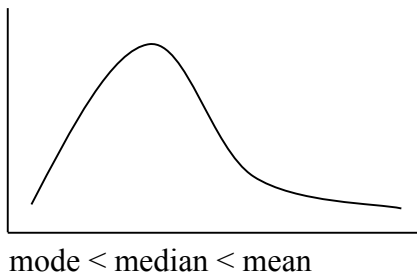


Relationships between Measures of Central Tendency and Distributions

In a **symmetrical distribution** the mean = the median = the mode. The right hand side of the distribution is exactly the same size as the left size. Thus, the median will be exactly in the middle of a symmetrical distribution because it is midpoint of a distribution. The mean will also be at the exact middle because averaging a score on the left side with the corresponding score of the right side will continually give you the middle score. The mode will also be the middle of the distribution since this is the point where the scores occur most often.



In a skewed distribution the measures are not all equal to each other. In a ***positively skewed distribution*** the peak is on the left side, which is the mode. The median is to the right of the mode, and mean is to the right of the median. These values are reversed in a ***negatively skewed distribution***.



Practice Questions: Set 5

- (1) What is the value of the mean, median, and mode for the following set of scores?
Scores: 1, 3, 5, 0, 1, 3
- (2) In a sample of $n = 6$, five individuals all have a score of 10 and the sixth person as a scores of $X = 16$. What is the mean for this sample?
- (3) After 5 points are added to every score in a distribution; the mean is calculated and found to be 30. What was the value of the mean for the original distribution?
- (4) For a perfectly symmetrical distribution with $m = 30$, the median would have a value of ___?
- (5) For the following set of scores, identify which measure would provide the best description of central tendency and explain your answer.
Scores: 0, 30, 31, 33, 33, 34, 35, 37, 38.

ANSWERS ON P. 150

Study Guide for Exam 1

Terms

bias	mode
central tendency	nominal scale
confound variable	observational study
continuous variable	operational definition
control group	ordinal scale
convenience sample	population
correlational method	probability
data	quasi-experiment
dependent variable	random assignment
discrete variable	random sampling
double-blind design	ratio scale
experimental method	reliability
experimenter bias	response variable
explanatory variable	sample
histogram	sampling techniques
hypothesis	scientific method
independent variable	simple random sample
inferential statistics	statistical significance
interval scales	stratified sampling
manipulate	subject variable
matching design	validity
mean	variability
median	variable

Formulas

Probability of a score in a distribution $p = \frac{f}{N}$

Mean for a population: $\mu = \frac{\sum X}{N}$ Mean for a sample: $\bar{X} = \frac{\sum X}{n}$

Sample problem with Calculation Procedures

You conduct a survey on how much your friends like a website and whether it is related to their GPA. Your survey's response scale runs from 0 = "not like at all" to 5 = "absolutely love". Your sample of 10 has these results for the survey (X): 5, 1, 2, 4, 3, 2, 4, 3, 0, 3.

a. Display results in a frequency distribution table and a histogram.

Describe each abbreviation in the table:

f =

p =

% =

cf =

c% =

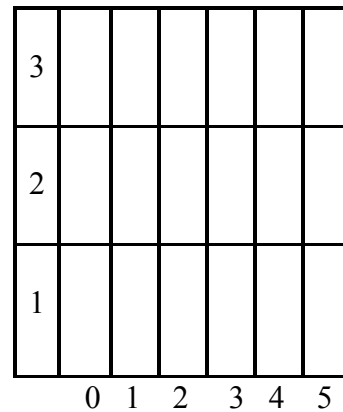
Frequency Distribution Table

Add column headings.

X					
Σ	?	?	?		

Histogram

Shade in the boxes.



b. Provide the following information about the sample:

$$p(X = 3) =$$

$$p(X > 0) =$$

$$p(0 < X < 4) =$$

$$p(X > 4) =$$

c. Find measures of central tendency for the sample:

Mean =

Median =

Mode =

ANSWERS ON P. 152

Chapter 8: Variability

Since not all scores in a distribution are the same, it is important to objectively describe the spread of scores in a distribution. **Variability** provides a quantitative measure of the degree to which scores in a distribution are spread out or clustered together.

- If scores are clustered close together, then the distribution is said to have small or low variability.
- If scores are greatly spread out, then the distribution is said to have large or high variability.
- If all scores are the same, then the distribution has no variability.

For this class we will just concentrate on three measures of variability: the **range**, the **variance**, and the **standard deviation**.

Range

The **range** is the simplest of the measures of variability to calculate. It is the difference between the largest (maximum) X value and the smallest (minimum) X value. For example, find the highest and lowest quiz score below.

1, 2, 4, 6, 8, 10

Since 10 is the highest score and 1 is the lowest score, you take $10-1$ to find the range, which is 9.

There is a downside to the range, which is that it does not take all scores in the distribution into account. Compare the range of the first example to range of this spread of quiz scores.

1, 8, 9, 9, 10, 10

As you can tell the range is also 9 in this distribution, but there are differences between the two distributions. If these scores represented points on a 10 point quiz, then the second distribution has all but one score above 80%, where as the first distribution has a wide variety of scores. It is important for a measure of variability to show this difference. Since the range is only based on the two most extreme values it cannot capture all scores in the distribution. Additionally, it becomes unstable if you repeatedly sample from the same population. Thus, the range is an unreliable measure of variability.

Variance and Standard Deviation

The other measures of variability, **variance** and **standard deviation**, do take all scores in a distribution into account and remain stable after repeated sampling. Thus, these are the most popular and most important measures of variability. Standard deviation is the one

most frequently reported for a set of scores because it is a more intuitive value, but you'll see later that we can use the variance to calculate some statistical values. The standard deviation measures how far off all of the scores in a distribution are from a standard, which is the mean of the distribution. **The variance is simply the standard deviation squared.** They are calculated a little differently depending on whether you are measuring a population distribution or a sample distribution.

For a Population Distribution

We will now construct a **deviation table**. The first step is to find the **deviation scores**. These are calculated by subtracting the mean from each score. Each deviation score tells us how far the score is from the mean. *It is the most important statistic about a score!*

Suppose we have a set of population scores: 1, 2, 3, 4, 5. In this case, $\mu = 15/5 = 3$. So we need to subtract 3 from each score to get our deviation scores.

Score	Deviation
X	$X - \mu$
5	$5 - 3 = 2$
4	$4 - 3 = 1$
3	$3 - 3 = 0$
2	$2 - 3 = -1$
1	$1 - 3 = -2$

Next you square all the deviations. This must be done because simply *adding them all the deviations together will equal out to 0*. This is because you are taking one side of the distribution and making it positive, and making the other side negative. Thus they will cancel each out. To get rid of the positive and negative signs we square the deviations and add them up. The final result is the called the **sum of squares (SS)**.

Sum of squares: $SS = \Sigma(X - \mu)^2$

Score	Deviation	Squared Deviation
X	$X - \mu$	$(X - \mu)^2$
5	2	2^2
4	1	1^2
3	0	0^2
2	-1	-1^2
1	-2	-2^2
Σ	0	10 (SS)

The next step is to find the **population variance**, which is the average of the squared deviations. To get this average we need to divide SS by the number of scores or individuals in the population (N). The symbol for the population variance is the Greek letter equivalent of s, *sigma*, which is written as σ , another notation is upper-case **SD**.

Variance for a population: $\sigma^2 = \frac{SS}{N}$

To get the standard deviation, we need to correct for all the squared deviation by taking the square root of the population variance. Thus, the **standard deviation is the square root of the mean squared deviation**.

Standard deviation for a population: $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{SS}{N}}$

From our example above the standard deviation is: $\sigma = \sqrt{\frac{10}{5}} = \sqrt{2} = 1.41$

This value 1.41, tells you *how much on average each score differs from the mean*.

To review: **Step 1:** Compute the deviation and sum of squares (SS)

Step 2: Determine the variance of the population

Step 3: Determine the standard deviation of the population

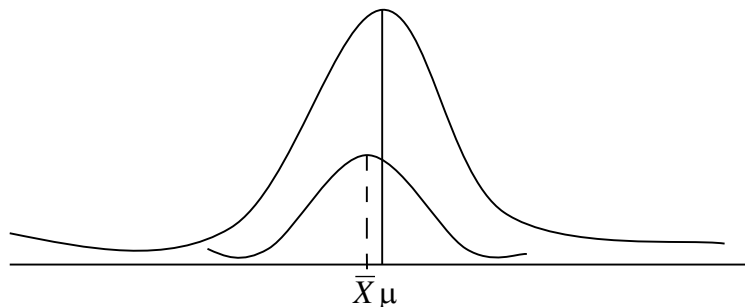
For a Sample Distribution

The computations for a sample are pretty much the same, but there are different notations (Roman letters) for the values because they are statistics instead of parameters.

s or sd = sample standard deviation

\bar{X} or m = mean

We also need to adjust the computation to take into account that a sample will typically be less variable than the corresponding population. See below.



If you have a good, representative sample, then your sample and population means should be very similar and the overall shape of the two distributions should be similar. However, notice that the variability of the sample is smaller than the variability of the population. To account for this the **sample variance is divided by $n - 1$** rather than just N .

What we're really doing here is trying to use a sample to make estimates about the nature of the population. But since we don't know things like what is the mean of the population, we really cannot measure our deviations from the population standard. So what we use is our best estimate of what the population mean is, and that is the sample mean. This will be important later in the course when we cover inferential statistics.

So what we're doing when we subtract 1 from n is using **degrees of freedom** to adjust our sample deviations to make an unbiased estimation of the population values. Programs to calculate the standard deviation (calculators, Excel, SPSS, and other spreadsheets) usually use the sample formula ($n-1$). For large N s, it doesn't make much difference.

What are degrees of freedom? Think of it this way. You know what the sample mean is ahead of time (you've got to figure out the deviations). So you can vary all but one item in the distribution to keep that same mean. But the last item is fixed. There will be only one value for that item to make the mean equal what it does. So $n - 1$ means all the values but one can vary. This is the degrees of freedom for our sample.

For example, suppose you know that the mean of your sample with $n = 5$ is 5. This means that the sum of the scores must be 25. If your first 4 items are 5, 4, 6, and 2 then what must the final number be to still have the sum of the scores equal 25?

$$5 + 4 + 6 + 2 + X = 25$$

There will be only one value of X that'll make this work: $X = 8$.

Let's do an example of computing the **standard deviation** of a sample of the following scores: 1, 2, 3, 4, 4, 5, 6, 7

Step 1: Compute the deviations and SS

Deviation Table

	Score X	Deviation (X - \bar{X})	Squared Deviation (X - \bar{X}) ²
	1	1 - 4 = -3	9
	2	2 - 4 = -2	4
	3	3 - 4 = 1	1
	4	4 - 4 = 0	0
	4	4 - 4 = 0	0
	5	5 - 4 = 1	1
	6	6 - 4 = 2	4
	7	7 - 4 = 3	9
Mean	4		
Sum		0	28 (SS)

Step 2: Determine the variance of the sample (remember it is a sample so we need to take this into account in the formula)

$$\text{Variance for a sample: } s^2 = \frac{SS}{n-1}$$

$$28/7 = 4$$

Step 3: Determine the standard deviation of the sample

$$\text{Standard deviation for a sample: } s = \sqrt{s^2} = \sqrt{\frac{SS}{n-1}}$$

$$s = \sqrt{\frac{28}{7}} = \sqrt{4} = 2$$

Properties of the Standard Deviation (Transformations)

1) Adding a constant to each score in the distribution will not change the standard deviation.

For example, if you add 2 to every score in the distribution, the mean changes by 2, but the variance stays the same (notice that none of the deviations would change because you add 2 to each score while the mean changes by 2)

2) Multiplying each score by a constant causes the standard deviation to be multiplied by the same constant.

For example, suppose your mean is 20 and that two of the scores in your distribution are 21 and 23. If you multiply 21 and 23 by 2 you get 42 and 46, and the mean also changes by a factor of 2 and is now 40. Before your deviations were $(21 - 20 = 1 \text{ \& } 23 - 20 = 3)$. Now your deviations are $(42 - 40 = 2 \text{ \& } 46 - 40 = 6)$. So your deviations got twice as big just as your mean got twice as big.

Practice Questions: Set 6

- (1) In a population of $N = 10$ scores, the smallest score is $X = 8$ and the largest score is $X = 20$. The range of the population is _____.
- (2) A sample of $n = 5$ scores, the mean is 20 and $s^2 = 4$. What is the sample standard deviation?
- (3) A population of scores has a mean of 50 and standard deviation of 12. If you subtract five points from every score in the population, then the value of the new standard deviation will be_____.
- (4) What is the value of SS for the following scores?
Scores: 1, 1, 1, 3
- (5) Compute the SS, variance, and standard deviation for the following population of scores.
Scores: 9, 1, 8, 6

ANSWERS ON P. 153

Chapter 9: z-scores



Descriptive statistics, like the mean and standard deviation, describe distributions by summarizing the center (central tendency) and spread (variability). While this isn't every detail about a distribution, it does give us a pretty good picture of what the distribution looks like.

For most bell-shaped curves (e.g., symmetrical and unimodal), the mean should be at the mid-point and the standard deviation should be somewhere half way between the mean and the most extreme values.

Our goal is to be able to find our raw scores within the distribution, and to be able to describe where it falls.

Locating a Score

Where is our raw score within the distribution?

A good point of reference is the mean (since it is usually easy to find). So a natural choice for describing the location of a data point would be the **deviation score**, which is found by subtracting the mean from the score ($X - \mu$).

- The direction is indicated by the negative or positive sign on the deviation score
- The distance from the mean is the value of the deviation score

If we are only concerned about a single distribution, then this seems to be pretty easy to do. But, if we want to compare two scores from two distributions, then the situation gets much harder. Consider the following situation

Example

You take the ACT test and the SAT test. You get a 25 on the ACT and a 620 on the SAT. The college that you apply to only needs one score. Which do you want to send them (that is, which score is better, 25 or 620)?

It is hard to do a direct comparison here because the two distributions have different properties: different means, and different variability.

How might we go about it?

- Look at the distribution graphs, locate the scores and compare -- still hard to tell.
- Think about cumulative percentiles and percentile ranks -- this might work.

- Try and take the deviations and standard deviations into account -- lets try this out!

Remember our example: you got a 25 on the ACT and a 620 on the SAT and you want to know which score is better. The population means and standard deviations for the ACT and SAT are provided.

ACT: Mean = 21, SD = 3, X = 25
Deviation = $25 - 21 = 4$
Then divide the deviation by the SD = $4/3 = 1.33$
Your ACT score is 1.33 SD above the mean

SAT: Mean = 500, SD = 100, X = 620
Deviation = $620 - 500 = 120$
Then divide the deviation score by the SD = $120/100 = 1.20$
Your SAT score is 1.20 SD above the mean

With this information, it is now easy to see which score is better! The ACT score is 1.33 SD above the mean, but the SAT score is only 1.20 SD above the mean, making the ACT score the better than the SAT score.

The comparison that we just did produced z-scores! They are discussed below.

Standardized Distribution

So to be able to make comparisons, one approach would be to transform both distributions into a standardized distribution.

A **standardized distribution** is composed of **transformed scores** that result in predetermined values for the mean and standard deviation, regardless of their values for the raw score distribution. Standardized distributions are used to make dissimilar distributions comparable.

In other words, we need to convert the two distributions into a form that allows us to make a comparison. For example, we can transform these data into z-scores. That is what we'll do: convert every score in the distribution into a standardized score, making the overall distribution standardized.

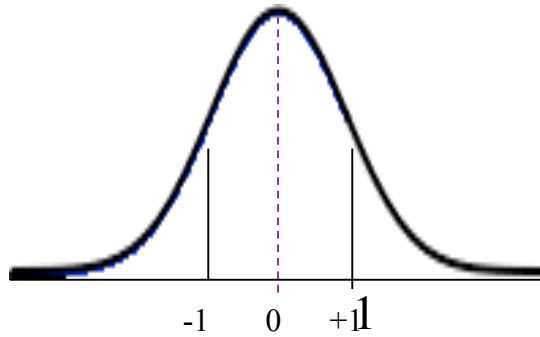
A **standard score** is a transformed score that provides information about its location in a distribution. A z-score is an example of a standard score.

A **z-score** specifies the precise location of each X value within a distribution. The sign of the z-score (+ or -) signifies whether the score is above the mean or below the mean. The numerical value of the z-score specifies the distance from the mean by counting the number of standard deviations between X and the mean.

For z-scores, the mean of the distribution is always 0 and the standard deviation is always 1.

$$\mu = 0$$

$$\sigma = 1$$



So for a z-score of 1, the data point is exactly 1 standard deviation away from the mean. If it is a positive 1, it is 1 standard deviation above the mean; if it is a negative 1, then it is 1 standard deviation below the mean.

How do we do this transformation?

For a score in a population

$$z\text{-score: } z = \frac{X - \mu}{\sigma}$$

For a score in a sample

$$z\text{-score: } z = \frac{X - \bar{X}}{s}$$

We are using the population formula for our examples, which are based on standardized scores with known population parameters. However, note that Excel and SPSS (and other programs and calculators) always use the sample formula for z-scores.

Now let's return to our ACT and SAT example. Notice what we did there: we subtracted the distribution means from the scores, and then we divided by their standard deviations. In other words what we did was transform them into z-scores. And then we made the comparisons based on those z-scores.

We can transform any and all observations or values from a distribution to a z-score if we know either the μ and σ , or \bar{X} and s .

We can also transform a z-score back into a raw score if we know the mean and standard deviation information of the original distribution. Let's look at the algebra to get from solving when z is unknown to solving when X is unknown.

$$z = \frac{(X - \mu)}{\sigma} \quad \rightarrow \quad (z)(\sigma) = (X - \mu) \quad \rightarrow \quad X = (z)(\sigma) + \mu$$

So suppose that you know somebody else who said that they got 2 SD above the mean on the SAT. How would we go about figuring out their score?

- 2 SD above = z of 2.0
- We know that the mean of SAT = 500, and the SD = 100, so we just plug in the numbers: $X = (z)(\sigma) + \mu = (2)(100) + 500 = 200 + 500 = 700$

Properties of the z-score distribution

Shape - the shape of the z-score distribution will be exactly the same as the original distribution of raw scores. Every score stays in the exact same position relative to every other score in the distribution.

Mean - when raw scores are transformed into z-scores, the mean will always = 0.

In the examples, enter the mean as the X score.

$$z = \frac{X - \mu}{\sigma} \begin{array}{l} \mu = 100, \sigma = 10; z = (100 - 100) / 10 = 0 \\ \mu = 200, \sigma = 10; z = (200 - 200) / 10 = 0 \\ \mu = 100, \sigma = 20; z = (100 - 100) / 20 = 0 \end{array}$$

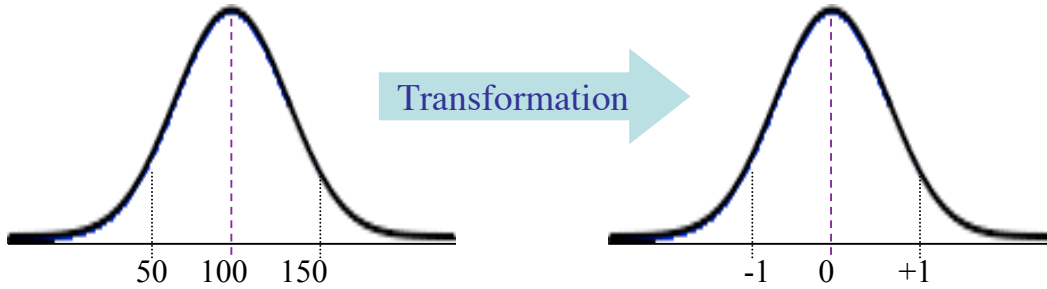
Standard deviation - when any distribution of raw scores is transformed into z-scores the standard deviation will always = 1.

In the examples, enter (mean +1 SD) as the X score.

$$z = \frac{X - \mu}{\sigma} \begin{array}{l} \mu = 100, \sigma = 10; z = (110 - 100) / 10 = 1 \\ \mu = 200, \sigma = 10; z = (210 - 200) / 10 = 1 \\ \mu = 100, \sigma = 20; z = (120 - 100) / 20 = 1 \end{array}$$

In other words: The transformation procedure really is just a way of re-labeling the axis of the distribution. So imagine that you leave the curve alone, but just draw new labels on the X-axis centering it on 0 and making each SD interval equal to 1.

Example: $\mu = 100, \sigma = 50$



$$X_{\text{Mean}} = 100$$

$$X_{+1\text{SD}} = 150$$

$$X_{-1\text{SD}} = 50$$

$$Z_{\text{Mean}} = \frac{100 - 100}{50} = 0$$

$$Z_{+1\text{SD}} = \frac{150 - 100}{50} = +1$$

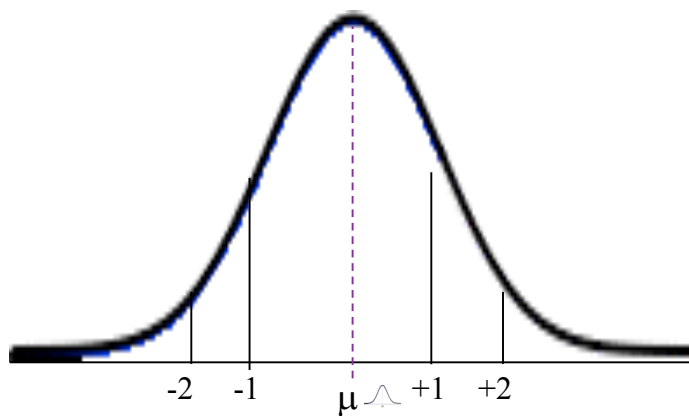
$$Z_{-1\text{SD}} = \frac{50 - 100}{50} = -1$$

Chapter 10: Normal Distributions

Normal Distributions

A normal distribution is a mathematical curve that provides a good model of relative frequency distributions found in behavioral research. In other words, it is a theoretical curve (generated by a formula) that provides a good fit to empirical findings.

The **Normal distribution** is a commonly found distribution that is symmetrical and unimodal.



Properties of a Normal Distribution

- UNIMODAL: the most frequently observed value of X is that value of X falling exactly at the mean of the distribution.
- BELL-SHAPED curve
- SYMMETRICAL about its mean. The scores above the mean form a mirror image of the distribution of scores below the mean
- MEAN, MODE, and MEDIAN of the distribution are the same

Other Technical Properties Normal Distributions

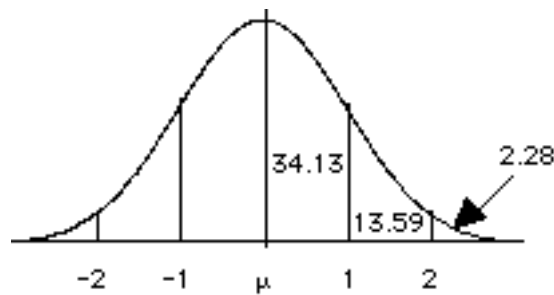
- Not all unimodal, symmetrical curves are normal, but a lot are.
- We'll assume that distributions we discuss are normal, and we won't worry about how close a distribution is to normal.

- A smooth curve like that above is referred to as a density curve and calculus (which we won't do) can solve for the percentage of areas under it.
- The area of segments under any density curve must sum to 1. This is consistent with proportions of frequency distributions totaling 1.

Normal distribution and z-scores

Like any other set of scores, the normal distribution can be transformed into z-scores. Calculus can solve for the proportion or percentage of the curve between any set of z-scores or beyond any z-score (that is, in the tail). Here are percentages for three prominent segments of the normal distribution.

- 34.13% of scores fall between μ and 1 SD.
-
- 13.59% of scores fall between 1 SD and 2 SD.
-
- 2.28% of scores fall between 2 SD and 3 SD.

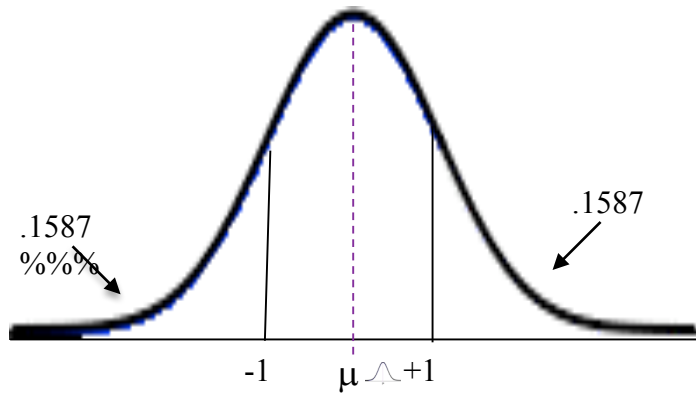


Fortunately, we don't have to figure out these percentages. An important tool in statistics is a table that provides such proportions. One such table is the **Unit Normal Table**. It provides solutions for z-scores of all the segments of a normal curve, that is, the area under the curve (and thus the probability of sampling) for that segment. There is one at the end of this reading packet, and others can be found in any statistics book.

Using the Unit Normal Table

Below is a portion of the table on pages 134-135. The first column is the z-score in question, which can be + or -. The rest of the columns give the **proportion** of the distribution **beyond the z-score** (the tail). The headings of the columns are the second decimal digit of the z-score (e.g., 1.00, 1.01, etc.).

z	.00	.01
0	.5000	.5040
⋮	⋮	⋮
⋮	⋮	⋮
1.0	.1587	.1562
⋮	⋮	⋮
⋮	⋮	⋮
2.0	.0228	.0222
⋮	⋮	⋮
⋮	⋮	⋮
3.0	.0013	.0013



At $z = 1$, 15.87% of the distribution lies in the tail beyond it (to the right if $+1$ or to the left if -1).

A. How to find the probability of a more extreme z-score (of being in a tail) from the Unit Normal Table. (This is what we will be doing in the rest of this course.)

Step 1: Sketch the distribution, showing the mean and standard deviation.

Step 2: Sketch the score (X) in question, being sure to place it on the correct side of the mean and roughly the correct distance from the mean.

Step 3: Read the problem again to see if you need the probability of getting a score greater than ($>$) or less than ($<$) X . Shade this area on your sketch.

Step 4: Translate the score (X) into a z-score.

Step 5: Look up the z-score **in the first column** and go across to the column for the second decimal place and find the probability.

Examples:

What is the probability of having an IQ of 85 or less?

$$p(X \leq 85)$$

$$\text{For IQ scores, } \mu = 100, \sigma = 15$$

$$z = (85 - 100)/15 = -1.0$$

Look up (-)1.00 in the table →
 $p = 0.1587$

What is the probability of having an IQ of 145 or above?

$p(X \geq 130)$

For IQ scores, $\mu = 100$, $\sigma = 15$

$z = (145 - 100)/15 = 3.0$

Look up 3.00 in the table →

$p = 0.0013$

B. How to find the z-score for a probability of being in the tail:

Step 1: Sketch the normal distribution.

Step 2: Shade the region corresponding to the required probability.

Step 3: Find the number closest to the probability **in the body of the table**.

Step 4: Go across to the first column to find the z-score having this probability.

Step 5: If needed, compute the corresponding raw score (X) from the z-score.

Example:

What IQ score do you need to have to be in the top 5% of the population?

The upper-tail is needed.

$p = 0.05$

Look in the body of the table for the number closest to 0.0500 and go across to the first column →

$z = 1.65$ (or 1.64)

Compute $X = (1.65)(15) + 100 = 124.75$

C. How to find the probability that X will fall between two scores (rather than above a score or below a score, as in A).

Step 1: Sketch the curve and shade the region of interest.

Step 2: Translate both scores to z-scores.

Step 3: Look up the probabilities of scoring $<$ or $>$ each of the two z-scores.

Step 4: Add (or subtract) the probabilities accordingly.

Example:

What is the probability of scoring between 300 and 650 on the SAT?

For SAT scores, $\mu = 500$, $\sigma = 100$

$z = (650 - 500)/100 = 1.5$

Look up in table $p(z \geq 1.5) \rightarrow 0.0668$ (upper tail to exclude)
 $z = (300 - 500)/100 = -2.0$
 Look up in table $p(z \leq -2.0) \rightarrow 0.0228$ (lower tail to exclude)
 Compute $p(-2.0 \geq z \geq 1.5) = 1 - 0.0668 - 0.0228 = 0.9104$

D. How to find the probability that X lies outside two points (the complement of C).

Example:

What is the probability of scoring lower than 300 or higher than 650 on the SAT?
 Same as above, but add the two probabilities together
 $p(-2.0 \leq z \geq 1.5) = 0.0668 + 0.0228 = 0.0896$.
 (Note that this probability and that in C add to 1.0 because they comprise the entire distribution.)

E. How to find percentile ranks and interquartile ranges.

Examples:

What is the interquartile range for the SAT?

For SAT, $\mu = 500$, $\sigma = 100$

Look up in the probabilities of 0.25 and 0.75

$0.25 \rightarrow z\text{-score} = -0.67$

$0.75 \rightarrow z\text{-score} = +0.67$

$X = z(\sigma) + \mu$

$= (-.67)(100) + 500 = 433$

$= (+.67)(100) + 500 = 567$

$IQR = 567 - 433 = 134$

What is your percentile rank if you have an IQ of 130?

For IQ, $\mu = 100$, $\sigma = 15$

$z = (130 - 100)/15 = 2.0$

Look up in the table, for $z = 2.0 \rightarrow$

$p = 0.9772$ or percentile rank or 97.72

(There is a short cut for figuring out the IQR. Since the range is always $\pm .67(\sigma)$, then you can compute the IQR as being $(2)(.67)(\mu)$. Example: for SAT: $(2)(.67)(100) = 134$.)

Practice Questions: Set 7

Use the following means and standard deviations: for ACT, $\mu = 21$, $\sigma = 3$
and for SAT, $\mu = 500$, $\sigma = 100$.

- (1) You take the ACT test and the SAT test. You get a 24 on the ACT and a 660 on the SAT. The college that you apply to only needs one score. Which do you want to send them (that is, which score is better, 24 or 660?). Why?
- (2) What is the probability of having an ACT score of 20 or less?
- (3) What SAT score do you need to have to be in the top 15% of the population?
- (4) What is the probability of scoring between 500 and 650 on the SAT?
- (5) What is your percentile rank if you have an ACT of 25.5?

ANSWERS ON P. 154

Chapter 11: Scatterplots & Correlations

Correlation

So far we've looked at how single variables act. However, we are often interested in how two (or more) different variables may be related to one another, that is, how they act together.

Correlation is a statistical technique that is used to measure and describe a relationship between two variables. Usually the two variables are observed as they exist naturally in the environment; there is no attempt to control or manipulate the variables.

Sometimes we are interested in whether there is a relationship between variables, but we don't plan to make any **causal** claims. That is, we aren't planning on making any conclusions like "*X causes Y*", but instead want to say something like "*X and Y are related*". If so, we can predict a person's score on one variable by knowing their score on the other one. A familiar example is weight and height.

However, suppose we are interested in the determinants of a person's height, that is, what makes a person tall, average height, or short. Here, a person's height is known as the **outcome variable**. It is the variable that we are interested in predicting or explaining. The other type of variable is a **predictor variable**. Here, we should think of all of the variables might influence or explain height. Let's consider the following partial list:

- Average of your parent's height
- Your current age
- Your gender
- Your weight

All of these variables are predictor variables because they are potential causes or influences on our response variable (your height). Notice that we can establish a value for each variable **for each individual**. (Remember how in SPSS we can code gender with numerical values.)

Regardless of whether we want to establish a causal relationship, there are steps to follow in examining any relationship between two variables:

- Make a scatterplot
- Look for overall patterns and deviations from those patterns
- Compute the Correlation Coefficient (r)

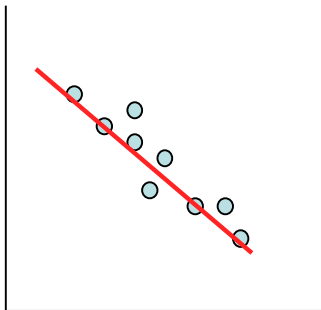
Scatterplots

In a **scatterplot** (also called a scattergram), you put two variables onto the Y and X axes, that is, you plot one variable against the other. Unless you have specified one as independent and one as dependent, it doesn't matter which variable is on the x axis and which variable is on the y axis. If you have specified variables, the independent one goes on the x axis and the dependent one goes on the y axis. **Each point represents a case or person**; the point is the intersection of the person's scores on the two variables. Imagining a line through the complete set of points is useful for seeing important features of the relationship.

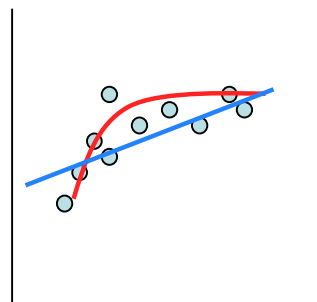
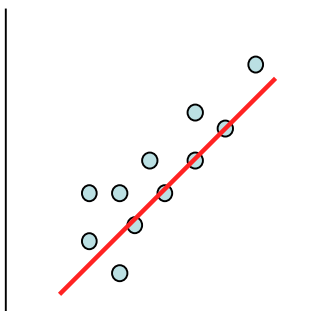
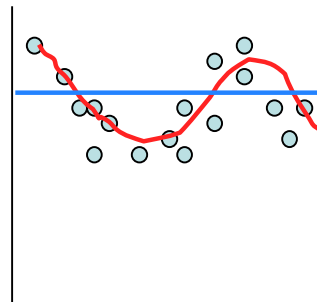
Form of the relationship

We will focus on **linear correlations** (straight lines), but there are also other forms that the relationship can take.

Linea



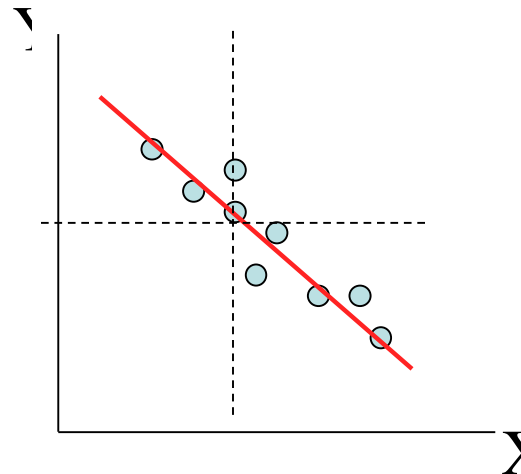
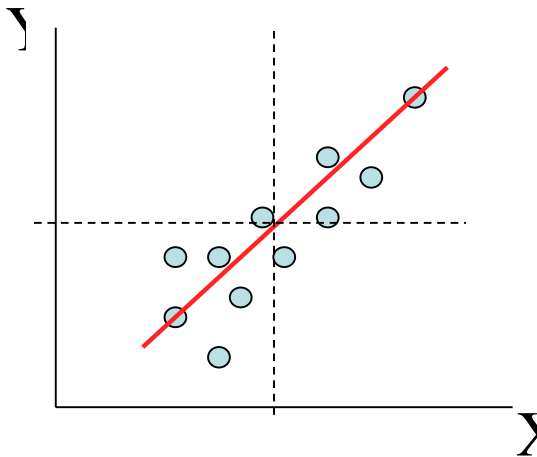
Non-linear



Direction of the relationship

Positive correlation means that the two variables tend to move in the same direction. That is, as one gets larger, so does the other.

Negative correlation means that the two variables tend to move in opposite directions. That is, as one gets larger, the other gets smaller.



Positive

- X & Y vary in the same direction
- As X goes up, Y goes
- Positive Pearson's r
- If we form quadrants by drawing lines through means of X & Y, most points are in the 1st & 3rd quadrants (more on this in class)

Negative

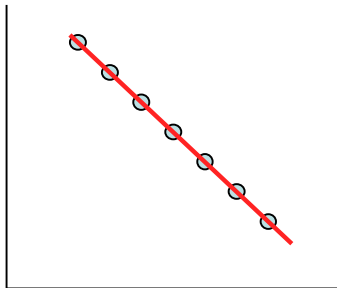
- X & Y vary in opposite directions
- As X goes up, Y goes down
- Negative Pearson's r
- If we form quadrants by drawing lines through means of X & Y, most points are in the 2nd & 4th quadrants (more on this in class)

Degree of the relationship

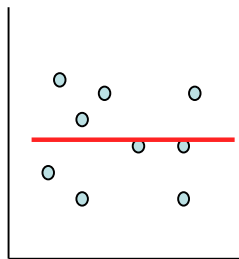
A correlation also measures the **strength** of the relationship between X and Y.

A correlation will have a **value between -1 and +1**.

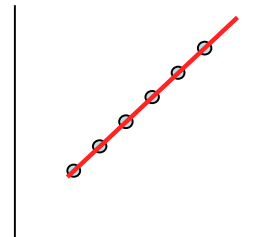
- 0 means there is no relationship; for each value of X, the best estimate of Y is the mean of Y (so knowing X adds no information)
- +1 means there is a perfect positive correlation; for each value of X, an exact value of Y can be predicted.
- -1 means there is a perfect negative correlation; again, for each value of X, an exact value of Y can be predicted.



$r = -1.0$
perfect negative
correlation



$r = 0.0$
no relationship

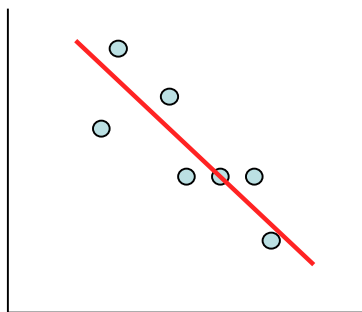


$r = 1.0$
perfect positive
correlation

Strength of the Relationship

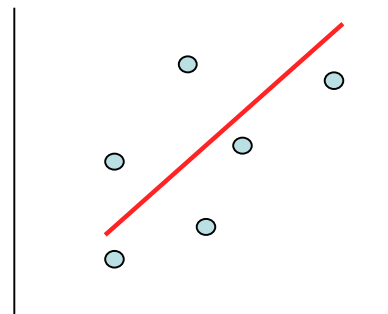
The further from zero, the stronger the relationship.

Relation A



$r = -0.75$

Relation B



$r = 0.40$

Which relationship is stronger? Relation A is stronger because it is further from zero than relation B. Remember, the further from zero the stronger the relationship.

Why do we use correlations?

Prediction: If we know that two variables are strongly related, then we may be able to predict the value of one based on the value of the other.

Example: If you know that ultrasound measurements of a baby's head are positively correlated with birth weight, then you can make an educated guess of the baby's birth weight by measuring the baby's head from an ultrasound.

Validity: If you develop a new test (TEST A) for X, and you want to know whether it is truly measuring X, then you can see if TEST A correlates with things that you already know correlate with X.

Example: If you discover a new formula for predicting birth weight (imagine some magic formula that includes the height and weight of the mother and father combined), then this formula should also correlate with the ultrasound estimates of birth weight.

Reliability: If you use the same test twice on the same individuals, you can correlate the two sets of scores. If the test is reliable, then it should give similar results both times, giving you a high correlation between time 1 and time 2.

Theory Verification: Many theories will predict that a relationship exists between different variables. So you can then go out, collect some data, and see if such a relationship exists.

What does a correlation mean mathematically?

How do we quantify the idea of correlation? There are a number of different correlations; we will focus on the most common measure, the **Pearson product-moment correlation**, represented by the notation, **r**. Here is a mathematical definition.

$$r = \frac{\text{degree to which X and Y vary together}}{\text{degree to which X and Y vary separately}}$$

$$r = \frac{\text{covariability of X and Y}}{\text{variability of X and Y separately}}$$

You should be able to understand this better later in the course, but here is the beginning of an explanation. Correlation or covariation means that as X changes, Y also changes. Remember that a perfect correlation is $r = 1.0$ (or -1.0). So, in the above formulas, if $r = \pm 1.0$, then the numerator must equal the denominator. On the bottom, we have two things,

how much does X change and how much does Y change irrespective of each other. On the top we have, how much X and Y change together. If their covariability is the same as their separate variability, then we can predict how much X changes when we know how much Y changes (or vice versa). As the covariability gets smaller, our ability to predict the changes in Y given the changes in X (or vice versa) gets less and less. If there is no covariability, the numerator is 0, and $r = 0$.

Computing Pearson's r (SS formula)

We need to construct a **bivariate deviations table**. It includes deviations for X and for Y as well as their joint variability.

Step 1. Variability of X and Y separately: We'll use the Sum of Squares as a measure of variability for X and for Y, just as we did when we calculated the variance and standard deviation.

$$SS_X = \Sigma(X - \bar{X})^2, \text{ where } \bar{X} \text{ is the mean of X values}$$

Step 2. Covariability of X and Y: We'll be computing something new, cross-products and the **Sum of the Products (SP)**. Note that SP is analogous to SS, where deviations are multiplied by themselves; here deviations of X are multiplied by deviations of Y. These are referred to as cross-products.

$$\text{Sum of the Products: } SP = \Sigma(X - \bar{X})(Y - \bar{Y})$$

Bivariate Deviations Table: Positive r

	X	Y	$(X - \bar{X})$	$(X - \bar{X})^2$	$(Y - \bar{Y})$	$(Y - \bar{Y})^2$	$(X - \bar{X})(Y - \bar{Y})$
	0	1	-6	36	-1	1	6
	10	3	+4	16	+1	1	4
	4	1	-2	4	-1	1	2
	8	2	+2	4	0	0	0
	8	3	+2	4	+1	1	2
Mean	6	2					
Σ	30	10	0	64 (SS_X)	0	4 (SS_Y)	14 (SP)

Step 3. Compute Pearson's Correlation Coefficient by inserting the above numbers we found for SS_X , SS_Y , and SP.

$$\text{Pearson's } r = \frac{SP}{\sqrt{SS_X SS_Y}} = \frac{14}{\sqrt{64 * 4}} = \frac{14}{16} = 0.875$$

This indicates that there is a fairly strong positive correlation: as X goes up we can predict that Y will too. Notice that *all of the cross-products are positive* because the X- and Y-deviations were either both positive or both negative. This results from their both being above their respective means or both below them.

Chapter 12: Correlations, continued

Computing different values of r using the deviations method

Fortunately, we do not have to go through the tedious calculation of r-values; Excel, SPSS, and advanced calculators do it for us. However, going through the calculation process helps in understanding how the correlation formula works. We'll now consider a negative correlation. We will start with the same data set above but *change Y values so there are in the opposite direction of their paired X value* with the constraint that the Y mean be unchanged. Changed numbers are *in italics*.

Bivariate Deviations Table: Negative r

	X	Y	$(X - \bar{X})$	$(X - \bar{X})^2$	$(Y - \bar{Y})$	$(Y - \bar{Y})^2$	$(X - \bar{X})(Y - \bar{Y})$
	0	4	-6	36	+2	4	-12
	10	0	+4	16	-2	4	-8
	4	2.5	-2	4	+0.5	0.25	-1
	8	1.5	+2	4	-0.5	0.25	-1
	8	2	+2	4	0	0	0
Mean	6	2					
Σ	30	10	0	64 (SS _X)	0	8.5 (SS _Y)	-22 (SP)

$$r = \frac{SP}{\sqrt{SS_X SS_Y}} = \frac{-22}{\sqrt{64 * 8.5}} = \frac{-22}{23.3} = -0.94$$

Notice that now *all the cross-products are negative (or zero)*. For every X that is below its mean the corresponding Y is above its mean, and for every X above its mean the corresponding Y is below its mean. SP is negative, which determines the sign of the correlation coefficient. The r-value is a little higher because we made sure all the Ys were similar in their deviations, just in the opposite direction.

As correlations approach 0, *cross-products will have a mix of positive and negative values*. Again, we will change Y values with the constraint that the Y mean be unchanged.

Bivariate Deviations Table: r = 0

	X	Y	$(X - \bar{X})$	$(X - \bar{X})^2$	$(Y - \bar{Y})$	$(Y - \bar{Y})^2$	$(X - \bar{X})(Y - \bar{Y})$
	0	2	-6	36	0	0	0
	10	0	+4	16	-2	4	-8
	4	1.5	-2	4	-0.5	0.25	1
	8	2.5	+2	4	+0.5	0.25	1
	8	4	+2	4	2	4	4
Mean	6	2					
Σ	30	10	0	64 (SS _X)	0	8.5 (SS _Y)	-2 (SP)

$$r = \frac{SP}{\sqrt{SS_X \cdot SS_Y}} = \frac{-2}{\sqrt{64 * 8.5}} = \frac{-2}{23.3} = -0.086$$

Notice how the cross-products cancel each other out so that SP is close to 0.

Finally, *as correlations approach 1, X- and Y-deviations become more and more alike.* When the correlation = 1, SS_X, SS_Y, and SP are the same.

Bivariate Deviations Table: r = 1

	X	Y	(X - \bar{X})	(X - \bar{X}) ²	(Y - \bar{Y})	(Y - \bar{Y}) ²	(X - \bar{X})(Y - \bar{Y})
	0	0.5	-6	36	-1.5	2.25	9
	10	3.0	+4	16	+1.0	1.00	4
	4	1.5	-2	4	-0.5	0.25	1
	8	2.5	+2	4	+0.5	0.25	1
	8	2.5	+2	4	+0.5	0.25	1
Mean	6	2					
Σ	30	10	0	64 (SS _X)	0	4 (SS _Y)	16 (SP)

$$r = \frac{SP}{\sqrt{SS_X \cdot SS_Y}} = \frac{16}{\sqrt{64 * 4}} = \frac{16}{16} = 1$$

Computing r from z-scores

An alternate method of calculating Pearson's *r* is with a z-formula. The advantage of this method is that both sets of scores are on the same scale, so it is easier to see the relationship between deviations. In the example above, it is not immediately apparent from the deviations that there is a perfect correlation. The table below has the same data; note the z-scores for each XY pair, in italics. They are the same for every X, Y pair! The sum of the z-cross-products add up to n-1!

Bivariate z Table: r = 1

	X	Y	(X - \bar{X})	<i>z_X</i>	(Y - \bar{Y})	<i>z_Y</i>	<i>z_Xz_Y</i>
	0	0.5	-6	<i>-1.5</i>	-1.5	<i>-1.5</i>	2.25
	10	3.0	+4	<i>1.0</i>	+1.0	<i>1.0</i>	1.00
	4	1.5	-2	<i>-0.5</i>	-0.5	<i>-0.5</i>	0.25
	8	2.5	+2	<i>0.5</i>	+0.5	<i>0.5</i>	0.25
	8	2.5	+2	<i>0.5</i>	+0.5	<i>0.5</i>	0.25
Mean	6	2					
Σ	30	10	0	0	0	0	4
s	4						

Formulas for s, z, & SS are needed. $s = \sqrt{\frac{SS}{n-1}}$ $z = \frac{X - \bar{X}}{s}$ $\sqrt{SS} = s\sqrt{n-1}$

Using the above formula for SS, we can get from the deviation formula to the z formula for Pearson's r. Note in the middle term how *the deviations over the standard deviation for X and Y resolves to their z-scores over n-1*. In the z-formula, r equals the sum of z products as a proportion of the degrees of freedom. The deviations formula and z-scores formula for Pearson's r are equal to one another and therefore calculate the same r-value.

$$r = \frac{SP}{\sqrt{SS_X SS_Y}} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{s_X s_Y (n-1)} = \frac{\sum z_X z_Y}{n-1} = \frac{4}{4} = 1$$

Here is another example from above, now with z scores.

Bivariate z Table: r = 0

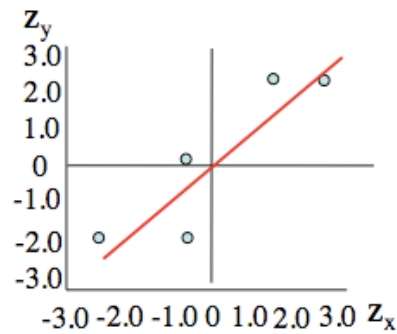
	X	Y	(X - \bar{X})	z_X	(Y - \bar{Y})	z_Y	$z_X z_Y$
	0	2	-6	-1.5	0	0	0
	10	0	+4	1.0	-2.0	-1.37	-1.37
	4	1.5	-2	-0.5	-0.5	-0.34	0.17
	8	2.5	+2	0.5	0.5	0.34	0.17
	8	4	+2	0.5	2.0	1.37	0.68
Mean	6	2					
Σ	30	10	0	0	0	0	-0.345
s	4	1.46					

$$r = \frac{\sum z_X z_Y}{n-1} = \frac{-0.345}{4} = -0.086$$

Scatterplots with z-scores

With z-scores, the center of the four quadrants is at 0,0. Each quadrant has different signed values for X and Y: quadrant 1 (+X, +Y), quadrant 2 (-X, +Y), quadrant 3 (-X, -Y), quadrant 4 (+X, -Y). The more closely related X and Y, the more similar will be their z scores.

Convert X and Y to z-scores



How to Interpret Correlations

We've already looked at basic properties of correlations: their form (linear or non-linear), direction (positive or negative), and strength (none, weak, strong, perfect). But there are some additional issues that we need to consider:

- Correlations are greatly affected by the range of scores in the data.
- Extreme scores (outliers) can have dramatic effects on correlations.
- Correlations describe a relationship between two variables, but DOES NOT explain why the variables are related.

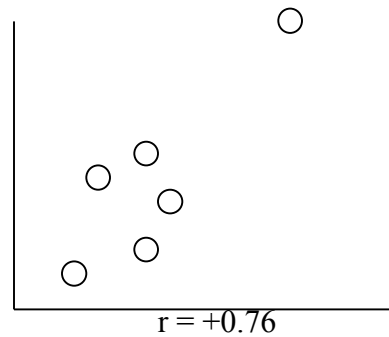
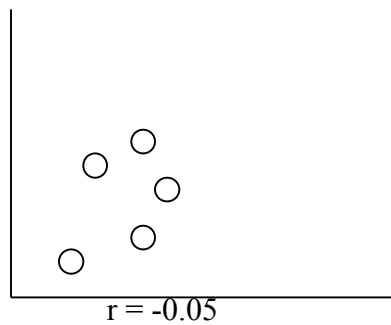
Correlations are greatly affected by the range of scores in the data

Suppose that in one study we look for a correlation between age and height, but we only test 0 to 10 yr olds. In a second study we look for the same relationship, but only test 20 to 30 yr olds. In the first case we will probably find a strong positive correlation, but in the later case we may find a near 0 correlation.

Which correlation is correct? Both are, if considered with respect to the range represented in the data. We should conclude that the strong positive correlation exists for a **restricted range**. That is, from years 0 to 10, there is a strong positive correlation between age and height. (Note: a non-linear function is appropriate for this relationship).

Extreme scores (outliers) can have dramatic effects on correlations

A single extreme score can dramatically change a correlation and affect the accuracy of the correlation. The 5 data points on the left show little relationship, but adding a 6th point at the high end of both variables produces a strong overall correlation.



It is important to observe points in a scatterplot for **outliers**. Often outliers represent erroneous or invalid data. If so, they should be omitted from further analysis. For example, there may have been a data entry error or the score may result from partially missing data. In such cases, the data do not contribute to our estimate of the relationship between the two variables but detract from it and therefore should be excluded.

Correlations describe how, but do NOT explain why, the variables are related

The basic underlying reason for this is that in a correlational study, we, the researchers, don't have control. That is, we are not manipulating one (or more) variable(s) while keeping everything else constant. As a result, we can't make causal claims.

Examples:

- Suppose that Dr. Steward finds that rates of spilled coffee and severity of plane turbulence are strongly positively correlated.

Correlationally speaking, one might argue that spilling coffee causes turbulence.

- Suppose that Dr. Cranium finds a positive correlation between head size and digit span (digit span is how many digits in order a person can repeat from memory).

Correlationally speaking, one might argue that people with bigger heads have bigger digit spans (instead of something like, head size and digit span increase with age).

- Suppose the Dr. Ruth finds a positive correlation between the number of babies born and the rate of stork sightings (I believe that such a correlation has been reported).

Correlationally speaking, one might interpret this as support for the hypothesis that storks bring babies to home.

Often what you may find is that there is **another variable Z, that causes both X and Y**, so X and Y may seem causally related, when they aren't.

Practice Questions: Set 8

(1) Create a scatterplot based on the following data:

Person	Height	Avg. Parents Height
A	65 in	68 in
B	60 in	64 in
C	69 in	70 in
D	59 in	65 in
E	72 in	67 in
F	67 in	65 in

(2) What is the direction of relationship (positive or negative) in this scatterplot? How strong of a relationship does there appear to be?

(3) For the data in Question 1, find \bar{X} , \bar{Y} , s_x , s_y , SS_x , SS_y , SP , and $\Sigma z_x z_y$.

(4) Now compute Pearson's Correlation Coefficient using the numbers found for SS_x , SS_y , SP , and $\Sigma z_x z_y$.

(5) Go back to the scatterplot and enter z-values on the X and Y axes.

(6) Interpret the r you just found. What is the direction and strength of the relationship? Does this match your interpretation based on the scatterplot?

ANSWERS P. 155

Study Guide for Exam 2

Terms

Correlation	Positive correlation
Cumulative percentage	Prediction
Extreme scores (outliers)	Range
Frequency	Scatterplot
Frequency distribution	Shape of distribution
Histogram	Standard deviation
Line graph	Standard scores
Linear relationship	Sum of squares
Negative correlation	Sum of products
Non-linear relationship	Symmetrical
Normal distribution	Transformations
Pearson's correlation coefficient	Unit normal table
Percentage	Variability
Percentile rank	Variance
	z-score

Formulas

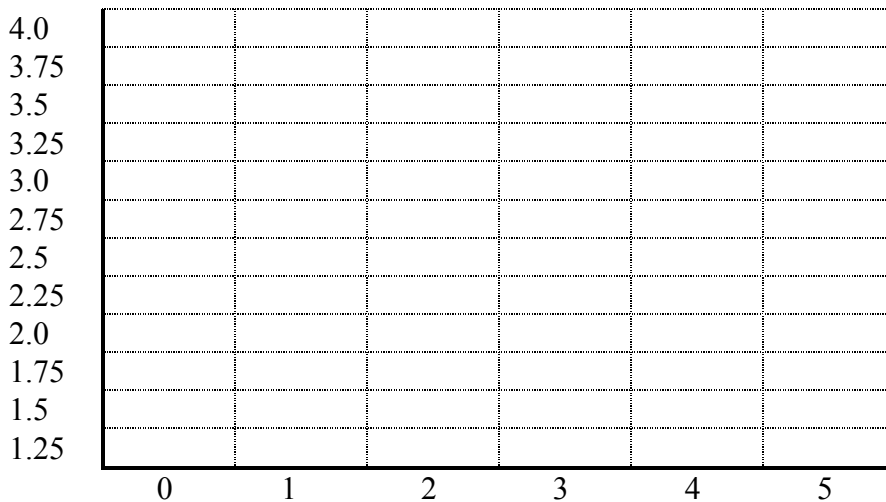
Statistic	For a population	For a sample
Variance	$\sigma^2 = \frac{SS}{N}$	$s^2 = \frac{SS}{n-1}$
Standard deviation	$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{SS}{N}}$	$s = \sqrt{s^2} = \sqrt{\frac{SS}{n-1}}$
Sum of squares	$\Sigma(X - \mu)^2$	$\Sigma(X - \bar{X})^2$
z score	$z = \frac{X - \mu}{\sigma}$	$z = \frac{X - \bar{X}}{s}$
Sum of Products		$\Sigma(X - \bar{X})(Y - \bar{Y})$
Pearson's correlation coefficient		$r = \frac{SP}{\sqrt{SS_x SS_y}} = \frac{\Sigma z_x z_y}{n-1}$

Sample Problem with Calculation Procedures

You conduct a survey on how much your friends like a website and whether it is related to their GPA. Your survey's response scale runs from 0 = "not like at all" to 5 = "absolutely love". Your assignment is to provide all descriptive statistics for the following dataset.

Person	Web liking	GPA	Person	Web liking	GPA
A	5	2.4	F	2	2.1
B	1	3.9	G	4	3.9
C	2	3.5	H	3	2.9
D	4	2.8	I	0	3.6
E	3	3.0	J	3	2.7

- a. **Make a scatterplot of both of your variables by entering the letter of each person on the proper place on the graph. Draw the best fitting line through the set of points.**



- b. **Complete a Bivariate Distribution and z Table (both deviations and z-scores) to get needed values in calculating measures of central tendency, variability, and correlation. (You don't have to do all of this this, just understand it. We'll do the calculations in Excel in lab.)**

Describe each abbreviation in the table:

$X =$

For what formula do you need $\sum(X)$?

$(X - \bar{X}) =$

What must the total be of $\sum (X - \bar{X})$?

$(X - \bar{X})^2 =$

What is $\sum (X - \bar{X})^2$ and where will you use it?

What is $(X - \bar{X})(Y - \bar{Y})$ and where will you use it?

What is $\Sigma z_x z_y$ and where will you use it?

X	$(X - \bar{X})$	$(X - \bar{X})^2$	z_x	Y	$(Y - \bar{Y})$	$(Y - \bar{Y})^2$	z_y	$(X - \bar{X})(Y - \bar{Y})$	$z_x z_y$

Σ
Mean

c. Find the following statistics for Xs and Ys. Show formulas and calculations.

Mode = X Y

Median =

Range =

M =

SS

Variance

SD =

d. Find the following statistics for Xs and Ys together. Show formulas and calculations.

r (SS formula) =

r (z formula) =

ANSWERS ON P. 157

III. Drawing Conclusions about Group Differences



To this point we've discussed the context from which we get our numbers, and a variety of methods that allow us to summarize and describe the distributions of these numbers. What we haven't discussed yet is how we interpret the numbers (and the summaries). In the final two sections of the course we're going to learn some statistical techniques for drawing conclusions from our data.

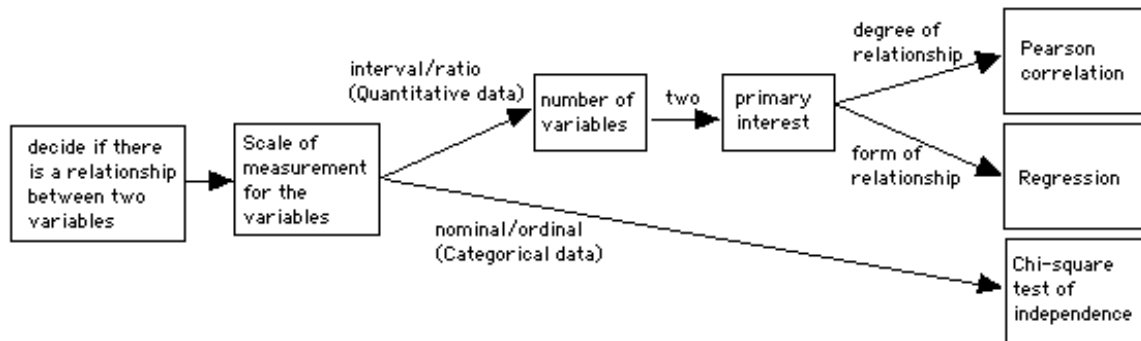
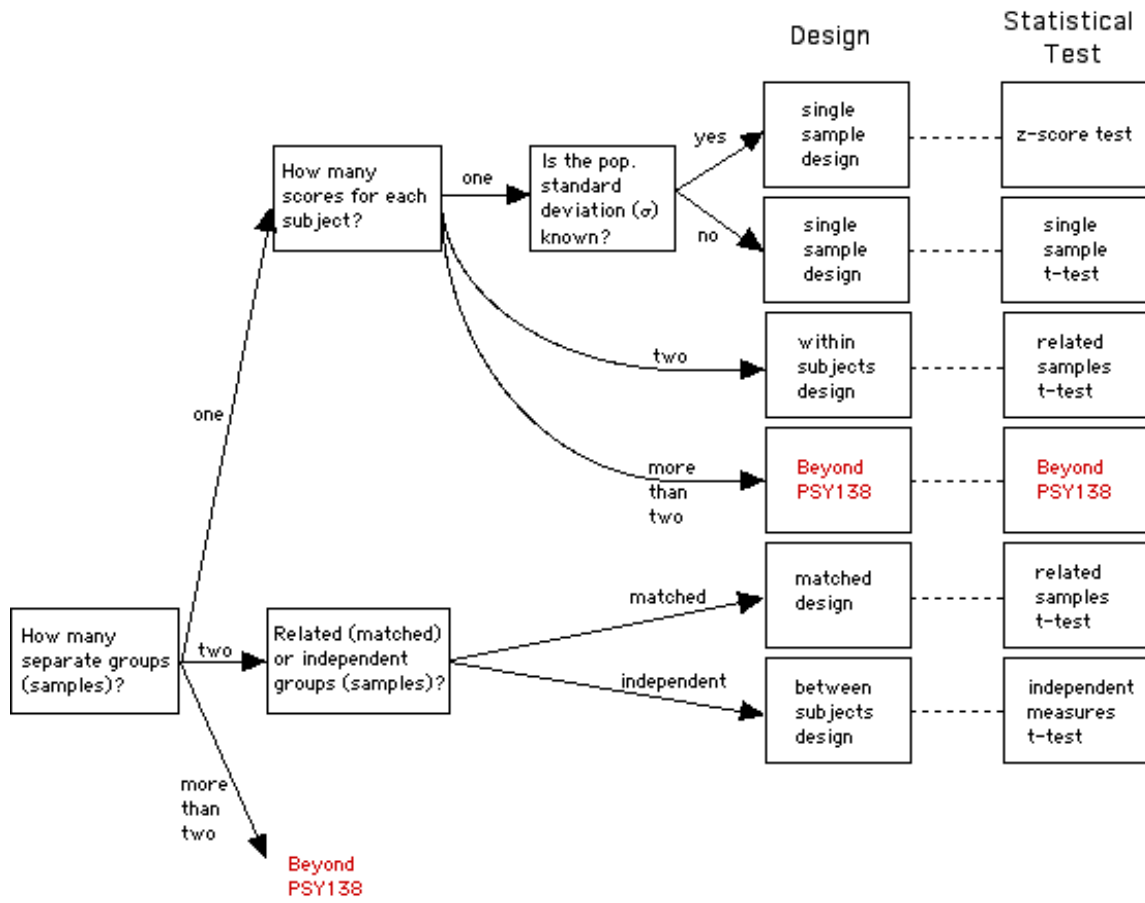
Nearly all of what we discuss in this portion of the course will involve a logical framework in which we ask and answer questions based on our data. Specifically, we're going to look at how we test **hypotheses** about populations using our samples and based on the likelihood of certain outcomes in given our sets of data. **The first two chapters in this unit (and accompanying classes) are the most important in the course; you must fully understand them or you will be lost for the rest of the course.**

This section will center on using a test statistic called the **t-statistic**. It, and the z-statistic, can be used to decide if there is a *difference between two groups*. In some cases, we compare a sample to a known population. In others, we compare two samples of scores or participants. We will see how the design of your research study determines how you calculate your t-statistic.

The final section will return to the topic of *relationship between two variables*, but this time within the framework of hypothesis testing. We will extend our discussion of correlation to how to test relationships found. We also will learn how to test relationships between categorical variables with a chi-square analysis.

The following decision tree includes on the right the hypothesis tests we will be covering. It includes the various features of research design that determine which test to use.

Decision Tree for Hypothesis Testing



Chapter 14: Hypothesis Testing in General

Hypothesis testing is an inferential procedure that uses sample data to evaluate the credibility of a hypothesis about a population. In other words, we want to be able to make claims about a population as a whole based on data that we collect from a single sample. We may find our sample to be different from the population, but is it different enough to be confident that we will get the same result from another sample? Might a difference we found just be by chance and so unlikely to occur repeatedly?

Let's work through an example. We can ask about the value of knowing about statistics. Suppose that we think that knowing about statistics helps people understand *USA Today*, the nationwide newspaper. We take a sample of students who have completed this class, and a sample of students from another class, none of whom have had a statistics course. To control motivation and interest in statistics, we only use students in majors that require statistics. Each person is given a copy of the paper, asked to read it, and is later tested for comprehension of the stories in the paper.

- What are the populations here?
 - 1) students who take statistics
 - 2) students who haven't yet taken statistics
- Our question: Are these two populations different? In other words, is there an effect of taking stats (on comprehending the paper)?
- Results of comprehension test (fictional):
 - 1) Mean for the statistics-class sample = 70% correct
 - 2) Mean for the no statistics-class-sample = 65% correct
- Problem: Is this 5% a "real" difference, or is it just due to sampling error. In the latter case, such a difference is likely to occur by chance, just as rolling 4 with dice isn't all that unlikely. We happened to draw a sample with a distance from the population mean that is small relative to the size of the standard deviation.
- If the difference is large enough to be very unlikely as a result of sampling error, then we can conclude that the two populations are different. This supports the hypothesis that statistics helps with reading the paper. If the difference is due to sampling error, then we should conclude that the populations are most likely the same, and that statistics knowledge does not help with understanding the newspaper.

Formalizing the Procedure of Hypothesis Testing

The advantage of formalizing is that if everyone uses the same techniques, then we know why they conclude what they do, and furthermore, we know about the **assumptions** that they are making. We will follow a **7-step procedure HC-STC-DC**. Make up your own mnemonic (memory aid) or use this one: **Hot Current-STeady Current-Direct Current**.

Step 1: *Hypotheses*

Our hypothesis is an educated guess/prediction about the effect of particular events/treatments/factors, which result in differences between populations. Our hypothesis may be general (e.g., this course will change comprehension abilities) or specific (e.g., this course will improve comprehension abilities by at least 10). We will learn about the logic of the null hypothesis.

Step 2: *Criterion for decision*

We need to decide how unlikely the difference we find would have to be by chance in order to accept it. We will select a small probability from a distribution, like $p \leq 0.05$; this is our criterion.

Step 3: *Sample statistics*

Then the experiment is done and data is collected from the sample on the measures being used. In this course, we won't actually collect data; instead, we will work with data supplied to us from imagined experiments. We will calculate the sample statistics that we have already learned: mean, sum of squares, and standard deviation. In addition, we will learn about statistics pertaining to the distribution of sample means.

Step 4: *Test statistic*

We will learn about a series of tests, each associated with a theoretical distribution: the z-test with the normal distribution, t-tests with the t distribution, and the χ^2 tests with the χ^2 distribution. The z- and t-tests compare the difference between our sample mean and the population mean to the difference expected by chance. We will learn a series of formulas and the circumstances that determine which test to use.

Step 5: *Compare observed to critical test value*

We already know how to use the Normal Unit table; we will learn how to look up values in the tables of t and χ^2 distributions. We will need to determine the degrees of freedom in the test in order to look up the correct critical test value in a distribution table.

Step 6: *Decide about null hypothesis*

If the observed test value is more extreme than the critical value, we have found a difference that is very unlikely by chance, so it must be “real.” Otherwise, we must decide that the difference found could have occurred by chance, that is, as a result of sampling error. We will learn how these findings relate to the null hypothesis.

Step 7: Conclude about relationship

Finally, we interpret how the findings relate to our original research question. Our experiment either provides support for the relationship between two variables hypothesized or it does not. If it does, we can calculate the size of the effect.

In the remainder of this chapter and the next one, we look at each of these steps in more detail.

Step1: Hypotheses

The logic that underlies hypothesis testing is that there are always (at least) two hypotheses: the *null hypothesis* and the *alternative hypothesis*

The *null hypothesis* (H_0) predicts that the independent variable (treatment) **has no effect** on the dependent variable for the population.

The *alternative hypothesis* (H_A) predicts that the independent variable **has an effect** on the dependent variable for the population. It is either directional or nondirectional.

The logic of hypothesis testing assumes that we are trying to *reject the null hypothesis*, **not** that we are trying to *prove the alternative hypothesis*.

Why? In everyday thinking, we talk of proving things to be true. But logical analysis demonstrates that to do so we would have to test every possible instance of our proposition and find the expected result in every single case. Otherwise, we are not proving all cases, but only making an unproven inference that that the cases we haven't tested are the same as the ones we have.

So, it is easier to prove that something is not true than to prove that it is. We just need to show that one instance is not true. We then can make an inference from the sample we have studied to the population that we have not. Our inference is a strong likelihood but not an absolute certainty. In science, we accept that this is the best we can do. Experientially we can only deal with **probable knowledge** (inference from sample studied to whole population); only ideally can we declare absolute truth (true of the whole population). So in scientific thinking, we start by stating a null hypothesis and

then try to reject it! The everyday ordinary hypothesis that there is a difference becomes our alternative hypothesis, which we can only indirectly support.

Example:

Null hypothesis: Taking a statistics course leads to no improvement in understanding *USA Today*.

Alternative hypothesis: Taking a statistics course improves understanding of *USA Today*.

To accept the alternative hypothesis: We need to test the whole population and find that every single student who took a statistics course improves in understanding *USA Today*. Furthermore, we would need to keep testing forever to be sure that every additional student who takes statistics improves in understanding *USA Today*. Explained in this detail, we see that this is obviously impossible. Our only justified conclusion would be to fail to accept the alternative hypothesis.

To reject the null hypothesis: Our sample must show so much improvement that we can conclude that we didn't just get lucky with this one group. Then we infer (but don't prove) that such improvement will occur in other samples drawn from the population of those who have taken statistics. In keeping with our probabilistic approach, we don't make the claim for every single individual (we know there is a lot of variability) but for every sample. When we reject the null hypothesis, we are claiming that almost all other samples of students who took statistics will improve in understanding *USA Today*. On the other hand, if our sample does not show enough improvement, we fail to reject the null hypothesis. Note that we haven't positively proven the null hypothesis; again, that would require testing everyone who ever took and will take statistics. Instead, we are claiming that other samples who have taken statistics would not show improvement in understanding *USA Today* either.

Note in the above example that the alternative hypothesis was directional. The statistics class must have a higher score in order to reject the null hypothesis. This is the most common situation; the experimenter predicts that the particular events/treatments/factors will result in a higher or lower score on the measure of interest in comparison to the score of the population in general. However, sometimes the experimenter is just looking to prove any effect of the particular events/treatments/factors. So, a difference in either direction is of interest and, if large enough, would support rejecting the null hypothesis.

Step 2: Criterion for decision

The next step is to set the criterion to use to either reject or fail to reject (remember, *not accept*) the null hypothesis. How unlikely does the effect have to be to claim it didn't occur by chance?

(Note that the word *criteria* is the plural of criterion, just as the word *data* is the plural of *datum*. These are Greek and Latin words, respectively, which accounts for the plurals

that are irregular by English standards. There are many such words throughout the disciplines of advanced learning, testifying to their origins and to the learning of these classical languages by scholars until the 20th century. Although we don't still learn those languages, we should use words from them accurately, and not say "a criteria" or "the data is.")

Consider the problem that we have. We have a sample, and its descriptive statistics are different from the population's parameters. How do we decide whether the difference that we observe is due to a "real" difference (which reflects a difference between two populations) or is due to sampling error?

To deal with this problem the researcher must **set a criterion in advance**. For example, think of the kinds of questions we were asking in the previous section. Given a population X with $\mu = 65$ and $\sigma = 10$, what is the probability that our sample (of size n) will have a mean of 80? We're going to be asking the same questions here, but taking it a step further and say things like, "The probability that my sample has a mean of 80 is 0.0002. That's pretty small. I'll bet that my sample isn't really from this population but is instead from another population."

Setting a criterion in advance is concerned with this part about saying "that's pretty small". When we set the criterion in advance, we are essentially saying how small a chance is small enough to reject the null hypothesis. Or in other words, how big a difference do I need to have to reject the null hypothesis. It is critical that this be done before running the study; we can't find a difference and then decide how big it needs to be. As we will discuss more below, the size of the difference also depends on whether our alternative hypothesis was directional or not.

There are various considerations that can influence the probability that is set. However, there often are conventional levels set within disciplines. For example, some fields may say that $p \leq 0.05$ is low enough to reject the H_0 , while other fields may chose $p \leq 0.01$ as the cut off. In psychology, $p \leq 0.05$ is commonly set as the criterion.

Errors in Decision Making

That's the big picture of setting the criterion; now let's look at the details:

What are the possible real world situations?

- H_0 is correct
- H_0 is wrong

What are the possible conclusions?

- H_0 is correct
- H_0 is wrong

So this sets up four possibilities (2 X 2):

- 2 ways of making mistakes
- 2 chances to be correct

Researchers' Decision-Making Table

Experimenter's Conclusion	Actual situation	
	H ₀ is correct	H ₀ is wrong
Reject H ₀	Oops! <i>Type I error</i>	Yay! <i>Correct</i> <i>(Statistical Power)</i>
Fail to reject H ₀	Yay! <i>Correct</i>	Oops! <i>Type II error</i>

The two kinds of error each have their own names, because they really are reflecting different things.

Type I error (α , alpha): H₀ is actually correct, but the experimenters rejected it. There really is only one population. Although the probability of getting a sample is really small, you just got one of those rare samples.

Type II error (β , beta): H₀ is really wrong, but the experimenters didn't reject it. Your sample really does come from another population, but your sample mean is so close to the original population mean that you can't rule out the possibility that there is only one population

Statistical Power: The *power* of a statistical test is its ability to detect these differences. Put in statistical terms, power is a statistic's ability to correctly reject the null hypothesis (Gravetter & Wallnau, 1996). A powerful statistic is more sensitive to true differences in your data than a less powerful statistic. Think of the power of a telescope. A telescope with low power will detect only the brightest stars; weak stars will remain unseen. Similarly, a test with low power will detect only large group differences and small ones will remain unseen. Power is defined as $1 - \beta$, when β is the probability of a Type II error.

A familiar example of this 2 X 2 decision-making matrix is in our justice system.

Courtroom Jury's Decision-Making Table

Jury's Conclusion	Actual situation	
	X is innocent	X is guilty
Guilty	Oops! <i>Type I error</i>	Yay! <i>Correct</i>
Not Guilty	Yay! <i>Correct</i>	Oops! <i>Type II error</i>

Notice that the initial hypothesis is that the accused is “innocent until proven guilty.” (This is not the case in other judicial systems, where instead of having the rights of a citizen who is accused of a crime, you are treated as a criminal who has been found out.) It is difficult to prove total innocence; after all, there is some evidence to support accusing the person. So, the jury has to prove “guilt beyond a reasonable doubt” if it is to reject the initial hypothesis that the person is innocent. Here the Type I error results in sending an innocent person to jail. A Type II error, by contrast, lets a guilty person go free. Because our society values freedom and seeks to avoid abuses of authority by the state, we set our criterion for the probability of mistakenly finding guilt very low (beyond a reasonable doubt). As a consequence, criminals are more likely to go free than in an authoritarian society, which is more concerned about keeping this probability very low even if innocents end up in jail.

In scientific research, we typically take a conservative approach, and set our criteria such that we try to minimize the chance of making a Type I error (concluding that there is an effect of something when there really isn't). In other words, scientists focus on setting an acceptable **alpha level** (α), or **level of significance**.

The **alpha level** (α), or **level of significance**, is a probability value that defines the very unlikely sample outcomes when the null hypothesis is true. Whenever an experiment produces very unlikely data (as defined by alpha), we will reject the null hypothesis. Thus, the alpha level also defines the probability of a Type I error, that is, the probability of rejecting H_0 when it is actually true. There is no *correct* alpha level, but it is typically small, such as 0.10, 0.05, and 0.01. **In psychology, α is usually set at 0.05.**

Step 3: Sample statistics

After we collect (or are given) data, we have to calculate statistics based on our sample. You may think that you can look at the difference between the sample mean and population mean in terms of the standard deviation, as we did with z-scores. But z-scores apply to the probability of *individual* scores in a distribution of *individual* scores. Consider the probability of getting a SAT score of 550; we know that the standard deviation is 100, so its z-score is 0.5, a pretty common score. The Unit Normal table shows that we expect to get a score that high or higher 30% of the time; it's at the 70th percentile. But what if we have a class with a *mean* SAT score of 550; how likely is that? Does that seem less likely than an individual with that score? Think of an individual of a sample of one. What happens as we increase sample size: from a class of 25 to 100 students? Now we are asking about the probability of a sample *mean*. So we need to consider a **distribution of sample means (or DSM)**, that is, a distribution where each score represents the mean of a sample (rather than a score of individual).

Distribution of Sample Means (DSM)

For a distribution of sample means, we should be able to find its mean and a measure of average dispersion. It turns out that the mean of the population of individual scores and of sample means is the same, so we don't need different terminology. But the measure of

average dispersion is different, so we will use another term, **standard error** of DSM. Note for individual scores, we refer to their *deviation* from the mean relative to the standard deviation; there is no error involved, just difference from the group average. However, the difference between a sample mean and the population mean is an *error*, specifically, an error of sampling. The size of this error relative to the standard error is what will enable us to decide if our sample is really different (from a different population) rather than just a sample drawn randomly from the original population.

What is the shape of DSM? If the population of individual scores is normal, so will the DSM. However, if the basic population is not normal, as samples get large ($n > 30$), DSM approaches a normal population. So with large enough sample sizes, we can conduct inferential statistics even on skewed populations.

How do we calculate the standard error? If we know the population standard deviation, we simply take it and divide by the square root of the sample size. **Note the subscript in the formula for the standard error to indicate that it refers to the DSM**; without the subscript we just have the standard deviation of a population of individual scores.

Standard error (σ known):
$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

For given size samples, we can determine the probability of getting certain means or of getting means the same or greater/less than certain means. In class, we will go through an example. Note from the formula that when $n = 1$, $\sigma_{\bar{X}} = \sigma$. What happens as n gets bigger? When a denominator gets larger relative to the numerator, the value of the fraction gets smaller ($1/8$ of a pie is less than $1/2$ of a pie). As n gets larger, $\sigma_{\bar{X}}$ gets smaller relative to σ , that is, the average dispersion in DSM gets smaller. And since this statistic is in the denominator of test-values, its getting smaller means the test-value will get bigger, that is, less likely to occur by chance.

Here is how this plays out for SAT scores. When $n = 1$, $\sigma_{\bar{X}} = \sigma = 100$, that is, 100 is the standard error from the population mean when you draw samples of just one person. But when $n = 25$, $\sigma_{\bar{X}} = 20$, that is, the standard error from the population mean is only $1/5$ as large when you draw samples of 25 persons. And when $n = 100$, $\sigma_{\bar{X}} = 10$, cutting the standard error in half from the sample of 25.

An individual with a score of 550 ($z = 0.50$) is at the 70th percentile. A class of 25 with a mean of 550 ($z_{\bar{X}} = 2.5$) is at the 0.6th percentile. A class of 100 with a mean of 550 ($z_{\bar{X}} = 5$) is beyond the 0.01st percentile. So, an individual score of 550 is not unusual, but a class of 25 with a mean of 550 would occur by sampling from the overall population less than 1/100 times; for a class of 100, it would occur less than 1/1,000. We could be confident that there really is something different about these classes; they are very unlikely to be random samples of SAT takers.

We have just run z-tests! The next chapter will go through all of this in greater detail. The point being made here is that if we know the population mean, the sample mean, and the standard error, the logic for finding an individual z-score applies to finding the z-value for a sample mean. We then can look in the table of the Unit Normal distribution and see what the probability of that z-value is. If it is more extreme than α , then we can reject the null hypothesis; if it is not more extreme, then we fail to reject the null hypothesis. We will go through these remaining steps of hypothesis testing in the next chapter.

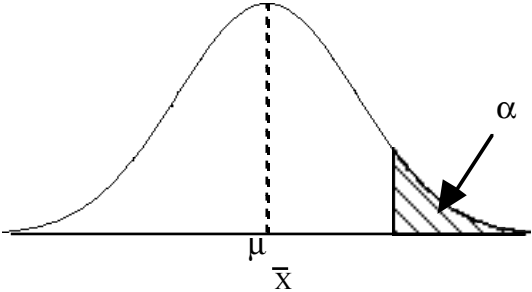
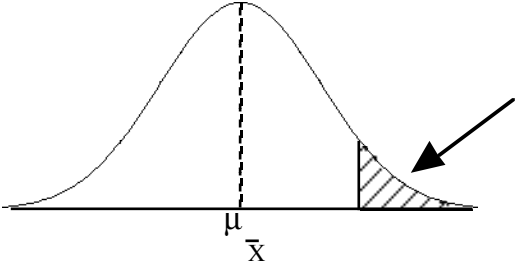
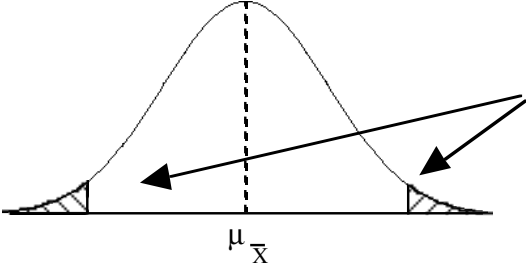
Chapter 15: Hypothesis Testing with z-tests

Remaining Steps

4 & 5: Test statistic & Compare observed to critical test values

6 & 7: Decide about null hypothesis & Conclude about relationship

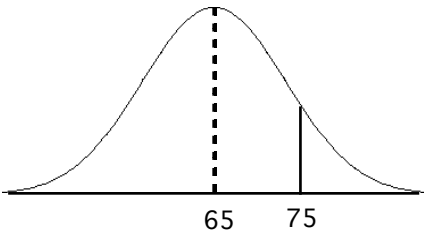
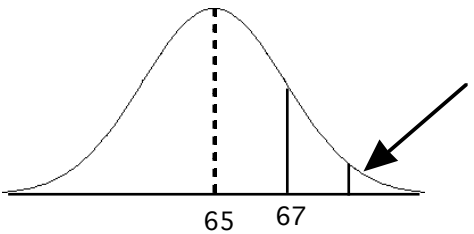
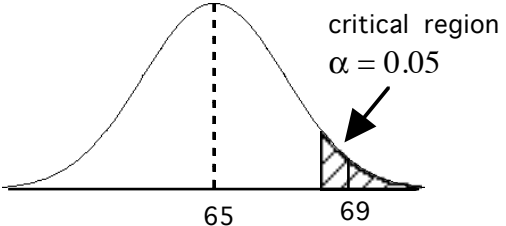
We are ready to look at pictures of distributions to try and connect them to what we've been discussing. Let's work through a series of population distributions and consider how different our sample mean (\bar{X}) has to be from the mean of the population distribution of sample means ($\mu_{\bar{X}}$) to be confident it is not just by chance.

	<p>α = probability of making a type I error</p>
	<p>Specific alternative hypothesis</p> <p>H_0: No difference H_A: Difference: New group has higher mean Test: 1-tailed $\alpha = 0.05$, all in 1 tail</p>
	<p>General alternative hypothesis</p> <p>H_0: No difference H_A: Difference: New group has higher or lower mean Test: 2-tailed $\alpha = 0.05$, 0.025 in each of 2 tails</p>

How do we interpret these graphs? If our sample mean falls in the shaded areas then we reject the H_0 . On the other hand, if our sample mean falls outside of the shaded areas, then we may not reject the H_0 . **The shaded regions are called the *critical regions*.**

A **critical region** is composed of extreme sample values that are very unlikely to be obtained if the null hypothesis is true. Thus, the more extreme the test statistic is (either very negative or very positive), the more likely it is to fall in a critical region. **The size of the critical region is determined by the alpha level.** Sample data that fall in the critical region will warrant the rejection of the null hypothesis. We can say that the difference we found is **statistically significant**, that is, that it is unlikely to have occurred by chance (sampling error) at our predetermined level (less than 5 times out of 100).

Example, one-tailed

<p>Population distribution</p> 	<p>$\mu = 65$ and $\sigma = 10$ (note line for 1 SD) Suppose a sample of $n = 25$ receives an experimental treatment with $\bar{x} = 69$. Did the treatment work? Let's assume an $\alpha = 0.05$ and that our alternative hypothesis is that the treatment improves performance (makes the mean higher).</p> <p>Will it work if we try another sample? What if we take sample after sample; how likely is $\bar{x} = 69$?</p>
<p>Distribution of sample means</p> 	<p>Find our sample mean in the distribution of sample means. We need to determine the probability of getting that mean or higher for the sample.</p> <p>How big a deviation is a sample mean that is $\mu + 4$? We need to compare 4 to the standard error--the standard deviation of sample means. Solving for it finds that $\sigma_{\bar{x}} = 10/\sqrt{25} = 10/5 = 2$. Note the line for 1 SE and how much smaller $\sigma_{\bar{x}}$ for a sample of 25 is than σ. Our sample mean is 2 SEs above the population mean. Is that in the critical region?</p>
	<p>What is the critical region?</p> <ul style="list-style-type: none"> • 1-tailed test, $\alpha = 0.05$ • In the Unit Normal table, find the area that corresponds to α. • $z = +1.65$, the critical test value • The critical region contains any z-value $+1.65$ or greater.

We already knew how to find the z-score corresponding to the critical region when the **criterion** is $\alpha = 0.05$; it is +1.65 (the sign of the value is important; it indicates if it is the upper or lower tail). This is the **critical value** of the test statistic. Now we need to use a new z formula for sample means to find our **observed test value**.

Below are the formulas for the standard error, which we already found, and z for the sample mean.

Standard error (σ known): $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

z observed

$$z_{\bar{X}} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

To find z observed for the sample mean in our example, we enter our statistics:

$$z_{\bar{X}} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{69 - 65}{2} = 2$$

We find that observed value of z is 2, while the critical value is 1.65, that is,

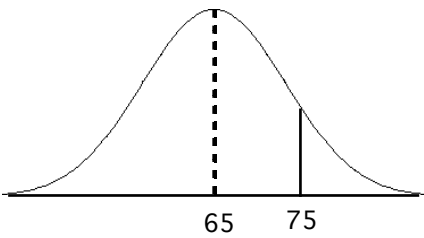
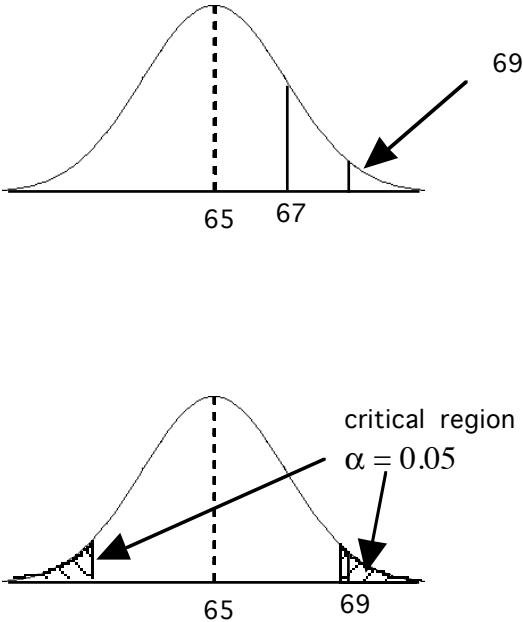
$$|z_{\text{observed}}| \geq |z_{\text{critical}}|$$

Note that we used the bars to indicate **absolute value**. Sometimes we may omit these when working only with positive numbers, but they indicate that it is the absolute relative sizes of test statistics that are at issue. If we had predicted a lower score in this example and found a sample mean of 61, the observed z would be -2 and the critical z-value would be -1.65. -2 is smaller, not greater, than -1.65; however, |-2| is greater than |-1.65|. Remember, we are interested in observed test values in the tails, that is, more extreme values, either positive or negative, than critical values.

Given that we found a more extreme test value than the critical value, our decision is to reject H_0 and our conclusion is that the treatment works to improve performance.

For the example that we just did, we made a hypothesis that the treatment would make a difference in a specific direction (that is, the treatment would increase the mean). However, a nondirectional hypothesis can be tested, namely, that the treatment will change the mean, either increase or decrease it. As we saw above, this will affect our critical region. In a 1-tailed test, it is all in the upper or lower tail, depending on whether we predicted the treatment to increase or decrease the score from the population mean. In a 2-tailed test, we must divide the critical region into the two tails of the distribution.

Example, two-tailed

<p>Population distribution</p>  <p>A normal distribution curve with a dashed vertical line at the mean $\mu = 65$ and a solid vertical line at 75.</p>	<p>Same as above: $\mu = 65$, $\sigma = 10$ (note line for 1 SD) sample of $n = 25$, treatment $\bar{X} = 69$.</p> <p>Did the treatment work? Does it affect the population of individuals?</p>
<p>Distribution of sample means</p>  <p>Two normal distribution curves. The top curve has a dashed vertical line at 65 and a solid vertical line at 67. The bottom curve has a dashed vertical line at 65 and a solid vertical line at 69. The tails of the bottom curve are shaded and labeled "critical region $\alpha = 0.05$".</p>	<p>Find our sample mean in the distribution of sample means. To determine the probability of getting that mean or higher for the sample, we need to solve for the Standard Error $\sigma_{\bar{X}}$, that is, the standard deviation of sample means: $\sigma_{\bar{X}} = 10/\sqrt{25} = 10/5 = 2$. Note the line for 1 SE. Our sample mean is 2 SEs above the population mean. Is that in the critical region?</p> <p>What is the critical region?</p> <ul style="list-style-type: none"> • 2-tailed test, $\alpha = 0.05$, so 0.025 in each tail • In the Unit Normal table, find the area that corresponds to $\alpha = 0.025$. • $z = \pm 1.96$, the critical test value <p>The critical region contains any z-value $+1.96$ or greater or -1.96 or smaller.</p>

The observed z we calculated above still applies. The kind of hypothesis we make has no affect on it. However, our comparison is different because we have a different critical z-value. Here we are comparing our observed z of 2 to a critical z-value of ± 1.96 . Again, our z is greater (in the upper tail), and again we need to indicate a comparison between **absolute values** because a more extreme z-value in either direction is in the critical region.

$$|z_{\text{observed}}| \geq |z_{\text{critical}}|$$

Thus, our decision again is to reject H_0 and our conclusion is that the treatment works to improve performance. The sample mean was large enough to be significantly different from the population mean for a directional or nondirectional hypothesis. However, it takes a larger difference to reach significance for a nondirectional hypothesis, since α is divided between two tails.

Consider what would happen in our example if the sample mean had been 68.5. The observed z would be 3.5 (difference between sample and population mean) divided by 2 (standard error) or 1.75. This is large enough to be in the critical region for a 1-tailed test (critical $z = +1.65$) but not a 2-tailed test (critical $z = \pm 1.96$). Consider the decision-making table from the last chapter. If the treatment really does improve performance, but we only made a nondirectional hypothesis, we would decide that it made no difference, a Type II error. If we expect an effect in one direction, we should make a directional prediction because it enables us to conduct a more powerful test of the hypothesis.

General Formula for Hypothesis Testing

We now want to consider the general form of the statistic used to test hypotheses. Below is the z-test formula and next to it the general form of all test statistics.

$$z_{\bar{X}} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \qquad \text{test statistic} = \frac{\text{observed difference}}{\text{difference expected by chance}}$$

The statistic is testing how likely the sample is to have been drawn from the population. The smaller the difference between the sample and population means, the smaller the test statistic. (As a numerator gets smaller, the value of the fraction gets smaller-- $\frac{3}{8}$ of a pie versus $\frac{1}{8}$ of a pie.) When the observed test value shrinks to zero, the test statistic equals zero. Remember that by definition a z-score of zero is the mean of a distribution. So a sample mean equal to the population mean results in a z-value of 0, and the probability of getting a mean of the size or greater is 50%. Obviously, we could not conclude that our sample is from a different population.

As the observed difference gets larger, the value of the test statistic gets larger. As with the distribution of z-scores, the larger the value (the further from the mean of 0), the lower the probability. According to the Unit Normal table, a sample mean large enough to produce a z-score of 2 is likely to occur less than 5 times out of 100 samples drawn, so we can conclude that the sample is from a different population.

Now, consider the denominator of the formula. It is a measure of the variability of the distribution of sample means, which is derived from the variability of the population. The smaller the variability, the smaller an observed difference is needed to be significant. For a steep, narrow distribution, a small difference from the population mean is significant; for a wide distribution, a much larger difference is needed. Since the variability is in the denominator, the value of the fraction changes in the opposite direction: smaller variability produces a larger test value, and larger variability produces a smaller test value. If the observed difference between sample and population means is 5, then variability of 2 will produce a test value of 2.5; for a z-score, the probability is less than 1/100. But if the observed difference and variability are both 5, then the test value is 1, a z-score of 0.16, indicating a more likely sample by chance. That is, you would expect to draw a sample with this mean or larger 16 out of 100 times, so you would reject the null hypothesis if $\alpha = 0.05$ so as to avoid the likelihood of a Type I error (claiming there is a difference when there really isn't one).

Effect Size

We need to add a caution about sample size and statistical significance. The larger the sample size, the smaller the standard error, so the smaller an observed difference needed to be significant. However, this also means that with very large sample sizes, even if the effect is really small, we're more likely to reject the null and decide there's an effect. Therefore, in the case of very large samples, we may detect effects that lack *practical* significance because they are so small they may not be important. For example, with very large census or poll samples, differences will occur on many variables, but may be of little practical significance. However, in basic scientific research, identifying a small but reliable difference may have great significance. We must be careful that when we find statistical significance that our findings also have practical significance, meaning the effect of the treatment is important.

As a way of evaluating the size of an effect irrespective of sample size, we can use Cohen's *d*. (There are other statistics that can be used for effect size, but they will not be covered in this course.) It provides the distance between the two means (the observed distance) relative to the **standard deviation** of the population (rather than the standard error, which is influenced by sample size). Here is the formula for z-tests, when we know the population standard deviation.

$$d = \frac{\bar{X} - \mu}{\sigma}$$

Cohen provided guidelines about effect size and these are generally recognized as conventions, much as $\alpha = .05$. **A small effect size is .20; a medium effect size is .50; and a large effect size is .80. Effect sizes > 1 are considered very large.** Effect sizes are absolute values, so the observed difference can be in either direction. An effect size of 1 indicates that the sample mean is 1 SD from the population mean. An effect size of .2 indicates that the sample mean is only 20% of a SD from the population mean.

Assumptions of Hypothesis Testing

- 1) *Random sample* - the samples must be representative of the populations. Random sampling helps to ensure the representativeness.
- 2) *Independent observations* - also related to the representativeness issue, each observation should be independent of all of the other observations. That is, the probability of a particular observation happening should remain constant and not be affected by other observations.
- 3) *σ is known and is constant* - the standard deviation of the original population must stay constant. Why? More generally, the treatment is assumed to be adding (or subtracting) a constant from every individual in the population. So the mean of that population may change as a result of the treatment, however, recall that adding (or subtracting) a constant from every individual does not change the standard deviation.
- 4) *The sampling distribution is relatively normal* - either because the distribution of the raw observations is relatively normal, or because of the Central Limit Theorem (or both).

Violations of any of these assumptions will severely compromise any conclusions that you make about the population based on your sample. (Various adjustments can be made or other kinds of inferential statistics used to deal with violations of various assumptions. Most of these are beyond this course.)

The framework that we have just discussed can be used for a wide variety of different situations. The rest of the course will cover some of the situations for which hypothesis testing is done with statistics. In each case we will follow the same basic steps of the framework, but the details of these steps vary depending upon the design of the research. The design determines the nature of the hypotheses, the way in which the test statistic is calculated, and ultimately how the conclusions are drawn. To assist with the decision making about which statistical test to use, we refer you to the decision tree presented at the beginning of this section (page 73). The tree provides a series of questions, for which the answers will guide you to the appropriate statistical test.

Practice Questions: Set 9

- (1) Discuss the errors that can be made in hypothesis testing (Types I and II).
- (2) Scores on the SAT test are Normally distributed with a $\mu = 500$, and a $\sigma = 100$. Dr. Ed Standards, the local district school superintendent, develops a new program that he believes should increase SAT scores for students. He selects 25 local high school students to take the program and then take the SAT test. His sample has an average SAT score of 559. Conduct a hypothesis test to determine whether this program works. Show all of your steps and state all of your assumptions.
- (3) Suppose that the school board tells Dr. Standards that the new program is too expensive to pilot on 25 students and asks that he reduce his sample size to 9 students. Assume that same properties for the population of SAT scores. Suppose that his sample of 9 students also has a mean score of 559. How does this reduction in sample size affect Dr. Standards' hypothesis test?

ANSWERS ON P. 159

Chapter 16: One-Sample t-test

The formulas

The one-sample t-test allows us to extend our hypothesis testing procedure to cases where we don't know the population standard deviation σ . Without the population standard deviation σ , we can't directly calculate the standard error $\sigma_{\bar{X}}$ (as we did for the z-test). In this case, we will take our *best guess* at what σ might be. That value is the sample standard deviation s , because without knowing the values in the entire population, our best guess at what the value is comes from the sample we are testing. That's what the sample statistics are supposed to be doing – representing the population values. So here we'll let the sample statistic s represent the population parameter σ in our test. And instead of calculating the standard error directly with σ , we'll calculate the estimated standard error, using the Roman alphabet to indicate that it is estimated from a sample: $s_{\bar{X}}$.

Recall the two relevant formulas for a one-sample z-test:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \qquad z_{\bar{X}} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

To get one-sample t-test formulas, we substitute s and $s_{\bar{X}}$ for σ and $\sigma_{\bar{X}}$. To solve for s , we need to recall the formulas for the standard deviation of a sample and degrees of freedom. That gives us these four formulas, two old and two new:

$$s = \sqrt{s^2} = \sqrt{\frac{SS}{n-1}}$$
$$df = n - 1$$

Standard error σ unknown: $s_{\bar{X}} = \frac{s}{\sqrt{n}}$

One sample t-observed: $t = \frac{\bar{X} - \mu}{s_{\bar{X}}}$

Effect size: $d = \frac{\bar{X} - \mu}{s}$

We use the t-test the same way we use the one-sample z-test (i.e., when we want to compare a treated sample with a known population), but in cases where we only know the population μ and not the population standard deviation σ .

Rule: When you know the value of σ , use a z-statistic. If σ is unknown, use s to estimate σ and use the t-statistic. The same is true for making estimations of the population mean (e.g., confidence intervals) that we will cover later.

Even though the formulas in the two situations are very similar, there is an important conceptual difference between the two situations. Because we are using the sample standard deviation s to **estimate** the population standard deviation σ , we need to take into account the fact that it is an estimate. This means that we must take the **degrees of freedom** into account because we are using a sample that is less variable than a population (we covered this already when we talked about variability). The formula for s uses $n - 1$ rather than n in the denominator because we have one fewer value free to vary. These are our degrees of freedom.

Degrees of freedom describe the number of scores in a sample that are free to vary. Because the sample mean places a restriction on the value of one score in the sample, there are $n - 1$ degrees of freedom for the sample.

This means that the higher the value of n , the more representative the sample will be of the population, which in turn means that s will be a better estimate of σ . It also has implications for the test statistic. The shape of the t-distribution varies as a function of the size of n (really it varies with the degrees of freedom). The bigger the n (the bigger the df), the closer the t-distribution is to a normal distribution.

Notice that we're talking about a new distribution here (or family of distributions, the t-distributions). This also means that we **won't** be using the Unit Normal table. Instead we'll have to use a different table, the t-distribution table. It is organized by both p values and by degrees of freedom (df). A copy of the t-table can be found at the end of this packet. Reading this table is different than reading the unit normal table. So let's talk about why first, and then how.

The answer to the first question is because the Unit Normal table is describing just one distribution--the normal distribution. The t-distribution table is actually describing several different t-distributions. This is because there is a different t-distribution for every different *degree of freedom* (although when df gets large, the differences become really small). So, each row corresponds to a different t-distribution. As a result of this, there also isn't enough space to put all of the probabilities corresponding to each possible t-score. Instead, what are listed are the t-scores at commonly used critical values (that is, at popular alpha levels).

The t-distribution, with infinite (or practically very large) dfs is equal to the normal distribution. That's why along the bottom of the table is a row of z-scores. The bottom

row of the table tells you which column to look in for confidence intervals. With smaller *dfs* the t distributions are shaped differently (although they are still unimodal and symmetrical with a mean = 0).

7-Step Procedure

We are ready to go through our 7-step procedure for hypothesis testing using a one-sample t-test. Remember from the decision tree (p. 73) that the research design involves one score per participant, no known population variance, and one sample.

Step 1: Hypotheses

Step 2: Criterion for decision

Step 3: Sample statistics

Step 4: Test statistic

Step 5: Compare observed to critical test score

Step 6: Decide about null hypothesis

Step 7: Conclude about relationship

Example: Suppose that your psychology professor, Dr. I. D. Ego, gave a 20 point true-false quiz to 9 students and wanted to know if they were different from groups in the past who have tended to have an average of 9.0. The scores from the current group were: 6, 7, 7, 8, 8, 8, 9, 9, 10. Did the current group perform differently from those in the past? Assume a significance level of $\alpha = 0.05$.

Step 1:

$H_0: \mu = 9.0$ and

$H_A: \mu \neq 9.0$

Note that the alternative hypothesis is nondirectional (asking if they are different, not specifically whether they are better or worse). Therefore, we will be conducting a two-tailed test.

Step 2:

$\alpha = 0.05$

Step 3:

$$\bar{X} = \frac{\sum X}{n} = 72/9 = 8$$

$$SS = \sum(X - \bar{X})^2 = 12$$

$$s = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{12}{8}} = \sqrt{1.5} = 1.225$$

$$n = 9, \text{ so } df = 9 - 1 = 8$$

Step 4:

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{1.225}{\sqrt{9}} = \frac{1.225}{3} = 0.41$$

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}} = \frac{8 - 9}{0.41} = -2.44$$

Step 5:

Find the critical t from the table: $df = 8$, 2-tailed, $\alpha = 0.05$
t-critical = ± 2.306 and t-observed = -2.44

$$|t_{\text{observed}}| \geq |t_{\text{critical}}| \text{ (observed t is more extreme)}$$

Step 6:

Reject the null hypothesis.

Step 7:

It looks as if the current students are different from past students. They are doing worse; we know this because the sample mean of 8 was lower than the population comparison mean of 9 and our test tells us this difference is large enough to be significant, that is, unlikely to have occurred because of sampling error. When we calculate the effect size, we find $d = \frac{\bar{X} - \mu}{\sigma} = \frac{8 - 9}{1.225} = 0.82$ that it is considered to be large, almost a full standard deviation lower than the population mean.

The practice questions and lab problems provide more examples for you to work through.

Practice Questions: Set 10

- (1) Suppose that your psychology professor, Dr. I. D. Ego, wants to evaluate people's driving ability after 24 hours of sleep deprivation. So she develops a test of driving skill (scores ranging from 1-bad driving to 10-excellent driving) and administers it to 101 drivers who have been paid to stay awake for 24 hrs. The scores from the group had a mean of 4.5 and a standard deviation of 1.6. Determine if the sleep-deprived group mean is significantly different from the known population mean of 5.8 for the driving test. Assume $\alpha = .05$.
- (2) Several years ago a school survey revealed that the average age at which students first tried an alcoholic beverage was $\mu = 14$ years. To determine if anything has changed, a random sample of 5 students was asked questions about alcohol use. The age at which drinking first began was reported as 11, 13, 14, 12, 10. Use these data to determine if there has been a change in the age at which drinking began. Use $\alpha = .05$.
- (3) A random sample of $n = 16$ scores has $M = 48$. Use this sample ($\alpha = .05$) to determine if the sample is different from the population with $\mu = 45$ for each of the following situations:
 - (a) Sample $SS = 60$.
 - (b) Sample $SS = 600$.
 - (c) How does the sample variability contribute to the outcome of the test?
- (4) A national company is attempting to determine if they need to hire more employees. One thing they are basing this decision on is the number of hours per week their current employees work. They collect a sample of average hours worked per week from 30 employees to compare with the national full-time work standard of 40 hours per week. The mean number of hours worked for their sample is 47.8 with $SS = 1020$. Using $\alpha = .05$, conduct a test to determine if this company's employees work more hours per week than the national standard.

ANSWERS ON P. 160

Chapter 17: Related-Samples t-test

A simple extension of the one-sample t-test can be made to handle research designs with related samples. Related samples are used when one group of subjects is tested more than once (e.g., before and after a treatment), called a *within-subjects or repeated measures design*, or when two groups of related subjects are used (e.g., twins, matched subjects). The reason researchers use related samples is that they increase the internal validity of the study. By testing subjects twice or matching subjects on some characteristic, you can reduce the variability due to confounding variables. This in turn will increase the power of your significance test. Go to page 73 to find these two designs that use the related-samples t-test in the decision tree.

Recall that the formula for a one-sample t-test is

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}}$$

We will use the same formula for the related-samples t-test, but the values will come from the set of difference scores (\bar{D}). In other words, we will determine the difference in the scores for each subject (or pair of subjects for matched samples) and conduct the test on these difference scores. The population mean if there are no differences is 0. So for the related samples test we have the formula

$$\text{Related-Samples observed } t: t_{\bar{D}} = \frac{\bar{D} - \mu_{\bar{D}}}{s_{\bar{D}}}$$

$$\text{Effect size: } d = \frac{\bar{D}}{s_D}$$

Let's look at an example of how this works:

Suppose we were interested in examining the effectiveness of a new therapy in reducing depression. We conduct a study to compare depression scores of clinic patients before and after the new therapy to determine the effect of the therapy. The depression score for each research participant is listed below (higher scores mean more depression).

Person	Before Therapy	After Therapy
A	98	90
B	72	72
C	86	75
D	90	86
E	77	78
F	65	63

The first step is to calculate the difference scores for each pair. We'll subtract the "after" score from the "before" score for each subject. *A positive difference means that the depression score has gone down, which is a positive outcome.* The set of difference scores is listed below.

Person	Difference Score
A	8
B	0
C	11
D	4
E	-1
F	2

We have been able to reduce our two samples to one sample of difference scores, and we can conduct our test as we did before. It is helpful to draw a distribution with the critical region marked. In this case, the upper tail is the critical region because a very positive difference score indicates successful treatment. We need the mean of the difference scores: \bar{D} .

$$\bar{D} = (8+0+11+4-1+2) / 6 = 4$$

The other statistics we need are analogous to those for the one-sample test:

Standard deviation of the differences: $s_D = \sqrt{\frac{SS_D}{n_D - 1}}$

Standard error of differences: $s_{\bar{D}} = \frac{s_D}{\sqrt{n_D}}$

$(D - \bar{D})$	$(D - \bar{D})^2$
4	16
-4	16
7	49
0	0
-5	25
-2	4

$$\Sigma 0 \quad SS_D = 110$$

$$s_D = \sqrt{110/5} = \sqrt{22} = 4.69$$

$$s_{\bar{D}} = 4.69/\sqrt{6} = 1.91$$

Now we can calculate the observed t-value.

$$t = \frac{4 - 0}{1.91} = 2.09$$

We use $\mu_D = 0$ because if there is no effect in the population, the average difference score should be 0 (indicating no difference between the scores). Before we can determine if this is a significant value, we have to calculate our degrees of freedom (n-1). Remember that we are testing the difference scores (n = 6), so we have 6 - 1 = 5 degrees of freedom (df = 5). Remember that we are conducting a one-tailed test.

Now we can use our t-table to find the critical value for $\alpha = .05$ and $df = 5$. The value is 2.015. When we compare this with our calculated t of 2.09, we find that our test value IS significant, because the observed t from the sample falls in the critical region, that is, it is more extreme than the critical value. Thus, we can conclude that the therapy is effective in reducing depression.

Now we can calculate the effect size.

$$d = \frac{\bar{D}}{s_D} = \frac{4}{4.69} = .85$$

It turns out to be a large effect. The posttreatment mean is almost a full standard deviation below the population mean of no differences.

Practice Questions: Set 11

- (1) For the sample difference scores below, determine if the sample differs from $m_D = 0$. Use $\alpha = .01$.

Difference scores (D): 4, 5, 4, 2, 4, 5, 3, 5, 4

- (2) A researcher was interested in the environmental effects on handedness. He measured the handedness of twins raised apart, where a positive score indicates more right-handedness and a negative score indicates more left-handedness (a score of 0 means the subject is ambidextrous). He used matched pairs of identical twins as subjects to rule out any genetic contribution to handedness scores (identical twins are the same genetically). The scores for each pair of twins are listed below. Use these data to determine if the twins differ in handedness score (indicating that environment plays a role in handedness). Use $\alpha = .05$. (Hint: A related-samples t-test is appropriate here.)

Handedness Score		
Pair	Twin A	Twin B
1	+10	+11
2	- 8	+ 3
3	-11	+11
4	+15	+10
5	0	+ 8
6	- 4	+ 7

- (3) Each of the following sets of sample statistics comes from a within-subjects design.

Set 1: $n = 10$, $\bar{D} = +4.0$, $s = 10$

Set 2: $n = 10$, $\bar{D} = +4.0$, $s = 2$

Find t-values. Even without looking up the critical t, for which set is it more likely to reject the H_0 indicating that the $\mu_D = 0$? Why? (Hint: calculate effect sizes.)

ANSWERS ON P. 162

Chapter 18: Independent-Samples t-test

This test is used in our final (and most common) design to test hypotheses about the differences between groups. In the design discussed in the last chapter, there were two samples but they were related. We were able to reduce them to a single set of difference scores because there were pairs of related scores. In the independent-samples design, we gather data from two unrelated samples and for the first time we have to conduct a test on two sets of scores. Because of this complication, the statistics are more complex and are not usually generated by hand but by SPSS. We are asking whether the sample means are likely to have come from a single population (null hypothesis) or from two different populations. We call these studies *between-subjects designs*, because the primary comparison of interest is made between groups of participants, who were formerly referred to as subjects. The logic of our test is similar to that of the previous t-tests we looked at, but the calculations get more complicated. We'll take it step by step.

Step 1: Hypotheses

The hypotheses are going to be a different because the situation is different. Remember that now we are making hypotheses about two different populations. For example, suppose that you want to compare two different treatments (e.g., two ways of studying, two different drugs, etc), or you want to compare two groups of people (e.g., men vs. women, young vs. old, etc.). So now, the hypotheses are about population A (women) and population B (men), and how they are different from one another.

Suppose that we are interested in how tall women and men are. We're looking at two populations here, A and B, where population A is the heights for women and population B is the heights for men.

$$\mu_A \quad \mu_B$$

So the H_0 hypothesis would be that women and men are the same height, or that there is no difference between the heights of women and men. That is,

$$H_0: \mu_A = \mu_B$$

- or -

$$H_0: \mu_A - \mu_B = 0$$

Our alternative hypothesis could be that women and men are different heights. That is,

$$H_1: \mu_A \neq \mu_B$$

- or -

$$H_1: \mu_A - \mu_B \neq 0$$

Is this a 1-tailed or 2-tailed hypothesis? It is not directional, so it is a two-tailed test. What might the hypotheses be for a 1-tailed test? Men are taller than women. $H_0: \mu_B \leq \mu_A$ & $H_1: \mu_B > \mu_A$.

Step 2: Criterion for decision

Figuring out your criterion is exactly the same process as before. You pick what your field has decided as being an accepted level of alpha (chance of making a type I error). For our example, let's assume $\alpha = 0.05$.

Steps 3 & 4: Sample and test statistics

Let's look at some sample data (in inches) for 9 women and 9 men:

Women's heights: 69, 63, 67, 64, 61, 66, 60, 63, 63

Men's heights: 67, 73, 74, 70, 70, 75, 73, 68, 69

We need to compute sample means and SS for each sample, because we have to consider the difference between these means and the variability for each sample as we conduct our test (just as we did in the one-sample case, but now we have two samples to consider). We do this just as we've done all along. If you need a review of how to calculate means and sum of squares, see chapters 8 and 9.

$$\bar{X}_A = 64.0, \bar{X}_B = 71.0 \qquad s_A = 8.25$$

$$SS_A = 66.0, SS_B = 64.0 \qquad s_B = 8$$

We'll look at the formula for this t-statistic, because that makes it easier to understand the formulas for variance and standard way. At the conceptual level, the formula is similar to previous t formulas. However, at the practical level, it is more complex because we have two samples, which means that we have two estimates. Here is full formula for the observed-t for independent samples with the formula for the one-sample t-test for comparison.

$$t_{obs} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{(\bar{X}_1 - \bar{X}_2)}} \qquad t = \frac{\bar{X} - \mu}{s_{\bar{X}}} =$$

We're interested in the difference between the two populations, so to compute the t statistic we need to see if the difference between our two samples is different from the difference between the two populations. So the numerator is pretty much straight forward: the difference between the two sample means minus the difference between the two populations. Remember that we are testing H_0 and in most cases the latter value is assumed to be 0, like it is in the related samples case)

The denominator is where things are more complex. What is $s_{\bar{x}_1 - \bar{x}_2}$ and how do we find it? We will go through it here to gain understanding of it, but normally we leave the calculations to SPSS or Excel.

Here are the relevant formulas for a one-sample t-test:

$$s = \sqrt{s^2} = \sqrt{\frac{SS}{n-1}} \qquad s_{\bar{x}} = \frac{s}{\sqrt{n}} = \sqrt{\frac{s^2}{n}}$$

The formula on the left is to get the standard deviation (square root of the variance) needed for the formula on the right. It is for the estimate of the standard error of the sample mean, which is the denominator in the formula for the one-sample t-statistic. Here we added an additional version of it with the variance for comparison purposes.

So, to answer the question above, $s_{\bar{x}_1 - \bar{x}_2}$ is the *estimate of the standard error from the two samples*. Recall that each sample will have some sampling error associated with it. What we need to do here is *pool* the error from the two samples. The reason that we want to pool the samples is to make the estimate of the standard error better. Basically, what we're doing is increasing the sample size that our estimate is based on, which will increase the precision of the estimate. So we *pool* the variances using the SS and *df* for each sample. Instead of solving for the standard deviation, like we did on the left above, here we solve for the analogous pooled variance.

$$\text{pooled variance} = s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2}$$

Note that if $n_1 = n_2$, this averaging formula can be used. $s_p^2 = \frac{s_1^2 + s_2^2}{2}$

Now we will use the pooled variance to solve for the estimated standard error from the two samples. Because each sample may be of different sizes (*n*'s), we need to *weight* the pooled variance by each sample's *n*.

$$s_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

If $n_A = n_B$, the standard error reduces to: $\sqrt{\frac{2s_p^2}{n}} = s_p \sqrt{\frac{2}{n}}$

So let's fill in the numbers from our example. Recall from above that $SS_A = 66.0$ and $SS_B = 64.0$; we also know that $n_A = 9$ and $n_B = 9$. So,

$$s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2} = \frac{66 + 64}{8 + 8} = 8.125 \qquad \text{or} \qquad s_p^2 = \frac{s_1^2 + s_2^2}{2} = \frac{8.25 + 8}{2} = 8.125$$

and

$$s_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = \sqrt{\frac{8.125}{9} + \frac{8.125}{9}} = 1.34 \quad \text{or} \quad \sqrt{\frac{2s_p^2}{n}} = s_p \sqrt{\frac{2}{n}} = \sqrt{8.125} * \sqrt{\frac{2}{9}} = 1.34$$

Now let's put together the whole t statistic (finishing *step 3*)

$$t_{obs} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{(\bar{x}_1 - \bar{x}_2)}} = \frac{(64 - 71) - 0}{1.34} = -5.22$$

Step 5: Compare observed to critical test value

So what is our critical t? First we need to know what our degrees of freedom are. Since we're using two samples, we need to take degrees of freedom for each sample into account. With one sample we used $n - 1$ because all of the values in the sample are free to vary but one, because we know the value of the sample mean. Now we've got two samples. How many values are free to vary?

Sample A: $n_A - 1$ and sample B: $n_B - 1$, so together there are $n_A + n_B - 2 = df$.

Looking back at our sample data above, $n_A = 9$ and $n_B = 9$, so df for our example is: $n_A + n_B - 2 = 9 + 9 - 2 = 16$

Now we can go to the t-table and look up the value for 2-tailed, $\alpha = 0.05$, $df = 16$.

We find $t_{crit} = \pm 2.12$, compared to $t_{obs} = 5.22$. So,

$$|t_{obs}| > |t_{crit}|$$

Steps 6 & 7: Decide about null hypothesis & Conclude about relationship

Our observed (computed) t statistic is greater than the critical t statistic, that is, it is in the critical region. Thus, we feel confident in rejecting the H_0 . There does seem to be a difference between the heights of men and women. The sample means tell us that men are taller and we know that this is a significant difference, so we also know that men are significantly taller. In calculating the effect size (note the pooled s), we find that it is very large. The sample mean difference is over three times the size of the pooled population standard deviation.

$$d = \frac{X_1 - X_2}{s_p} = \frac{64 - 71}{\sqrt{8.125}} = \frac{7}{2.85} = 3.46$$

Practice Questions: Set 12

(1) A between-subjects design was conducted to compare two groups. Data were the following:

$$\bar{X}_A = 58, \bar{X}_B = 52$$

$$n_A = 4, n_B = 4$$

$$SS_A = 84, SS_B = 108$$

- (a) Calculate the variance for each sample and then compute the pooled variance. You should find that the pooled variance is exactly halfway between the two sample variances. Why is this true for this particular study?
- (b) Do these data indicate a significant difference between the groups? Use a two-tailed test with $\alpha = .05$.

(2) For two samples, one sample has $n = 6$ and $SS = 500$, while the other sample has $n = 9$ and $SS = 670$. If the sample mean difference is 15 points, is this difference large enough to be significant for $\alpha = .05$ with a two-tailed test?

(3) Two people are arguing about the size of different breeds of dogs. One believes that German Shepherds are larger than Huskies, while the other person believes the opposite is true. So they conduct a study to see which one of them is correct. They sample the weights of 10 dogs of each breed. The data are as follows:

German Shepherds: 55, 72, 61, 43, 59, 70, 67, 49, 55, 63

Huskies: 48, 77, 46, 51, 60, 44, 53, 61, 52, 41

- (a) Should a 1-tailed or 2-tailed test be conducted? Why?
- (b) Conduct the appropriate test with $\alpha = .05$. Which breed is larger or are they the same?

ANSWERS ON P. 164

Study Guide for Exam 3

Terms

Alpha level	One-sample t-test
Alternative hypothesis	One-tailed test
Beta level	Pooled variance
Critical region	Related-samples t-test
Decision criteria	Standard error
Degrees of freedom	Statistical power
Difference scores	t-distribution
Effect size	Test statistic
Estimated standard error	Two-tailed test
Hypothesis testing	Type I error
Independent-samples t-test	Type II error
Null hypothesis	

Worksheet: Finding critical values in distribution tables:

z, 1-tailed, $\alpha = 0.05$; z =
z, 2-tailed, $\alpha = 0.05$; z =
z, 1-tailed, $\alpha = 0.01$; z =
z, 2-tailed, $\alpha = 0.01$; z =
t, 1-tailed, $\alpha = 0.05$, df = 29; t =
t, 2-tailed, $\alpha = 0.05$, df = 29; t =
t, 1-tailed, $\alpha = 0.01$, df = 29; t =
t, 2-tailed, $\alpha = 0.01$, df = 29; t =

Worksheet (including formulas): z-test

The z-test allows us to test a hypothesis about a sample against a population mean, when *the population standard deviation is also known*. It can be conducted in Excel with ZTEST, which returns only the *p*-value, but other statistics are available from other formulas.

Steps for conducting a z-test

- 1) Hypotheses (Is H_A directional or not? Is test 1- or 2-tailed?)
- 2) Criterion for decision (If not told otherwise, assume $\alpha = 0.05$.)
- 3) Sample statistics (In Excel, AVERAGE & STDEV but need formula for SE)
- 4) Test statistic (In Excel, need formula)
- 5) Compare observed to critical test value (In Excel, compare *p*-value to α -value)
- 6) Decide about null hypothesis
- 7) Conclude about relationship (If there is one, calculate effect size)

Example: A class of 25 students averages 23 for ACT scores while the population statistics are $\mu = 21$ and $\sigma = 3$. Is the class above average in ACT scores?

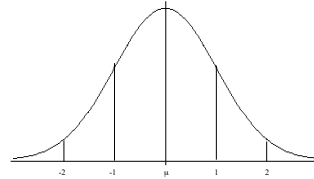
1) Hypotheses

H_0

H_A :

2) Criterion (shade in critical region)

- a. $\alpha =$
- b. 1- or 2-tailed H_A ?



3) Sample statistics (give the name of each)

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

4) Test statistic

$$z_{\bar{X}} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

5) Compare observed to critical test value

- a. Is t_{obs} in critical region?
- b. Which has the larger absolute value?

6) Decide about null hypothesis

- a. Reject or fail to reject H_0 ?

7) Conclude about relationship

- a. Support an effect or not?
- b. If so, calculate effect size

$$d = \frac{\bar{X} - \mu}{\sigma}$$

Worksheet (including formulas): One-Sample t-test

The one sample t-test allows us to extend hypothesis testing procedures to cases where we are testing a sample against a population mean, but we **DO NOT** know the population standard deviation. It should be run in SPSS.

Steps for conducting a one sample t-test

Same as for z-test, except need to calculate degrees of freedom.

Example: The quiz average for a class is 7.5. The teacher predicts that if classical music is playing in the background the quiz scores will increase. She tries this out with 5 students and they score 7, 8, 8, 9, and 9. Using alpha set at 0.05, see if the teacher's hypothesis is supported.

Student	Score		
A	7		
B	8		
C	8		
D	9		
E	9		

1) Hypotheses

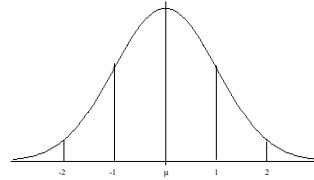
H_0

H_A :

2) Criterion (shade in critical region)

a. $\alpha =$

b. 1- or 2-tailed H_A ?



3) Sample statistics (give the name of each)

$$\bar{X} = \frac{\sum X}{N} =$$

$$SS = \sum (X - \bar{X})^2 =$$

$$s = \sqrt{\frac{SS}{n-1}} =$$

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} =$$

4) Test statistic

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}} =$$

5) Compare t -observed to t -critical

- a. Is t_{obs} in critical region?
- b. Which has the larger absolute value?

6) Decide about null hypothesis

- a. Reject or fail to reject H_0 ?

7) Conclude about relationship

- a. Support an effect or not?
- b. If so, calculate effect size

$$d = \frac{\bar{X} - \mu}{s}$$

Worksheet (including formulas): Paired-Samples t-test

The paired (or related) samples t-test is used when two groups of paired participants are tested or one group of participants is tested twice. It should be run in SPSS, but it can be run in Excel, which returns only the p -value.

Steps for conducting a paired samples t-test

Same as for one sample t-test, except hypotheses are about differences between means and you compute sample statistics for differences)

Example: Here is data illustrating comparing college student's motivation scores before and after Thanksgiving break to see if there is an effect of a week off school.

Student	Before	After			
A	65	70			
B	68	69			
C	50	55			
D	75	73			
E	80	82			

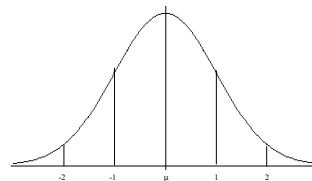
1) Hypotheses

H_0 :

H_A :

2) Criterion (shade in critical region)

- a. $\alpha =$



b. 1- or 2-tailed H_A ?

3) Sample statistics (give the name of each)

$$\bar{D} = \frac{\sum D}{n} =$$

$$SS_D = \sum (D - \bar{D})^2 =$$

$$s_D = \sqrt{\frac{SS_D}{n_D - 1}}$$

$$s_{\bar{D}} = \frac{s_D}{\sqrt{n_D}} =$$

4) Test statistic

Remember that the mean (μ_D) we are testing our hypothesis against is zero (0) because H_0 predicts no difference.

$$t_{\bar{D}} = \frac{\bar{D} - \mu_{\bar{D}}}{s_{\bar{D}}}$$

5) Compare t -observed to t -critical

- Is t_{obs} in critical region?
- Which has the larger absolute value?

6) Decide about null hypothesis

- Reject or fail to reject H_0 ?

7) Conclude about relationship

- Support an effect or not?
- If so, calculate effect size

$$d = \frac{\bar{D}}{s_D}$$

Worksheet (including formulas): Independent-Samples t-test

This design allows for a comparison between two different, unrelated groups of participants. Independent samples t-tests do not control for individual differences, and thus have more error than a paired samples t-test. It should be run in SPSS and a bar

graph created. Preliminary results can be run in Excel, but it returns only the p -values for tests with and without equal variance without an easy way to test which is the case.

Steps for conducting an independent-samples t-test

Same as for one sample t-test, except hypotheses are about differences between means and you compute pooled variance and the standard error of differences.

Example: Here is data illustrating the motivation scores of students in a required course compared to scores of a different group of students in an elective course at the same level. Let's test the hypothesis that students in elective courses are more motivated than those in required courses. Assume $\alpha = 0.05$.

X_1 (Req)			X_2 (Elec)		
5			2		
3			4		
4			1		
5			3		
3			2		

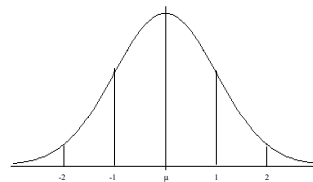
1) Hypotheses

H_0 :

H_A :

2) Criterion (shade in critical region)

- $\alpha =$
- 1- or 2-tailed H_A ?



3) Sample statistics (give the name of each)

$$\bar{X} = \frac{\sum X}{n}$$

X_1

X_2

$$SS = \sum (X - \bar{X})^2$$

$$df = n - 1$$

$$s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2}$$

$$s_p^2 = \frac{s_1^2 + s_2^2}{2}$$

$$s_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

$$s_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{2s_p^2}{n}} = s_p \sqrt{\frac{2}{n}}$$

4) Test statistic

Remember that the difference between means (μ_1 & μ_2) in H_0 is zero (0).

$$t_{obs} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{(\bar{x}_1 - \bar{x}_2)}}$$

5) Compare t -observed to t -critical

- Is t_{obs} in critical region?
- Which has the larger absolute value?

6) Decide about null hypothesis

- Reject or fail to reject H_0 ?

7) Conclude about relationship

- Support an effect or not?
- If so, calculate effect size using pooled s , which = $\sqrt{s_p^2}$

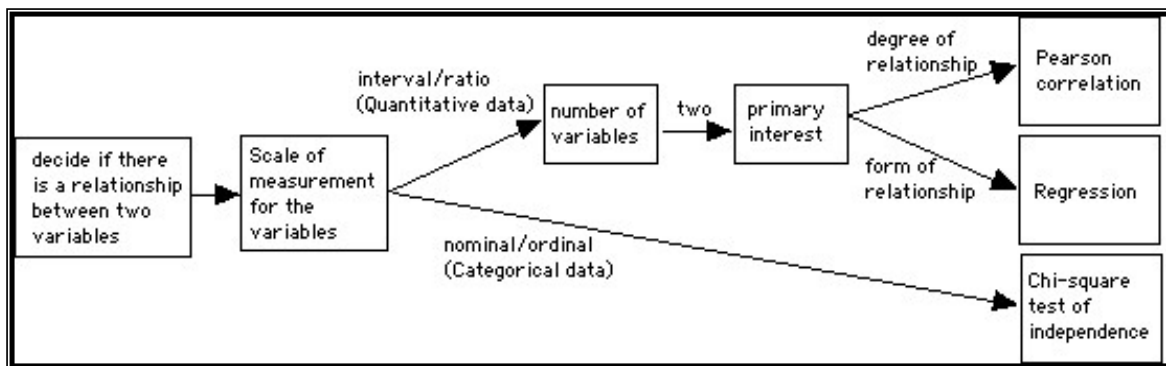
$$d = \frac{\bar{X}_1 - \bar{X}_2}{s_p}$$

ANSWERS ON P. 166-171

IV. Drawing Conclusions about Relationships between Variables and Population Parameters

In the previous section we examined research designs in which we typically examined the relationship between two variables, a continuous dependent variable (e.g., exam scores) and a discrete independent variable (e.g., study method: crammed or distributed). We used the IV to split the DV into two groups and compared the two groups to see if the IV had an effect on the DV (i.e., does study method impact exam scores? = did the crammed study group differ from the distributed study group).

In this section we'll discuss how to use the hypothesis testing framework to analyze data from different research designs. The hypotheses for these designs generally test whether or not two variables are systematically related to one another. These analyses are commonly found for observational (correlational) research designs.



In the last portion of the course, we'll also examine a different kind of inferential statistical framework: Estimation. Within this framework, rather than test hypotheses about populations, we try to estimate population parameters using data collected from samples.

Chapter 20: Hypothesis Testing with Correlation

Back in lab 12 we discussed Pearson's correlation coefficient (r) as a descriptive statistic used to describe the relationship between two continuous variables. We may also use this statistic within the hypothesis testing framework as an inferential statistic. The hypotheses that we test are whether or not there is a relationship and even about what direction the relationship has. At the population level, a relationship is represented by rho (ρ), and at the sample level by our familiar r .

So when is a correlation the appropriate analysis? Check the decision tree.

The logic is the same as before.

- Determine the hypotheses, and critical level of alpha.
- Find your df , and the critical value of r from a table.
- Compute your observed r .
- Compare the critical and observed r 's,
- Make your decision about the null hypothesis.

Example

Suppose that we wanted to know if students who near campus have higher GPAs than students who live farther away and commute to campus. We could measure students' GPAs and also measure how far away they live by measuring the distance to their residence from the middle of the quad. These are the two measured variables we're interested in.

Now let's go through our hypothesis testing steps:

Step 1: Hypotheses

Two-tailed:

$H_0: \rho = 0$; there is no relationship between X & Y .

$H_A: \rho \neq 0$; there is a relationship between X & Y .

We are making our predictions as a comparison with 0, because 0 would indicate no relationship. However, close inspection of the example ("*to know if students who near campus have higher GPAs than students who live farther away and commute to campus*") reveals that a set of *one-tailed* hypotheses would be more appropriate. Here we predict a negative correlation: as distance decreases, GPA increases).

One-tailed:

$H_0: \rho \geq 0$; the relationship between X & Y is positive or zero.

$H_A: \rho < 0$; the relationship between X & Y is negative.

Step 2: Criterion for decision

We'll use the conventional $\alpha = .05$.

Step 3: Sample statistics

Here are our sample data:

Subject	GPA	Distance from campus (in miles)
A	3.45	1.3
B	3.03	0.8
C	2.67	5.7
D	2.50	0.5
E	3.16	2.9
	Mean _{GPA} = 2.96	Mean _{distance} = 2.24

Recall the deviation formula for a Pearson r statistic. (We'll focus on this rather than the z formula because we need SP to calculate the slope for regression in the next chapter.)

$$r = \frac{SP}{\sqrt{SS_X SS_Y}}$$

The denominator of the formula requires us to calculate the sum of squares (SS) for each measure individually, and the numerator of the formula requires calculation of the sum of products of the two variables (SP). For a review of how to compute these components, see Chapter 12.

For our example, we get

$$SP = -0.63, SS_{GPA} = .58, SS_{distance} = 18.39$$

Plugging these SS and SP values into our r equation gives us

$$r = 0-.19$$

Step 4: Test statistic

Now we need to find our critical value of r using a table as we did for our z and t -tests. A table of critical r 's is included at the end of this PIP packet. We need to know our degrees of freedom, because like t , the r distribution changes depending on the sample size. For an r -value,

$$df = n - 2$$

What is n ? It is the number of individuals in our sample. Here it is 5; just as in

within-subjects related-samples t-tests, two scores are being taken for each participant, but n refers to the number of participants not scores. Why subtract 2? Because we know two values, X & Y , so we lose two degrees of freedom.

For our example, we have $df = 5 - 2 = 3$. Now, with $df = 3$, $\alpha = .05$, and a one-tailed test, we can find r_{critical} in the table of Pearson r values.

The $r_{\text{crit}} = -0.805$ (negative because we are doing a one-tailed test looking for a negative relationship). Notice in the table that for small degrees of freedom very large values of r are needed to be confident of significance. It is preferable to have at least 30 participants for calculating an r -value. With that sample size, an r -value of about .30 is significant at the .05-level.

Step 5: Compare observed to critical test value

$$|r_{\text{obs}}| < |r_{\text{crit}}|$$

The observed r of -0.19 is not in the critical region that begins at -0.805 ,

Step 6: Decide about null hypothesis

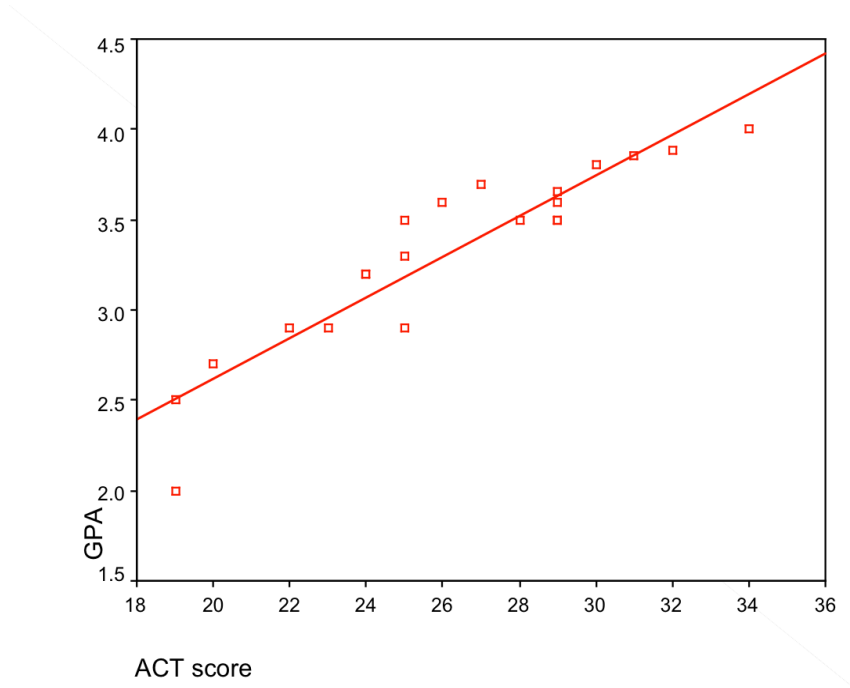
Since the size of the relationship observed in this sample is likely to occur by chance, we cannot reject the null hypothesis.

Step 7: Conclude about relationship

We conclude that we have no evidence to support a relationship between GPA and distance from campus.

Chapter 21: Regression

In the previous chapter, we worked on the correlation between two variables. In this lab we are going to extend our knowledge of correlations by looking at a **best-fit line** between two variables. Let's look at an example of hypothetical data on the relationship between ACT scores and college GPA. As you can see there is a pretty strong positive relationship between the two variables. Drawing a **best fit line** through the data can tell us even more about the data.



Some reasons to look at the best-fit line include:

- 1) In cases of weak relationships, the best-fit line can make the sign of the relationship easier to identify
- 2) The line is *like a mean of a set of scores*. Thus, it gives a representative view of the data, even when the data points are not shown.
- 3) The most important purpose of the line is to provide prediction. A college admissions office can tell what GPA an incoming freshman is likely to achieve based on their ACT scores. For example, a student with an ACT score of 24 is likely to achieve a GPA of 3.0.

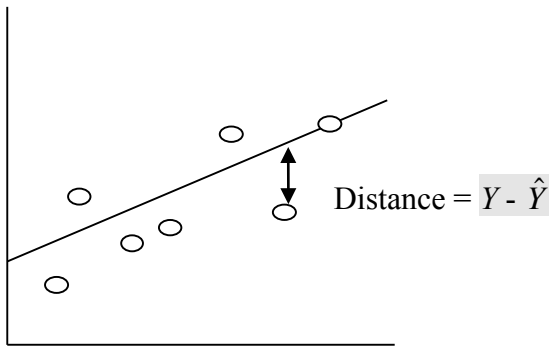
The statistical procedure for calculating the best fitting line is called **regression**. The best fitting line is then called the **regression line**. This line can be expressed in an equation as well as in a graph.

The line can be described by the following linear equation:

$\hat{Y} = bX + a$, where \hat{Y} = the predicted value of the Y variable, a = the intercept and b = the slope.

As you can see in the example above, the line does not cross through every data point. It is important to find a line that will differ the least from all the data points, in order to reduce error. Thus it is important to identify how well the line actually fits the data points. This is done by basically averaging out the distance from each data point to the line or finding the **least-squares solution**, comparable to how we found the variance and standard deviation for a single variable.

This is found by comparing the actual value of Y for each data point with the predicted value of \hat{Y} , termed “Y-hat.”



Since some distances will be positive and some will be negative, the second step is to square each of the distances and add them together to find the overall squared error. Note that this is analogous to the Sum of Squares of deviations, but here we are squaring predicted values rather than observed values.

$$\text{Total squared error} = SS_{error} = \sum(Y - \hat{Y})^2$$

Here are the other relevant formulas.

$$\text{Regression line} = \hat{Y}$$

$$\text{slope} = b = \frac{SP}{SS_x}$$

$$\text{intercept} = a = \bar{Y} - b\bar{X}$$

Let's put that all together with an example looking at the relationship between homework and quiz scores for students in a class. X represents quiz scores and Y represents homework scores for 5 students.

X	Y	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$
7	11	2	5	10	4
4	3	-1	-3	3	1
6	5	1	-1	-1	1
3	4	-2	-2	4	4
5	7	0	1	0	0
				16 = SP	10 = SS _X

For this example, the mean for X is 5 and the mean for Y is 6. These means were used to calculate deviation scores for each variable in the third and fourth columns. For example, for the first X data point the distance between 7 and the mean of 5 is 2. The fifth column contains the products of deviation scores to calculate the sum of these products (SP) and the final column has the total squared deviations for X (SS_X). We can now calculate our regression equation.

$$b = \frac{SP}{SS_X}$$

$$= 16/10 = 1.6$$

$$a = \bar{Y} - b\bar{X} = 6 - 1.6(5) = -2$$

Thus, the regression equation is:

$$\hat{Y} = 1.6X - 2$$

We can then use this equation to predict a homework score based on a quiz score. For example, if someone scored a 5.5 on the quiz, their predicted homework grade would be

$$\hat{Y} = 1.6(5.5) - 2 = 6.8$$

There are a few things to keep in mind when using a regression equation for prediction:

- 1) There is always error in prediction (unless there is a ± 1.0 correlation). The **standard error of the estimate** (a value calculated from the total squared error) describes the error between the regression line and the actual data points.
- 2) Regression should not be used to make predictions beyond the range of values of X included in the data set. Thus we could not use quiz scores below 3 and above 7 in our example from above to predict homework scores. The relationship might not be linear beyond these values.

Finally, keep in mind that the regression equation alone does not tell us how accurate we are in our predictions. Thus, we need to compute the **standard error of the estimate**, in order to have an idea of the accuracy in our predictions. Standard error of the estimate is similar to standard deviation, just like the regression line is similar to the mean when we use one variable. First we need to compute the **total squared error**. Remember the equation for this is:

$$SS_{error} = \sum(Y - \hat{Y})^2$$

Then we'll divide that by our degrees of freedom to get an average error value (like when we calculate the variance).

$$\frac{SS_{error}}{df}$$

In the case of regression our degrees of freedom = $n - 2$. It is $n - 2$, and not $n - 1$ because we need 2 points to make a straight line. There will be no error between these points and the line, so we have 2 restrictions.

Finally we take the square root of the whole thing:

$$\text{Standard error of the estimate} = s_{est} = \sqrt{\frac{SS_{error}}{df}}$$

We'll use our original example to find the standard error of the estimate.

X	Y	Predicted Y values $\hat{Y} = 1.6X - 2$	Error $(Y - \hat{Y})$	Squared Error $(Y - \hat{Y})^2$
7	11	9.2	1.8	3.24
4	3	4.4	-1.4	1.96
6	5	7.6	-2.6	6.76
3	4	2.8	1.2	1.44
5	7	6.0	1.0	1.00
			0	$SS_{error} = 14.40$

Note: the sum of errors from a line is 0, just as the sum of deviations from a mean is 0.

Since we have 5 subjects, $df = 5 - 2 = 3$.

The standard error of the estimate is then

$$\sqrt{\frac{SS_{error}}{df}} = \sqrt{\frac{14.4}{3}} = 2.19$$

The value of 2.19 tells us the average distance on the Y axis between the actual data points and the predicted regression line.

Practice Questions: Set 13

(1) A set of X and Y scores have a mean of X of 4, SS_x of 15, mean of Y of 5, and SP of 30.

- (a) What is the regression equation for predicting Y from X?
- (b) What are the predicted Y scores for the following X scores: 3, -2, 5, 6

(2) Find the regression equation for predicting Y from X for the following set of scores. (Show your work for each step)

<u>X</u>	<u>Y</u>
0	9
1	7
2	11

(3) Find the regression equation and standard error of estimate for the following set of data. (Show your work for each step)

<u>X</u>	<u>Y</u>
4	1
7	16
3	4
5	7
6	7

(4) When a correlation is close to ± 1.0 , then the standard error of the estimate will be _____. When the correlation is close to 0, then the standard error of estimate will be _____.

- (a) large, small
- (b) close to 1.0, close to 0
- (c) small, large
- (d) cannot tell from the information given

ANSWERS ON P. 172

Chapter 22: Chi-Square Test

The last test we'll discuss is used for different types of data than those we've been looking at. So far our tests have been used for continuous data (i.e., data from interval and ratio scales). The *chi-square test for independence* is used when you are testing hypotheses with categorical (i.e., nominal) data. This test is done on data that have been organized using *cross tabulation*. The test is named after the Greek letter *chi*, and we'll see later that it is a squared value; the lower-case symbol is χ^2 . The theoretical distribution is different from those for z, t, and r tests.

Suppose, for example, that we want to know whether people with different work statuses (e.g., full-time, retired) differ in happiness? This amounts to tabulating frequencies for work status and happiness. Cross tabulation or *crosstabs* gives frequencies for one variable separately for each level of another variable. In other words, cross tabulation is a statistical technique used to display a breakdown of the data by these two variables (that is, it is a table that displays the frequency of different majors broken down by gender). To create the crosstabs, we count up the number of people that fit each category.

Let's look at some data. Suppose that we surveyed people and asked them about their work status (full-time or retired) and if they were happy (yes or no). Listed below are the results of the survey:

<u>Subject #</u>	<u>Word Status</u>	<u>Happy</u>
1	full-time	yes
2	retired	yes
3	retired	no
4	retired	yes
5	full-time	no
6	full-time	yes
7	retired	no
8	full-time	yes
9	retired	no
10	retired	no
11	full-time	no
12	full-time	yes
13	retired	yes
14	full-time	yes
15	retired	no
16	retired	no
17	full-time	no
18	full-time	yes
19	full-time	yes
20	retired	no

From these data, we can create a table with work status in the columns and happiness in the rows by counting the number of subjects in each category. The observed frequencies go in the body of the table. Our crosstabs would be:

	Work status	
Happiness	Full-time	Retired
Yes	7	3
No	3	7

For example, this table tells us that 7 people in our survey were full-time workers and happy. So cross tabulation is a way to organize data.

We can also use crosstabs as a first step in conducting a hypothesis test on our two categorical variables to see if they are related to one another (i.e., to see if happiness depends on work status and vice versa). What we're learning about here is if a relationship exists, not if one variable causes a change in the other variable. The χ^2 test can help us test for a relationship in this case. It works by comparing observed frequencies with the frequencies that are expected if there is no relationship between the variables (i.e., if they are independent). A significant χ^2 means there is a relationship because the observed and expected frequencies are sufficiently different.

So let's work through the hypothesis testing steps for χ^2 and our example data above.

Step 1: Hypotheses

Our null hypothesis for χ^2 is that there is no relationship, which is the same as saying the variables are independent. The hypotheses can be stated either way.

- H₀: Happiness is independent of work status OR No relationship between happiness and work status
- H_a: Happiness is NOT independent of work status OR There is a relationship between happiness and work status

Step 2: Criterion for decision

We'll set $\alpha = 0.05$

Step 3: Sample statistics

Crosstabs:

	Work status	
Happiness	Full-time	Retired
Yes	7	3
No	3	7

These are our observed data. We now have to use these data to estimate the frequency of the cells expected by chance (if there is no relationship).

Part 1: Obtain row and column totals, also called the *marginals* (in bold below).

Work status/Happiness	Full-time	Retired	Marginals
Yes	7	3	10
No	3	7	10
Marginals	10	10	

Part 2: Compute the **expected frequencies** (f_e) to add to the table. Calculate these values by multiplying the appropriate marginals and dividing by the total number of subjects. These are how many subjects that would be expected from the marginals if there is no relationship between the variables. Another version of same formula takes the proportion (or percentage) in the row for a cell.

$$f_e = \frac{f_{row} f_{column}}{n} \quad \text{or} \quad f_e = \frac{f_{row}}{n} * f_{column}$$

- Full-time/Yes = $f_e = (10*10)/20 = 5$
- Full-time/No = $f_e = (10*10)/20 = 5$
- Retired/Yes = $f_e = (10*10)/20 = 5$
- Retired/No = $f_e = (10*10)/20 = 5$

NOTE: These numbers will not always be the same, as they are in this example because the marginals are equal!

Work status/Happiness		Full-time	Retired	
Yes	Observed	7	3	10
	Expected	5	5	
No	Observed	3	7	10
	Expected	5	5	
		10	10	

Step 4: Test statistic

Now we're ready to calculate the χ^2 test statistic from the observed (f_o) and expected (f_e) frequencies. We're going to treat each cell as a test case and add them up. The formula is the familiar observed differences over expected difference by chance.

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

So for our example, we have:

$$\begin{aligned}\chi^2 &= \sum \frac{(7-5)^2}{5} + \frac{(3-5)^2}{5} + \frac{(3-5)^2}{5} + \frac{(7-5)^2}{5} \\ &= 4/5 + 4/5 + 4/5 + 4/5 = 3.2\end{aligned}$$

Degrees of freedom

$$df = (\#Columns - 1) * (\#Rows - 1)$$

$$df = (2-1)(2-1) = 1$$

Step 5: Compare observed to critical test value.

Go to the χ^2 table and find the critical value.

For this example, with $df = 1$, and $\alpha = 0.05$, the critical χ^2 value from the table is 3.84. As with other test statistics, as the difference between observed and expected values gets larger, the observed test value gets larger and the probability of such a value gets smaller. In this case, our χ^2 value too small to be unlikely by chance.

$$|\chi^2_{\text{observed}}| < |\chi^2_{\text{critical}}|$$

Step 6: Decide about null hypothesis

Since the observed χ^2 is not in the critical region, we cannot reject H_0 .

Step 7: Conclude about relationship

We have no evidence here of a relationship between happiness and work status.

Practice Questions: Set 14

(1) New research seems to suggest that kids raised in homes with pets tend to have fewer allergies than kids raised without pets. A survey study was conducted to test this finding. A sample of 100 adults were asked if they had allergies (yes/no) and how many pets they had between the ages of 1 and 10 years old (0/1/2 or more). Use the crosstabs table below to conduct a chi-square test with $\alpha = .05$. Indicate whether these data support the previous findings or not.

# Pets/Allergies	0	1	2 or more
No	10	25	35
Yes	15	10	5

(2) For the following voting survey data, create a crosstabs table and then conduct a test to determine if the two variables are related. Use $\alpha = .01$.

Gender Plans to Vote For

Male Bush
 Male Bush
 Female Kerry
 Female Bush
 Female Kerry
 Male Kerry
 Male Bush
 Female Kerry
 Female Bush
 Male Kerry
 Male Bush
 Female Bush
 Male Kerry
 Female Kerry
 Male Bush
 Female Kerry
 Male Bush
 Male Kerry
 Female Bush
 Female Bush
 Male Kerry
 Female Kerry
 Male Bush
 Female Kerry
 Female Bush
 Male Bush
 Female Kerry
 Male Bush
 Female Kerry
 Male Bush

ANSWERS ON P. 173

Chapters 23: Estimation of Population Means

Everything that we did in the last four chapters is related to this chapter. However, the logic of what we are doing here, estimation, is *different* from the logic used in hypothesis testing.

In the last several chapters we tested the a null hypothesis that basically asked the question, is this different from that? Estimation asks a different question. With estimation we are making educated guesses as to the value of a population parameter.

When do we use estimates?

- 1) After we do hypothesis testing and have rejected the H_0 .
“So we reject that there is no difference due to the treatment, but we still want to know how much of a difference is there”
- 2) You just want to know some basic information about a population, but you can't measure the whole group, so instead you take a sample.

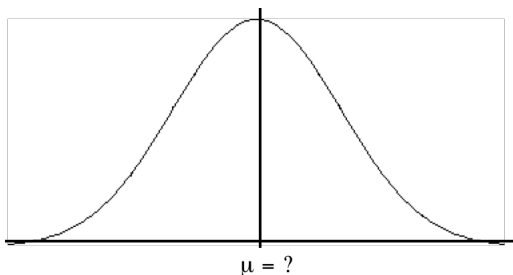
In fact, we've already done a lot of estimating. Each time we computed a t-statistic, we used an estimate of the variability of the population. Whenever we found a significant difference, we estimated the effect size. In this section we'll extend our use of estimation to include estimates of the population mean.

Two kinds of estimates of the population mean.

- 1) *point estimates* of the mean: using a single number as your estimate of an unknown quantity
- 2) *interval estimates* (confidence intervals) of the mean: using a range of values as your estimate of an unknown quantity. When an interval is accompanied with a specific level of confidence (or probability) , it is called a confidence interval.

Both kinds of estimates are determined by the same equation, the difference is that for the point estimates, we'll just compute a single number (that's why it is called a point estimate), but for the interval estimate, we'll compute an interval between two points.

Let's start at the conceptual level. Consider the following population distribution.



Suppose that we guess that the mean is 85? How confident are we in this guess?

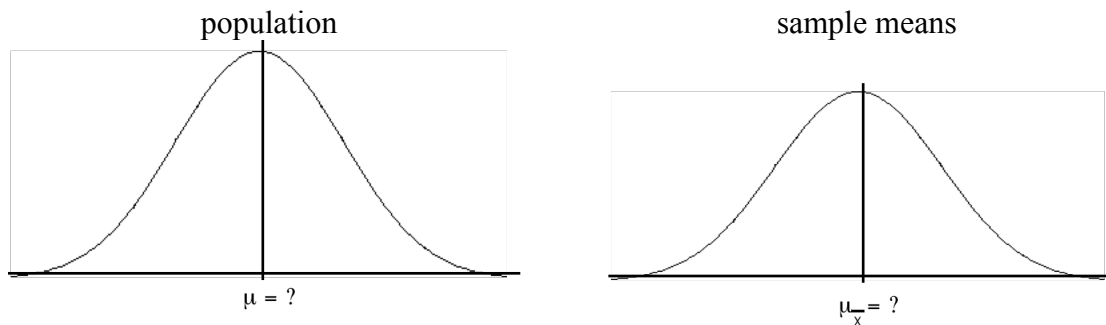
Suppose that we guess that the mean is somewhere between 71 & 99? How confident are we in this guess?

Hopefully, you will think that you'd be more confident in the range. This difference corresponds to the difference between point and interval estimations.

	<u>point estimate</u>	<u>interval estimate</u>
Disadvantages	it doesn't convey any sense of how much precision we have in making that estimate.	we often need to have one specific value, a range of possible values just may not be enough

Okay, now let's begin with a point estimate of the mean. What will be the best single estimate of the population mean?

The mean of the distribution of sample means



However, suppose that all we have is a single sample. Now what is our best guess?

The sample mean. So how good is it?

- 1) It is the only piece of evidence that we have, so it is our best guess.
- 2) Recall, that most of our sample means will be pretty close to the population mean, so we have a good chance that our sample mean is close.

How can we get an estimate where we'd have a better chance of being right? Instead of giving a point estimate, we can estimate an interval. Again, consider the distribution of sample means. If we think in terms of z-scores, and pick a range of ± 1 z-units. Then what we can say is that about 68% of the possible means are within that range. So we can be pretty confident that our population mean fits into that range.

Now let's formalize things a bit. Let's first talk about the logic of estimation, and then move onto the actual formulas that we'll use.

Step 1: You begin by making a reasonable estimation of what the z (or t) value should be for your estimate.

For a point estimation, you want what? z (or t) = 0, right in the middle. For an interval, your values will depend on how confident you want to be in your estimate

Step 2: You take your “reasonable” estimate for your test statistic, and put it into a formula and solve for the unknown population parameter. Because you use a reasonable estimate for your test statistic, then you should get a reasonable estimate of the population parameter.

The formula

It is the same one(s) that we’ve been using all along, but we do a little algebra to move it around. Since, instead of solving for an observed z or t score, we are solving for the population parameter, we change to the mean of the DSM; that is how we will estimate the population mean. Since we need to enter a known z or t scores; we use a critical score corresponding to the confidence interval we have selected. We look up a critical test value, which can be plus or minus. This is so we get a high and low value for our interval.

For an example, let’s assume that $\bar{X} = 85$, $\sigma = 5$, $n = 25$

$$z_{\bar{X}} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \quad \rightarrow \quad z_{\bar{X}}(\sigma_{\bar{X}}) = \bar{X} - \mu \quad \rightarrow \quad \mu_{\bar{X}} = \bar{X} \pm z_{crit}(\sigma_{\bar{X}})$$

Step 1: To estimate μ , we make a reasonable estimate of z . Our best guess will be when $z = 0$. So, we plug that into the formula.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{25}} = 1$$

$$\mu_{\bar{X}} = \bar{X} \pm 0(1) = 85$$

Step 2: We see that $\mu_{\bar{X}} = \bar{X}$ is our most reasonable estimate.

Okay, that’s the formula for point estimation. What about for an **interval estimation**? We use the same formula, but we note the two changes above. So, the first thing that we want to do is decide how confident do we want to be in our estimate. Let’s chose 95%, which is analogous to using .05 for an alpha level. So we need to go to the unit normal table and figure out between what two z -scores do 95% of the sample means lie. Since 5% won’t be between, we want two tails with 2.5% in each, so the z -scores are ± 1.96 .

$$\mu_{\bar{X}} = \bar{X} \pm z_{crit}(\sigma_{\bar{X}}) = 85 + 1.96\left(\frac{5}{\sqrt{25}}\right) = 86.96$$

$$\mu_{\bar{X}} = \bar{X} \pm z_{crit}(\sigma_{\bar{X}}) = 85 - 1.96\left(\frac{5}{\sqrt{25}}\right) = 83.04$$

Chapter 24: Estimation Combined with Hypothesis Testing

Our initial example of population estimation used z-scores. The same logic applies to using t-statistics.

How do you know which test statistic to use? We follow the same logic as before: It depends on the design.

How many groups and scores are sampled?

If one group with one score, then is the population standard deviation known?

If so, then use the z-test.

If not, then use the one-sample t-test.

If one group and two scores per person, then use related-samples t-test.

If two groups, then are the samples independent?

If independent, then use independent-samples t-test

If dependent, then use related-samples t-test

If more than two groups, then an F-test is needed, which goes beyond this course.

Formulas for estimation of a population mean

One sample (σ known)	$\mu_{\bar{X}} = \bar{X} \pm z_{crit}(\sigma_{\bar{X}})$
------------------------------	--

One sample (σ unknown)	$\mu_{\bar{X}} = \bar{X} \pm t_{crit}(s_{\bar{X}})$
--------------------------------	---

Related samples	$\mu_{\bar{D}} = \bar{D} \pm t_{crit}(s_{\bar{D}})$
-----------------	---

Independent samples	$\mu_1 - \mu_2 = \bar{X}_1 - \bar{X}_2 \pm t_{crit}(s_{X_1 - X_2})$
---------------------	---

Note: when using t-tests, make sure that you use the appropriate *dfs*.

Examples

One-sample t-test. Consider the data given above, $\bar{X} = 85$ and $n = 25$, with the sample $s = 5$. There are 24 *dfs*, so the critical t-value for $p = .025$, 1-tailed (.05 total error) is ± 2.064 .

$$\mu_{\bar{X}} = \bar{X} \pm t_{crit}(s_{\bar{X}}) = 85 \pm (2.064)\left(\frac{5}{\sqrt{25}}\right) = 85 \pm 2.064$$

$$= 82.94 \text{ to } 87.06.$$

Related-sample t-test. Dr. S. Beach reported on the effectiveness of cognitive-behavioral therapy as a treatment for anorexia. He examined 12 patients, weighing each of them before and after the treatment. Estimate with 95% confidence the average

population weight gain for those undergoing the treatment. Differences (posttreatment - pretreatment weights) are 10, 6, 3, 23, 18, 17, 0, 4, 21, 10, -2, 10. They average 10 and $s_D = 8.24$.

$$s_{\bar{D}} = \frac{s_D}{\sqrt{n}} = \frac{8.24}{\sqrt{12}} = 2.38$$

For $df = 11$, $p = .025$, 1-tailed (.05 total error), $t_{crit} = \pm 2.201$

$$\mu_{\bar{D}} = \bar{D} \pm t_{crit}(s_{\bar{D}}) = 10 \pm (2.201)(2.38) = 10 \pm 5.24$$

= 4.76 to 15.24

Independent-samples t-test. Dr. Mnemonic develops a new treatment for patients with a memory disorder. He randomly assigns 8 patients to one of two samples. One sample (A) receives the new treatment while the other (B) receives the old treatment. He then tests both groups with a memory test to see if there is a difference. Estimate with 95% confidence the population difference between the two groups. Data provided is $\bar{X}_A = 44.5, s_A = 7.19, \bar{X}_B = 50, s_B = 9.13$.

$$s_p^2 = \frac{(s_A^2 df_A) + (s_B^2 df_B)}{df_A + df_B} = 67.52$$

$$s_{X_1 - X_2} = \sqrt{\frac{2 * s_p^2}{n}} = \sqrt{\frac{2 * 67.52}{4}} = \sqrt{33.76} = 5.81$$

For $df = 6$, $p = .025$, 1-tailed (.05 total error), $t_{crit} = \pm 2.45$.

$$\mu_1 - \mu_2 = \bar{X}_1 - \bar{X}_2 \pm t_{crit}(s_{X_1 - X_2}) = 5.5 \pm (2.45)(5.81)$$

= -8.73 to 19.73

Combining estimation and hypothesis testing

Since the same formula is being used, it may have occurred to you that we could conduct population estimation and hypothesis testing in combination. That is the case.

In the final example above, the CI (95%) included 0. What does that signify? In hypothesis testing for independent samples, 0 is the mean of the population of differences for the null hypothesis. That is, if there is no difference between the two samples, we expect a mean difference for the distribution of sample means to be 0. If 0 is within the confidence interval, it is a likely value, so we cannot reject the null hypothesis.

If we conduct the hypothesis test for the same example, we reach the same conclusion.

$$t_{obs} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{X}_1 - \bar{X}_2}} = \frac{5.5 - 0}{5.81} = 0.95$$

For $df = 6$, $p = .05$, 2-tailed, $t_{crit} = \pm 2.45$ (note that this is the same t_{crit} for the 95% CI).

Since $t_{obs} < t_{crit}$, we fail to reject the null hypothesis. There is more than a 5% chance that the difference could result from sampling error in the population of differences ($\mu = 0$) posited by the null hypothesis.

We can state as a general rule that if the confidence interval for estimating the population mean includes the mean posited by the null hypothesis (as long as both are using the same critical test value), we cannot reject the null hypothesis.

You have probably guessed that if the confidence interval does not include the mean posited by the null hypothesis, then we can reject it. That is correct (as long as both are using the same critical test value).

The related-samples example is such a case, although its 1-tailed hypothesis test creates a complication. The 95% CI was from 4.76 to 15.24 for weight gain resulting from the new treatment for anorexia. That the mean of 0 predicted by H_0 is not included suggests that we can reject the null hypothesis.

$$t_{obs} = \frac{\bar{D} - \mu_{\bar{D}}}{(s_{\bar{D}})} = \frac{10}{2.38} = 4.2$$

We compare this observed t to a critical t for $df = 11$, $p = .05$, 1-tailed: $t_{crit} = \pm 1.796$. (The critical value for estimation was 2.201. This complication occurs when running a 1-tailed hypothesis test and estimating a 95% CI. If it were a 90% CI, the critical test values would be the same.)

Since $t_{obs} > t_{crit}$, we reject the null hypothesis. There is less than a 5% chance that the difference could result from sampling error in the population of differences posited by the null hypothesis ($\mu \leq 0$).

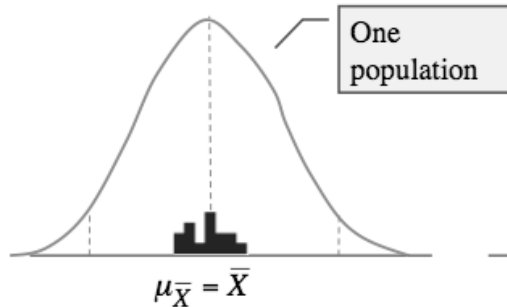
Inferences from one or two populations

When we estimate the mean of a population, we are working with only one population, the one we are making inferences about (from the sample). In contrast, when we test hypotheses, we are working with two populations, the one posited by the null hypothesis and the alternate one posited by our research hypothesis.

The figure below illustrates the case of failing to reject the null hypothesis. Each population is shown with its mean and 2.5% cutoffs in the tails. On the left is the population inferred from the sample, centered on its mean. On the right are the two populations in hypothesis testing. Remember that we test the null hypothesis. In this

case, the sample mean falls well within the cutoffs of the H_0 population, so we cannot reject the null hypothesis. Note that the confidence interval of the estimated mean of the sample's population includes the mean of the H_0 population. In this case, we are deciding that there is only one population, the H_0 population, and its mean is known, so we do not need to estimate it from the sample.

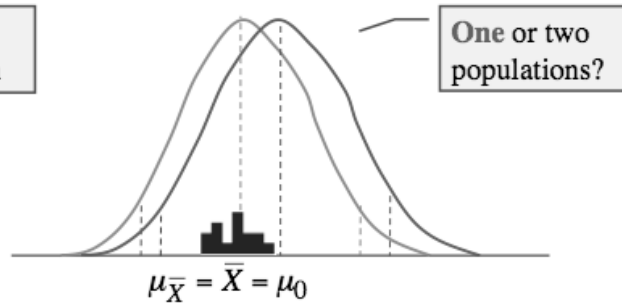
Estimate population mean from sample mean



One population—that sample is centered on

Test null hypothesis: $\mu_{\bar{X}} = \mu_0$

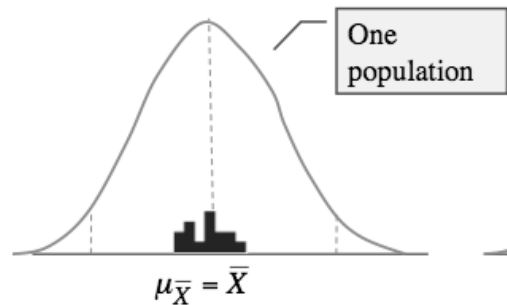
H_0 : fail to reject (no treatment effect)



Sample's population includes mean of H_0 population, so it must be one and the same population

The figure below illustrates the opposite case: rejecting the null hypothesis. The population on the left is the same as above, but on the right the two populations for hypothesis testing are more separated than above. In this case, the sample mean falls outside the cutoffs of the H_0 population, that is, in the critical region. Thus, we can reject the null hypothesis.

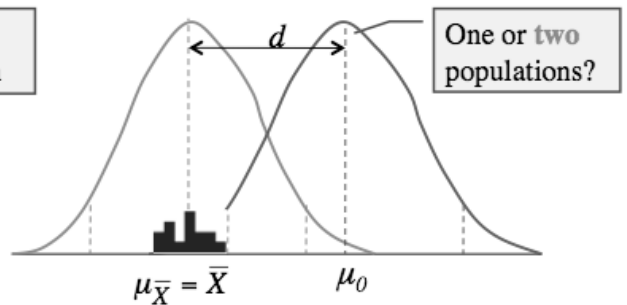
Estimate population mean from sample mean



One population—that sample is centered on

Consider alternative hypothesis: $\mu_{\bar{X}} \neq \mu_0$

H_0 : reject (a treatment effect)



Sample's population does not include mean of H_0 population (at cutoff point shown), so it must be a second population

Note that the confidence interval of the estimated mean of the sample's population does not include the mean of the H_0 population. In this case, we are deciding that there are two populations, the H_0 population and the one posited by the alternate hypothesis, and that the sample is drawn from this second population. We need the confidence interval to estimate the population mean, and we can calculate the effect size from the distance between the means of the two populations.

With this figure, which is on the cover of this text, we have reached the end of our coverage of descriptive and inferential statistics in this course. Subsequent courses extend the concepts we have covered to comparison of three or more means with the analysis of variance and F tests, multivariate correlation and regression for three or more variables, and multivariate prediction models. You now have entered the modern world of reasoning and decision-making using statistics. Your statistical knowledge and skills should be of regular use to you as an educated person in everyday living and in your career.

Practice Questions: Set 15

- (1) A sample of $n = 25$ scores is obtained from an unknown population. The sample has a mean $\bar{X} = 200$ and a standard deviation $s = 20$.
- (a) Use the sample data to compute a 90% confidence interval (CI) estimate of the population mean
 - (b) Use the sample data to compute an 80% CI estimate of the population mean.
 - (c) Describe the relationship between level of confidence of the estimate and the width of the confidence interval.
- (2) Two groups of pollsters ask voters how many political commercials they had seen in the last month.
- (a) The first group asked a sample of 64 voters and had a sample mean $\bar{X} = 200$ and a standard deviation $s = 20$. Compute a 95% CI estimate of the population mean.
 - (b) The second group asked a sample of 36 voters and had a sample mean $\bar{X} = 200$ and a standard deviation $s = 20$. Compute a 95% CI estimate of the population mean.
 - (c) Compare the confidence intervals for the two groups. Which gives a “better” estimate of the population mean. Explain what you mean by “better.”
- (3) Dr. Brainiac conducted an experiment examining how people use examples to solve problems. He asked two groups of participants to solve the same problem. The groups differed with respect to what kind of sample problem they were given prior to solving the test problem. The first group was given a sample problem that, on the surface seemed very different, but the underlying solution was analogically similar to that of the test problem. The second group received a sample problem that on the surface seemed related to the test problem, but had a very different solution. Dr. Brainiac measured how many seconds were required to solve the test question.

Analogically similar solution group	Different solution group
$n = 15$ $\bar{X} = 66$ sec. $SS = 2030$	$n = 10$ $\bar{X} = 78$ sec. $SS = 1420$

- (a) Compute a point estimate of the mean difference between the two groups.
- (b) Compute a 95% CI of the mean difference between the two groups.
- (c) Can Dr. Brainiac reach a conclusion from the results of his experiment?

ANSWERS ON P. 174

Study Guide for Exam 4 and Final Exam

Listed below are many of the concepts discussed in the Drawing Conclusions Part II component of the course. Remember that while Exam 4 covers essentially what is new since the last exam (you still need to from before that the new material builds upon, e.g., means and standard deviations), the Final exam is considered cumulative, and covers all of the course material (so study all four of the review pages).

Terms

- Best fit line
- Chi-square test of independence
- Confidence intervals
- Cross tabulation
- Estimation
- Least squares linear regression
- Margin of error
- Pearson's correlation coefficient
- Point estimates
- Rho (ρ)
- Scatterplot
- Standard error of the estimate
- Total squared error

Formulas (provide the name of each and what it is used for)

$$r = \frac{SP}{\sqrt{SS_X SS_Y}}$$

$$\hat{Y} = (X)(\text{slope}) + (\text{intercept}) = Xb + a = bX + a$$

$$b = \frac{SP}{SS_X}$$

$$a = \bar{Y} - b\bar{X}$$

$$SS_{error} = \sum(Y - \hat{Y})^2$$

$$s_{est} = \sqrt{\frac{SS_{error}}{df}}$$

$$f_e = \frac{f_{column} f_{row}}{n} \quad \text{or} \quad f_e = \frac{f_{row}}{n} * f_{column}$$

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$\mu = \bar{X} \pm z_{crit}(\sigma_{\bar{X}})$$

$$\mu = \bar{X} \pm t_{crit}(s_{\bar{X}})$$

$$\mu_D = \bar{D} \pm t_{crit}(s_{\bar{D}})$$

$$\mu_1 - \mu_2 = \bar{X}_1 - \bar{X}_2 \pm t_{crit}(s_{(X_1 - X_2)})$$

Finding critical values in distribution tables

r, 1-tailed, $\alpha = 0.05$, $df = 28$; r =

r, 2-tailed, $\alpha = 0.05$, $df = 28$; r =

r, 1-tailed, $\alpha = 0.01$, $df = 28$; r =

r, 2-tailed, $\alpha = 0.01$, $df = 28$; r =

χ^2 , $\alpha = 0.05$, $df = 2$; $\chi^2 =$

χ^2 , $\alpha = 0.01$, $df = 2$; $\chi^2 =$

Sample problems

Complete the five steps for hypothesis testing for each problem. When the null hypothesis is rejected, calculate the effect size and estimate the population mean. **Assume $\alpha = 0.05$ for all hypotheses and the 95% confidence interval for estimation.** BE COMPLETE.

1. The 30 students taking statistics one semester have SAT quantitative scores averaging $\bar{X} = 480$. Given the known population parameters ($\mu = 500$, $\sigma = 100$), is the class below average on this test? What if the same mean had been from two classes totaling 100? Extra credit: Explain any difference in outcome.

2. The same 30 students average $\bar{X} = 98$ ($s = 5$) on a mathematics test with a known population mean $\mu = 100$. Is the class below average on this test? Extra credit: Explain any difference in outcome for the significance of the class's 2-point difference on this test compared to its 20-point difference on the SAT.

3. The same 30 students kept a log, which revealed that they studied statistics 4 hours per week before the midterm and averaged an increase of $\bar{D} = 2$ hours per week after the midterm. The standard deviation of the 30 pairs of difference scores (after-before) is $s_D = 1$. Did their studying increase significantly?

Extra credit: How do the critical t-values compare for testing the hypothesis and estimating the population parameter? When are they the same and when are they different?

4. The two lab sections ($n = 15$ in each) of the above class were compared on a common exam. The data for section 1 are $\bar{X} = 80$, $SS = 170$, and for section 2 are $\bar{X} = 78$, $SS = 200$. Is there a statistically significant difference?

Extra credit: Explain any difference in outcome for the significance of the sections' 2-point difference on this exam and the class's 2-point difference from the norm on a mathematics test (problem 2).

5. The 30 students in the statistics class kept a log of study hours (noted above) to determine if they are positively related to exam scores (noted above). Data for study hours are $\bar{X} = 5$ and $SS_X = 30$. Data for exam scores are $\bar{Y} = 79$ and $SS_Y = 185$. $SP = 25$ and $SS_{\text{error}} = 270$. Find the correlation and the amount of variance accounted for, and test for significance. Find the regression line and standard error of the estimate. What are the predicted exam scores for the following amounts of weekly study: 0, 3, 5, and 10?

Extra credit: Why does the regression predict only a few more points on the exam when study time is doubled?

6. Test the relationship between grades in the statistics course and graduating as a psychology major. The following workspace provides data from one semester's students followed until graduation.

Psych Graduate	Grades									
	A		B		C		D		F/W	
	Obs	Exp	Obs	Exp	Obs	Exp	Obs	Exp	Obs	Exp
Yes	20		30		20		5		0	
No	5		25		40		25		10	

ANSWERS ON P. 175

V. Statistical Tables

The Unit Normal Table (z)

	Second decimal place of z									
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2297	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233

2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010

The t Distribution

<i>df</i>	Proportion in One Tail					
	0.25	0.10	0.05	0.025	0.01	0.005
	Proportion in Two Tails					
	0.50	0.20	0.10	0.05	0.02	0.01
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707
7	0.711	1.415	1.895	2.365	2.998	3.499
8	0.706	1.397	1.860	2.306	2.896	3.355
9	0.703	1.383	1.833	2.262	2.821	3.250
10	0.700	1.372	1.812	2.228	2.764	3.169
11	0.697	1.363	1.796	2.201	2.718	3.106
12	0.695	1.356	1.782	2.179	2.681	3.055
13	0.694	1.350	1.771	2.160	2.650	3.012
14	0.692	1.345	1.761	2.145	2.624	2.977
15	0.691	1.341	1.753	2.131	2.602	2.947
16	0.690	1.337	1.746	2.120	2.583	2.921
17	0.689	1.333	1.740	2.110	2.567	2.898
18	0.688	1.330	1.734	2.101	2.552	2.878
19	0.688	1.328	1.729	2.093	2.539	2.861
20	0.687	1.325	1.725	2.086	2.528	2.845
21	0.686	1.323	1.721	2.080	2.518	2.831
22	0.686	1.321	1.717	2.074	2.508	2.819
23	0.685	1.319	1.714	2.069	2.500	2.807
24	0.685	1.318	1.711	2.064	2.492	2.797
25	0.684	1.316	1.708	2.060	2.485	2.787
26	0.684	1.315	1.706	2.056	2.479	2.779
27	0.684	1.314	1.703	2.052	2.473	2.771
28	0.683	1.313	1.701	2.048	2.467	2.763
29	0.683	1.311	1.699	2.045	2.462	2.756
30	0.683	1.310	1.697	2.042	2.457	2.750
40	0.681	1.303	1.684	2.021	2.423	2.704
60	0.679	1.296	1.671	2.000	2.390	2.660
120	0.677	1.289	1.658	1.980	2.358	2.617
∞	0.674	1.282	1.645	1.960	2.326	2.576

Critical Pearson r Values

	One-tailed test			
	0.05	0.02	0.01	0.005
	Two-tailed test			
df	0.10	0.05	0.02	0.01
1	.988	.997	.9995	.9999
2	.900	.950	.980	.990
3	.805	.878	.934	.959
4	.729	.811	.882	.917
5	.669	.754	.833	.874
6	.622	.707	.789	.834
7	.582	.666	.750	.798
8	.549	.632	.716	.765
9	.521	.602	.685	.735
10	.497	.576	.658	.708
11	.476	.553	.634	.684
12	.458	.532	.612	.661
13	.441	.514	.592	.641
14	.426	.497	.574	.623
15	.412	.482	.558	.606
16	.400	.468	.542	.590
17	.389	.456	.528	.575
18	.378	.444	.516	.561
19	.369	.433	.503	.549
20	.360	.423	.492	.537
21	.352	.413	.482	.526
22	.344	.404	.472	.515
23	.337	.396	.462	.505
24	.330	.388	.453	.496
25	.323	.381	.445	.487
26	.317	.374	.437	.479
27	.311	.367	.430	.471
28	.306	.361	.423	.463
29	.301	.355	.416	.456
30	.296	.349	.409	.449
35	.275	.325	.381	.418
40	.257	.304	.358	.393
45	.243	.288	.338	.372
50	.231	.273	.322	.354
60	.211	.250	.295	.325
70	.195	.232	.274	.302
80	.183	.217	.256	.283
90	.173	.205	.242	.267
100	.164	.195	.230	.254

Critical Values of the Chi Square Distribution

<i>df</i>	Level of Significance		
	.05	.025	.01
1	3.84	5.02	6.64
2	5.99	7.38	9.21
3	7.81	9.35	11.34
4	9.49	11.14	13.28
5	11.07	12.83	15.09
6	12.59	14.45	16.81
7	14.07	16.01	18.48
8	15.51	17.53	20.09
9	16.92	19.02	21.67
10	18.31	20.48	23.21
11	19.68	21.92	24.72
12	21.03	23.34	26.22
13	22.36	24.74	27.69
14	23.68	26.11	29.14
15	25.00	27.49	30.58
16	26.30	28.85	32.00
17	27.59	30.19	33.41
18	28.87	31.53	34.80
19	30.14	32.85	36.19
20	31.41	34.17	37.57
21	32.67	35.48	38.93
22	33.92	36.78	40.29
23	35.17	38.08	41.64
24	36.42	39.36	42.98
25	37.65	40.65	44.31
26	38.88	41.92	45.64
27	40.11	43.19	46.96
28	41.34	44.46	48.28
29	42.56	45.72	49.59
30	43.77	46.98	50.89
40	55.76	59.34	63.69
50	67.50	71.42	76.15
60	79.08	83.29	88.38
70	90.53	95.02	100.42
80	101.88	106.63	100.43
90	113.15	118.14	124.12
100	124.34	129.56	135.81

VI. Solutions to Practice Problems

Practice Question Solutions: Set 1

- (1) (a) The four players: Dunwoody, Osuna, Pettitte, & Sosa
- (b) The variables are:
Name (non numeric values)
Team (non numeric values)
Position (non numeric values)
Age (numeric value)
Salary (numeric value)
- (c) Sammy Sosa's salary is listed not in dollars, but rather in \$1,000s (that is, he was making not \$9,000 a year, but rather \$9,000,000 a year)
- (2) (a) In the first situation, rather than randomly assigning people to an exercising condition, the researcher is "observing" people who are already exercising or not exercising. So the values for that variable are pre-existing. In the second example, the researcher randomly assigns people to the two exercise conditions. So, their pre-existing exercise habits don't really matter, there is an equal chance that they'll be either group.
- (b) The additional information gained comes from the fact that the experimenter makes the decision (with random assignment) about the people's exercise routines. In this way he can be reasonably certain that any differences between groups are due to the assigned exercise conditions rather than to some other pre-existing difference between the individuals in the two groups.
- (continues on next page)

(3) The key here is the kind of methods used by the researchers.

In the first example, the researcher is performing an experiment (i.e., assigning rats to one of the two conditions). So the researcher can be pretty confident that any differences between the rats with respect to body size are likely to be due to the independent variable (whether they got the hormone or not). In addition, the researcher is interested in the effect of a growth hormone. To get this, he measures the weight (size) of the rats. In this case the dependent measure (weight) is pretty clearly a good indicator of the effectiveness of growth hormone (there is a huge body of literature that have demonstrated that growth hormones are causally related to body size).

In the second example, the researchers are using an observational study. Here they are not assigning mothers to a *work* or *don't work* condition. So it is possible that any differences found between the two groups of daughters may be due to the mother's employment status, or to other differences between the groups (e.g., suppose that the mothers who chose to work differ with respect to femininity and that they pass these traits on to their daughters. So it is these traits, not the employment status of the moms that count).

(4)

- a. This suggests that there is an order to the scores, but that the different scores may be of different sizes. This means that the scale is probably *ordinal* (if it were interval or ratio, we'd know how much larger it was).
- b. Here we know that Peter's is larger and we know that a ratio of the scores is interpretable. So the scale of measurement is *ratio*.
- c. Here we know what scores they got, but can't make a lot of comparisons (other than the type of score). This suggests that the scale of measurement used is *nominal*.

Practice Question Solutions: Set 2

(1)

- (a) Drug type
- (b) Drug type: 2 levels – get the drug, get the placebo
- (c) Drinking behavior after 6 months
- (d) Random variable (although if they decide to analyze the age variable, then you could treat it as an explanatory variable and the entire design as a quasi-experiment)
- (e) Confound variable – because you won't be able to know whether any change in drinking behavior is due to the drug condition or the age condition

(2)

The hope is that any random variables will end up being distributed roughly equally in all of the different experimental conditions. That way they do not become confound variables.

Practice Question Solutions: Set 3

(1)

- (a) Answers will vary. One example of a “bad” sample using a voluntary response method might be to put an question in the student paper asking people to log into a website and rate how they feel about the parking situation. Here it is likely that those with strong opinions (and in this situation, probably those who have had bad experiences) are likely to constitute an overly large percentage of your sample.
- (b) Answers will vary. One example of a “bad” sample would be to stand by the parking control office and ask students who go in an out of there what their opinion of the parking situation would be. This would be bad because there is a good chance that the people going in an out of that office may be a biased sub set of all students in general (e.g., those paying parking tickets).
- (c) Answers will vary. One approach would be to get access to all of the students enrolled at the college, randomly select 700 of those students and call each one and ask their opinion about the parking. Because the respondents are randomly selected from the entire population, you are greatly reducing the potential of getting a biased sample.

(2) (a) $1 / 35 = 0.029$

(b) $25 / 35 = 0.71$

(c) $4 / 35 = 0.114$

Practice Question Solutions: Set 4

(1)

X	f	p	%	cp	cf
5	3	0.125	12.5	1.0	24
4	5	0.208	20.8	0.874	21
3	8	0.333	33.3	0.666	16
2	5	0.208	20.8	0.333	8
1	3	0.125	12.5	0.125	3

N=24

(2)

66.6%

(3)

The distribution is symmetrical and unimodal

Practice Question Solutions: Set 5

(1)

$$\text{Mean} = (1+3+5+0+1+3)/6 = 2.167$$

Median = 0,1,1,3,3,5 even # of scores, so average of middle two $(1+3)/2 = 2.0$

Mode: most frequent score; here there are two, 1 & 3

(2)

$$\bar{X} = (10+10+10+10+10+16)/6 = 11.0$$

(3)

If 5 points were added to every score, then the mean would increase by 5. So the original distribution must have had a mean $\mu = 30 - 5 = 25$.

(4)

For a perfectly symmetrical distribution, the mean is equal to the median. So the median would be = 30.

(5) For the following set of scores, identify which measure would provide the best description of central tendency and explain your answer.

There is an extreme score (the 0) in this distribution leading to negative skew. So the best measure of center would probably be the median.

Problem Solutions: Study Guide for Exam 1

a. Display results in a frequency distribution table and a histogram.

Describe each abbreviation in the table:

f = frequency of each value

p = probability of value = f/n

% = percentage of value = $100(f/n)$

cf = cumulative frequency: add *f*s from bottom

c% = cumulative percentage: add %s from bottom

Frequency Distribution Table

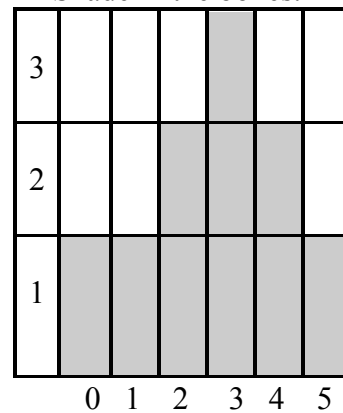
Add column headings.

X	f	p	%	cf	c%
5	1	.1	10	10	100
4	2	.2	20	9	90
3	3	.3	30	7	70
2	2	.2	20	4	40
1	1	.1	10	2	20
0	1	.1	10	1	10

Σ 10 1.0 100

Histogram

Shade in the boxes.



b. Provide the following information about the sample:

$$p(X = 3) = .3$$

$$p(X > 0) = .9$$

$$p(0 < X < 4) = .6$$

$$p(X > 4) = .1$$

c. Indicate measure of central tendency for the sample:

$$\text{Mean} = 27/10 = 2.7$$

$$\text{Median} = 3$$

$$\text{Mode} = 3$$

Practice Question Solutions: Set 6

(1) The range is $\text{max} - \text{min} = 20 - 8 = 12$.

(2) Sample standard deviation is the square root of the variance, so here it is $\sqrt{4} = 2$.

(3) Subtracting a constant from every score in the distribution keeps every score in the same place (relative to each other), so the standard deviation will not change.

(4)

$$\text{Mean} = (1+1+1+3)/4 = 1.5$$

$$SS = \sum(X - \bar{X})^2 = (1-1.5)^2 + (1-1.5)^2 + (1-1.5)^2 + (3-1.5)^2 = 3$$

(5)

$$\text{Mean} = (9+1+8+6)/4 = 6.0$$

$$SS = \sum(X - \bar{X})^2 = (9-6)^2 + (1-6)^2 + (8-6)^2 + (6-6)^2 = 38$$

$$s^2 = SS/N = 38/4 = 9.5$$

$$s = \sqrt{s^2} = 3.08$$

Practice Question Solutions: Set 7

(1)

To directly compare them, transform both into z-scores and then compare the two z-scores.

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

$$\text{ACT: } z = (24 - 18)/6 = 1.0$$

$$\text{SAT: } z = (660 - 500)/100 = 1.6$$

(2)

$$\text{ACT: } z = (20 - 21)/3 = -0.33,$$

go to the Unit Normal Table and look up the proportion associated with this score.

$$z(.33) \rightarrow p = 0.3707.$$

(3)

Go into the table first and find 15% (or 0.1500). The z-score that corresponds to this is 1.04 (.1492 is closer to .1500 than is .1515). Now transform this z-score back into a raw SAT score.

$$(z)(\sigma) + \mu = X$$

$$1.04(100) + 500 = 604$$

(4)

Figure out the z-score for each score and then determine the area between the two scores.

$$\text{SAT: } z = (500 - 500)/100 = 0.0 \quad p \text{ of } z(0) = .5000$$

$$\text{SAT: } z = (650 - 500)/100 = 1.5 \quad p \text{ of } z(1.5) = .0668$$

Here you can subtract out the smaller tail from the larger one and get the desired area (You should make a sketch of the distribution and shade in the various regions.) $.5 - .0668 = 0.4332$

(5)

What is your percentile rank if you have an ACT of 25.5?

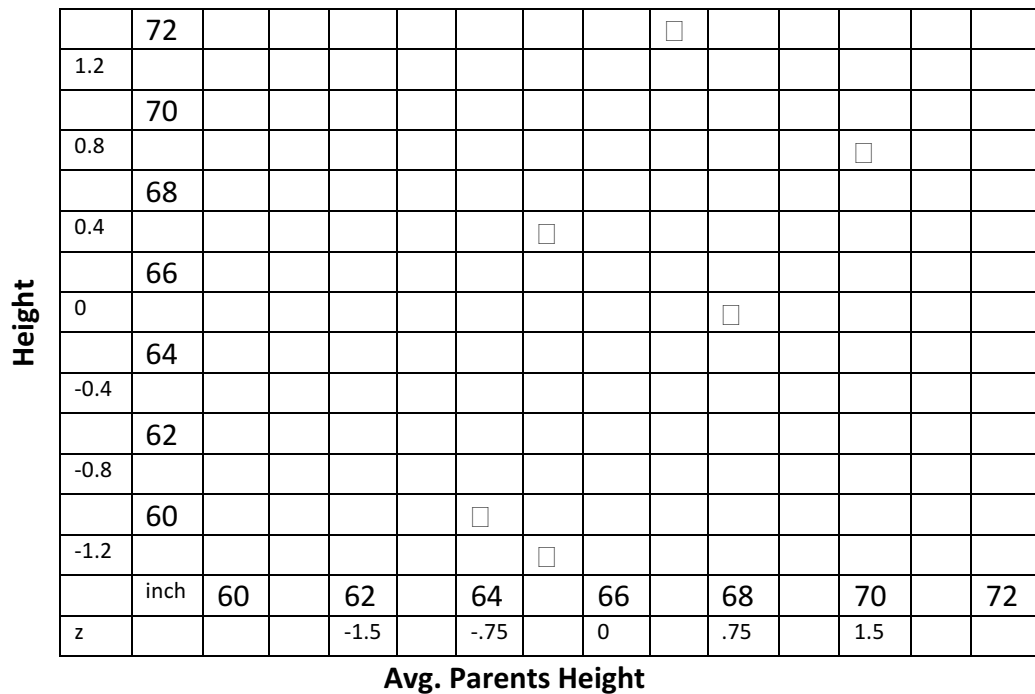
Rephrase as "what proportion of the scores are 25.5 or worse?"

$$\text{ACT: } z = (25.5 - 21)/3 = 1.5 \quad p \text{ of } z(1.5) = 0.0668$$

This value is the top tail; we want the complement of this, so we subtract it from 1.0, which is 0.9332 or 93.22 percentile.

Practice Question Solutions: Set 8

(1)



(2) A line through the points would be positively sloping, and the points would be fairly close to the line suggesting a moderate to strong correlation.

(3)

X (APH)	Y (Ht)	$(X - \bar{X})$	$(X - \bar{X})^2$	z_x	$(Y - \bar{Y})$	$(Y - \bar{Y})^2$	z_y	$(X - \bar{X})(Y - \bar{Y})$	$z_x z_y$
68	65	1.5	2.25	0.66	-0.33	0.11	-0.06	-0.50	-0.04
64	60	-2.5	6.25	-1.10	-5.33	28.44	-1.05	13.33	1.15
70	69	3.5	12.25	1.55	3.67	13.44	0.72	12.83	1.12
65	59	-1.5	2.25	-0.66	-6.33	40.11	-1.24	9.50	0.82
67	72	0.5	0.25	0.22	6.67	44.44	1.31	3.33	0.29
65	67	-1.5	2.25	-0.66	1.67	2.78	0.33	-2.50	-0.22
$\bar{X} =$	$\bar{Y} =$		$SS_X =$			$SS_Y =$		$SP =$	$\Sigma =$
66.5	65.3		25.5			129.33		36	3.12
$s_x =$	$s_y =$								
2.26	5.09								

(4)
continued on next page

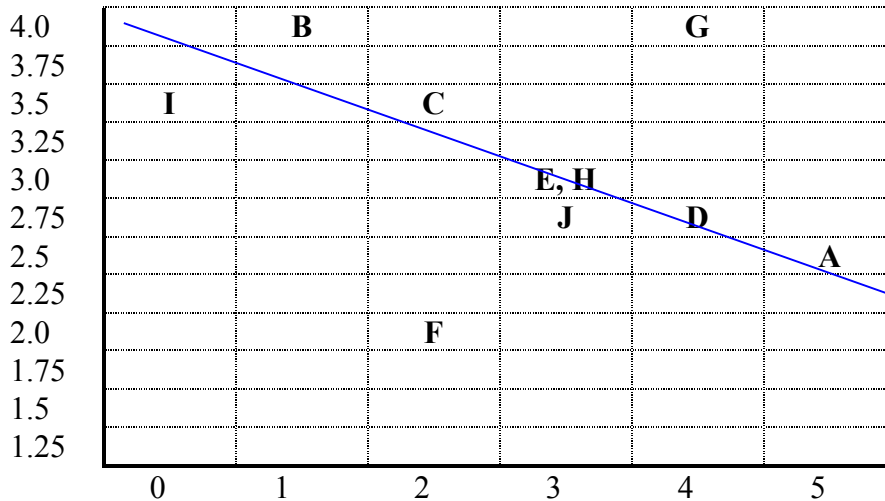
$$r = \frac{SP}{\sqrt{SS_X SS_Y}} = \frac{36}{\sqrt{25.5 * 129.33}} = \frac{36}{57.4} = 0.63$$

$$r = \frac{\sum z_x z_y}{n-1} = \frac{3.12}{5} = 0.63$$

(5) $r = 0.63$ is a moderately strong positive relationship between height and average parents height (suggesting that both variables move in the same direction). The scatterplot, especially the z-scores, show similar deviations, but less variability in the parents' average height.

Problem Solutions: Study Guide for Exam 2

a. Make a scatterplot of both of your variables by entering the letter of each person on the proper place on the graph. Draw the best fitting line through the points.



b. Describe each abbreviation in the table:

\bar{X} = **Mean, average of scores**

Where will you use the $\sum(X)$? **In formula for Mean**

$(X - \bar{X})$ = **Deviation from mean**

What must the $\sum (Y - \bar{Y})^2$ total? **0**

$(X - \bar{X})^2$ = **Deviation from mean squared**

$\sum (X - \bar{X})^2$ **Sum of Squares (SS), in formula for variance and SD**

$(X - \bar{X})(Y - \bar{Y})$ **Sum of Products, in deviation formula for r**

$\sum z_x z_y$ **Sum of z Products, in z formula for r**

c. Find the following statistics for Xs and Ys. Show formulas and calculations.

	X	Y
Mode	3	3.9
Median	3	3.0
Range	0-5	2.1-3.9
M	$\frac{\sum X}{N} = \frac{27}{10} = 2.7$	$\frac{\sum Y}{N} = \frac{30.8}{10} = 3.08$
SS	$\Sigma (X - \bar{X})^2 = 20.1$	$\Sigma (Y - \bar{Y})^2 = 4.04$
Variance = SS/n-1	$20.1/9 = 2.44$	$4.04/9 = 0.45$
SD = $\sqrt{SS/n-1}$	$\sqrt{2.44} = 1.49$	$\sqrt{.45} = 0.67$

d. Find the following statistics for Xs and Ys taken together. Show formulas and calculations.

Pearson's r

$$r = \frac{SP}{\sqrt{SS_X SS_Y}} = \frac{-3.42}{\sqrt{20.1 * 4.04}} = \frac{-3.42}{9.01} = -0.38$$

$$r = \frac{\sum z_x z_y}{n-1} = \frac{-3.42}{9} = -0.38$$

Practice Question Solutions: Set 9

(1)

A Type I error is concluding that there is an effect (a “difference”) when there really isn’t one.

A Type II error is concluding that there is not an effect, when there really is one.

(2)

Step 1: SAT Hypotheses

$H_0: \mu \leq 500$ (one-tailed because superintendent believes the program should increase SAT scores)

$H_A: \mu > 500$

Step 2: Criterion for decision: $\alpha = 0.05$, the standard level.

Step 3: Sample statistics (given in the question)

$\mu = 500$, $\sigma = 100$, $n = 25$, sample mean = 559

Step 4: Test statistic

Here we have one sample and know the population standard deviation, so we conduct a one sample z-test.

$$z = (X - \mu) / (\sigma / \sqrt{n}) = (559 - 500) / (100 / \sqrt{25}) = 2.95$$

Step 5: Compare observed to critical test value.

With $\alpha = 0.05$ and 1-tailed test, the critical z-score is 1.65. Here the computed z is 2.95, which is in the critical region (beyond the 1.65 critical score).

Step 6: Decide about null hypothesis: reject the H_0

Step 7: Conclude about relationship: The new program does significantly increase SAT scores. The effect size is $59/100$ or .59, which is considered moderate.

(3)

This would have an impact on the standard error (making it larger). So instead of dividing by $\sqrt{25}$, you would divide by $\sqrt{9}$. The end result would be that your computed z-score would be = 1.77. This score is still within the critical region and Dr. Standard would still reject the null hypothesis and come to the same conclusion as in 2 above. Note that the effect size does not change because it is independent of sample size.

Practice Question Solutions: Set 10

(1)

Step 1: Hypotheses

$H_0: \mu = 5.8$ (two-tailed because looking to see if sleep deprived group was different)

$H_A: \mu \neq 5.8$

Step 2: Criterion for decision: $\alpha = 0.05$

Step 3: Sample statistics (given in the question)

$\mu = 5.8, n = 101, \bar{X} = 4.5, s = 1.6$

Step 4: Test statistic

Here we have one sample and know the population standard deviation, so we conduct a one-sample z-test.

$t = (\bar{X} - \mu) / (s / \sqrt{n}) = (4.5 - 5.8) / (1.6 / \sqrt{101}) = -8.17$

$df = n - 1 = 101 - 1 = 100$

Step 5: Compare observed to critical test value

With $\alpha = 0.05$ and 2-tailed test the critical t-score is ± 2.0 (the table doesn't have $df = 100$, so we'll use the highest level in the table that isn't greater than our actual df , here $df = 60$). Here the computed t is -8.17 , which is in the critical region.

Step 6: Decide about null hypothesis: reject H_0

Step 7: Conclude about relationship: The sleep-deprived group is significantly different than the known population. The effect size is $1.3/1.6 = .81$, which is large.

(2)

Step 1: Hypotheses

$H_0: \mu = 14$ (two-tailed because looking to see if there is any change)

$H_A: \mu \neq 14$

Step 2: Criterion for decision: $\alpha = 0.05$

Step 3: Sample statistics (given in the question)

$\mu = 14, n = 5$

$\bar{X} = \Sigma X / n = 12$

$s = \sqrt{SS / n - 1} = 1.58$

Step 4: Test statistic

Here we have one sample and don't know the population standard deviation, so we conduct a one-sample t-test.

$t = (\bar{X} - \mu) / (s / \sqrt{n}) = (12 - 14) / (1.58 / \sqrt{5}) = -2.83$

$df = n - 1 = 5 - 1 = 4$

Step 5: Compare observed to critical test value

With an $\alpha = 0.05$ and 2-tailed test the critical t-score is ± 2.77 . Here the computed t is -2.83 , which is in the critical region.

Step 6: Decide about null hypothesis: reject H_0

Step 7: Conclude about relationship: The average age at which drinking first began is significantly younger. The effect size is $2/1.58 = 1.27$, which is very large.

(3)

In both cases, the critical t is the same ($df = 16-1 = 15$, two-tailed) = 2.131. The difference in (a) and (b) comes from the different estimated standard error (due to different sample standard deviations).

$$(a) s = \sqrt{60/15} = 2, s_x = 2/\sqrt{16} = .5, t = 3/.5 = 6$$

$t_{obs} > t_{crit}$ so reject H_0 ; effect size = $3/2 = 1.67$, which is very large

$$(b) s = \sqrt{600/15} = 6.32, s_x = 6.32/\sqrt{16} = 1.58, t = 3/1.58 = 1.9$$

$t_{obs} < t_{crit}$ so fail to reject H_0

(c) As the sample variability increases, so does the average difference expected by chance. As a result, in (a) the difference is 1.67 SD & we reject the H_0 , but in (b) the difference is less than $\frac{1}{2}$ SD ($3/6.32$) & we fail to reject the H_0 .

(4)

Step 1: Hypotheses

$H_0: \mu = 40$ (1-tailed because looking to see if they work more than 40 hrs/wk)

$H_A: \mu \neq 40$

Step 2: Criterion for decision: $\alpha = 0.05$

Step 3: Sample statistics (given in the question)

$$\mu = 40, n = 30$$

$$\text{solve } \bar{X} = 47.8$$

$$\text{solve } s = 5.93$$

Step 4: Test statistic

Here we have one sample and don't know the population standard deviation, so we conduct a one sample t-test.

$$s = \sqrt{1020/29} = 5.9; s_x = 5.9/\sqrt{30} = 1.08$$

$$t = (\bar{X} - \mu)/(s_x) = (47.8 - 40)/(1.08) = 7.24$$

$$df = n - 1 = 30 - 1 = 29$$

Step 5: Compare observed to critical test value

With $\alpha = 0.05$ and 1-tailed test, the critical t-score is 1.699. Here the computed t is 7.20, which is in the critical region.

Step 6: Decide about null hypothesis: reject H_0

Step 7: Conclude about relationship: The workers are working significantly greater than 40 hrs per week. The effect size is $7.8/5.9 = 1.3$, which is very large.

Practice Question Solutions: Set 11

(1)

Step 1: State the Hypotheses

$H_0: \mu_D = 0$ (two-tailed because looking for a “difference”)

$H_A: \mu_D \neq 0$

Step 2: Criterion for decision: $\alpha = 0.01$

Step 3: Sample statistics (given in the question)

$\mu_D = 0, n_D = 9$

solve for $\bar{D} = 4.0$

solve for $s_D = \sqrt{7/(9-1)} = 0.94$

Step 4: Test statistic

Here we have one sample and don't know the population standard deviation, so we conduct a one sample t-test.

$t = (\bar{D} - \mu_D) / (s_D / \sqrt{n}) = (4 - 0) / (0.94 / \sqrt{9}) = 12.83$

$df = n - 1 = 9 - 1 = 8$

Step 5: Compare observed to critical test value

With $\alpha = 0.01$ and 1-tailed test, the critical t-score is ± 3.355 . Here the computed t is 12.83, which is in the critical region.

Step 6: Decide about null hypothesis: reject H_0

Step 7: Conclude about relationship: groups do differ. The effect size is $4/.94 = 4.2$, which is very large.

(2)

A	B	D = (B-A)	$D - \bar{D}$	$(D - \bar{D})^2$
10	11	1	-7	49
-8	3	11	3	9
-11	11	22	14	196
15	10	-5	-13	169
0	8	8	0	0
-4	7	11	3	9
		$\bar{D} = 8$		$SS_D = 432$

Step 1: Hypotheses

$H_0: \mu_D = 0$ (2-tailed because looking for a “difference”)

$H_A: \mu_D \neq 0$

Step 2: Criterion for decision: $\alpha = 0.05$

(continues on next page)

Step 3: Sample statistics (given in the question)

$$\mu_D = 0, n_D = 6$$

$$\bar{D} = 8.0$$

$$s_D = \sqrt{432/(6-1)} = 9.3$$

Step 4: Test statistic

Here we have two samples, but they are paired, so we conduct a related-samples t-test.

$$t = (\bar{D} - \mu_D) / (s_D / \sqrt{n}) = (8.0) / (9.3 / \sqrt{6}) = 2.18$$

$$df = n - 1 = 6 - 1 = 5$$

Step 5: Compare observed to critical test value

With $\alpha = 0.05$ and 2-tailed test, the critical t-score is ± 2.571 . Here the computed t is 2.18, which is not in the critical region.

Step 6: Decide about null hypothesis: fail to reject H_0

Step 7: Conclude about relationship: Twins do not differ in handedness. Note that difference is almost 1 SD ($8/9.3$), which is a large effect size. This suggests that the experiment is worth redoing with a larger sample to provide more power.

(3)

$$t_{\text{set1}} = 4 / (10 / \sqrt{9}) = 1.2, \text{ effect size} = 4 / 10 = .4, \text{ medium}$$

$$t_{\text{set2}} = 4 / (2 / \sqrt{9}) = 6, \text{ effect size} = 4 / 2 = 2, \text{ very large}$$

Set 2 has the smaller standard deviation (and thus standard error) and thus a larger t, which is more likely to allow rejecting the H_0 .

Practice Question Solutions: Set 12

(1a) $s^2 = SS/n-1$
A: $s^2 = 84/3 = 28$
B: $s^2 = 108/3 = 36$
Pooled: $s^2 = (84+108)/(3+3) = 192/6 = 32$
1/2-way between the sample variances because sample ns are equal

(1b) $H_0: \bar{X}_A - \bar{X}_B = 0$
 $H_A: \bar{X}_A - \bar{X}_B \neq 0$
 t_{critical} (2-tailed, $\alpha = .05$, $df = 6$) = 2.447
 $s_{x-x} = \sqrt{s^2_{\text{pool}}/4 + s^2_{\text{pool}}/4} = \sqrt{32/4 + 32/4} = \sqrt{16} = 4$
 $t_{\text{observed}} = (58-52)/4 = 6/4 = 1.5$
observed < critical, so fail to reject null hypothesis and conclude no difference between groups

(2) $s^2_{\text{pool}} = (SS_A + SS_B)/(df_A + df_B) = (500+670)/(5+8) = 1170/13 = 90$
 $s_{x-x} = \sqrt{90/6 + 90/9} = \sqrt{15 + 10} = \sqrt{25} = 5$
 $t_{\text{observed}} = 15/5 = 3$
 t_{critical} (2-tailed, $\alpha = .05$, $df = 13$) = 2.16
observed > critical, so reject null hypothesis and conclude a difference between groups. Using s_p , the effect size is $15/\sqrt{90} = 15/9.5 = 1.58$, which is very large.

(continues on next page)

(3) Workspace

	X_A	X_B	$X_A - \bar{X}_A$	$(X_A - \bar{X}_A)^2$	$X_B - \bar{X}_B$	$(X_B - \bar{X}_B)^2$
	55	48	-4.4	19.36	-5.3	28.09
	72	77	12.6	158.76	23.7	561.69
	61	46	1.6	2.56	-7.3	53.29
	43	51	-16.4	268.96	-2.3	5.29
	59	60	-0.4	.16	6.7	44.89
	70	44	10.6	112.36	-9.3	86.49
	67	53	7.6	57.76	-0.3	0.09
	49	61	-10.4	108.16	7.7	59.29
	55	52	-4.4	19.36	-1.3	1.69
	63	41	4.6	21.16	-12.3	151.29
Sum	594	533	0	768.6 = SS_A	0	992.1 = SS_B
Mean	59.4	53.3				

$$s^2_{\text{pool}} = (SS_A + SS_B) / (df_A + df_B) = (768.6 + 992.1) / (9 + 9) = 1760.7 / 18 = 97$$

$$s_{x-y} = \sqrt{97.82/10 + 97.82/10} = \sqrt{19.56} = 4.42$$

$$t_{\text{observed}} = (59.4 - 53.3) / 4.42 = 6.1 / 4.42 = 1.38$$

t_{critical} (2-tailed, $\alpha = .05$, $df = 18$) = 2.101 (2-tailed because not only one direction hypothesized)

observed < critical, so fail to reject null hypothesis and conclude no difference between groups

Problem Solutions: Study Guide for Exam 3

Worksheet: Finding critical values in distributions

z, 1-tailed, $\alpha = 0.05$; z = **1.64**

z, 1-tailed, $\alpha = 0.01$; z = **2.32**

t, 1-tailed, $\alpha = 0.05$, df = 29; t = **1.699**

t, 1-tailed, $\alpha = 0.01$, df = 29; t = **2.462**

z, 2-tailed, $\alpha = 0.05$; z = **1.96**

z, 2-tailed, $\alpha = 0.01$; z = **2.59**

t, 2-tailed, $\alpha = 0.05$, df = 29; t = **2.045**

t, 2-tailed, $\alpha = 0.01$, df = 29; t = **2.756**

Worksheet: z-test

1 & 2) Hypotheses and criterion for decision

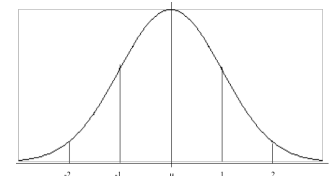
Ho: $\bar{X} \leq 21$

Ha: $\bar{X} > 21$

Critical region:

a. $\alpha = 0.05$

b. **1-tailed (upper)**



z-critical = 1.65

3) Sample statistics

Standard error

$$\begin{aligned}\sigma_{\bar{X}} &= \frac{\sigma}{\sqrt{n}} \\ &= 3/\sqrt{25} = 3/5 = 0.6\end{aligned}$$

4) Test statistic

z-observed

$$\begin{aligned}z_{\bar{X}} &= \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \\ &= (23 - 21) / 0.6 = 3.33\end{aligned}$$

5) Compare observed to critical test value: $z_{\text{crit}} = 1.65$

z-observed > z-critical, that is, it is in the critical region

6) Decide about null hypothesis: **reject null hypothesis**

7) Conclude about relationship: **The class has above average ACT scores. Effect size = $2/3 = .67$, which is large.**

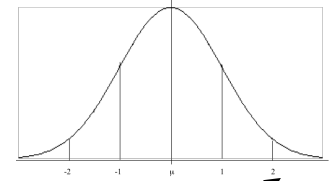
Worksheet: One-Sample t-test

Example: The quiz average for a class is 7.5. The teacher predicts that if classical music is playing in the background the quiz scores will increase. She tries this out with 5 students and they score 7, 8, 8, 9, and 9. Using alpha set at 0.05 see if the teacher's hypothesis is supported.

1 & 2) Hypotheses and criterion for decision

Ho: $\bar{X} \leq 7.5$, Mean with music less than or equal to pop. mean.

Ha: $\bar{X} > 7.5$, Mean with music greater than population mean.



Critical region: a. $\alpha = 0.05$
 b. 1-tailed (upper)

t-critical = 2.132

3) Sample statistics

Student	Score	$X - \bar{X}$	$(X - \bar{X})^2$
A	7	-1.2	1.44
B	8	-0.2	0.04
C	8	-0.2	0.04
D	9	0.8	0.64
E	9	0.8	0.64
Σ	41	0	2.8
\bar{X}	8.2		

Mean $\bar{X} = \Sigma X/N = 41/5 = 8.2$

Sum of Squares $SS = \Sigma (X - \bar{X})^2 = 2.8$

Standard deviation $s = \sqrt{\frac{SS}{n-1}} = \sqrt{2.8/4} = \sqrt{0.7} = 0.8366$

Standard error $s_{\bar{X}} = \frac{s}{\sqrt{n}} = 0.8366/\sqrt{5} = 0.8366/2.236 = 0.374$

$df = 5-1 = 4$

4) Test statistic

One-sample t-observed $t = \frac{\bar{X} - \mu}{s_{\bar{X}}} = (8.2 - 7.5)/0.37 = 1.89$

5) Compare observed to critical test value: $t_{crit} = 2.132$

t-observed < t-critical, that is, it is not in the critical region

6) Decide about null hypothesis: **fail to reject null hypothesis**

7) Conclude about relationship: **Evidence not support positive effect of classical music on quiz scores.**

Worksheet: Paired-Samples t-test

Example: Here are data comparing college student's motivation scores before and after Thanksgiving break to see if there is an effect of a week off school.

Student	Before	After	D(A-B)	$D - \bar{D}$	$(D - \bar{D})^2$
A	65	70	5	2.8	7.84
B	68	69	1	-1.2	1.44
C	50	55	5	2.8	7.84
D	75	73	-2	-4.2	17.64
E	80	82	2	-0.2	0.04
Σ			11	0	$\Sigma SS_D^2 = 34.76$
\bar{D}			2.2		

Note: the order of subtraction (A-B) will produce positive scores if studying increased after break.

1 & 2) Hypotheses and criterion for decision

$H_0: \mu_D = 0$, There is no difference between the two periods.

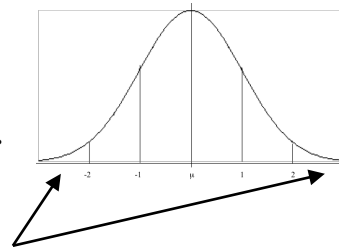
$H_a: \mu_D \neq 0$ or There is a difference between the two periods.

Critical region:

a. $\alpha = 0.05$

b. 2-tailed

t-critical = 2.776



3) Sample statistics

Mean of differences $\bar{D} = \frac{\sum D}{n} = \frac{11}{5} = 2.2$

Sum of Squares $SS_D = \sum (D - \bar{D})^2 = 34.76$

Standard deviation

$$s_D = \sqrt{\frac{SS_D}{n_D - 1}}$$

$$= \sqrt{34.76/4} = \sqrt{8.69} = 2.95$$

Standard error $s_{\bar{D}} = \frac{s_D}{\sqrt{n_D}} = 2.95/\sqrt{5} = 1.32$

$df = 5-1 = 4$

4) Test statistic: Remember that the mean (μ_D) we are testing our hypothesis against is zero (0) because H_0 predicts no difference.

Related-Samples observed-t $t_{\bar{D}} = \frac{\bar{D} - \mu_{\bar{D}}}{s_{\bar{D}}} = 2.2 - 0/1.32 = 1.67$

5) Compare observed to critical test value: $t_{crit} = 2.776$

t-observed < t-critical, that is, it is not in the critical region

6) Decide about null hypothesis: **fail to reject null hypothesis**

7) Conclude about relationship: **Evidence not support any difference between motivation scores during two periods.**

Worksheet: Independent-Samples t-test

Example: Here is data illustrating the motivation scores of students in a required course compared to scores of a different group of students in an elective course at the same level. Let's test the hypothesis that students in elective courses are more motivated than those in required courses. Assume alpha = 0.05

X ₁ (Req)	X ₁ -M	(X ₁ -M) ²	X ₂ (Elec)	X ₂ -M	(X ₂ -M) ²
5	1	1	2	-0.4	0.16
3	-1	1	4	1.6	2.56
4	0	0	1	-1.4	1.96
5	1	1	3	0.6	0.36
3	-1	1	2	-0.4	0.16
$\sum X_1 = 20$	0	SSX ₁ = 4	$\sum X_2 = 12$	0	SSX ₂ = 5.2
$\bar{X}_1 = 4$			$\bar{X}_2 = 2.4$		

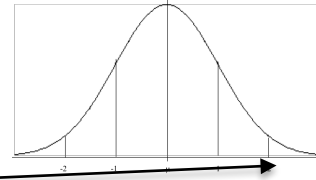
1) Hypotheses and criterion for decision

H₀: X₂ - X₁ ≤ 0, The motivational difference between elective and required courses is less than or equal to zero.

Ha: $X_2 - X_1 > 0$, The motivational difference between elective and required courses is greater than zero.

NOTE: SINCE ELECTIVE SCORES ARE LOWER, THE NULL HYPOTHESIS IS NOT GOING TO BE REJECTED!

Critical area: a. $\alpha = 0.05$
 b. 1-tailed (upper)



critical value **2.132**

3) Sample statistics

	X_1	X_2
Mean	$\bar{X} = \frac{\sum X}{n} = 20/5 = 4$	$= 12/5 = 2.4$
Sum of Squares	$SS = \sum (X - \bar{X})^2 = 4$	$= 5.2$
Degrees of freedom	$df = n - 1 = 4$	$= 4$
Pooled variance	$s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2} = (4 + 5.2)/4+4 = 9.2/8 = 1.15$	
Standard error	$s_{(\bar{X}_1 - \bar{X}_2)} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = \sqrt{1.15/5 + 1.15/5} = \sqrt{2.3/5} = \sqrt{0.46} = 0.678$	

(When adding fractions, do not add the denominators!)

4) Test statistic

(Remember that the difference between means (μ_1 & μ_2) in Ho is zero (0)).

Independent-Samples observed t

$$t_{obs} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{(\bar{X}_1 - \bar{X}_2)}} = ((2.4 - 4) - 0)/0.678 = -1.6/0.678 = -2.35$$

5) Compare observed to critical test value: $t_{\text{crit}} = 2.132$

t-observed is not in the critical upper region (it's all the way at the other end).

6) Decide about null hypothesis: **cannot be rejected.**

7) Conclude about relationship: **The evidence does not support that motivational scores are higher in elective courses.**

Practice Question Solutions: Set 13

(1a) $b = SP/SS_x = 30/15 = 2$
 $a = \bar{Y} - b\bar{X} = 5 - 2*4 = -3$
 $\hat{Y} = bX + a = 2X - 3$

(1b) $X \quad \hat{Y} = 2X - 3$
 3 3
 -2 -7
 5 7
 6 9

(2)

	X	Y	X- \bar{X}	Y- \bar{Y}	CrossProd	(X- \bar{X}) ²
	0	9	-1	0	0	1
	1	7	0	-2	0	0
	2	11	1	2	2	1
Sum	3	27	0	0	2 = SP	2 = SS _x
Mean	1	9				

$b = 2/2 = 1$
 $a = 9 - 1*1 = 8$
 $\hat{Y} = X + 8$

(3) Workspace as above plus for Y-hat, predictive error, and squared error

	X	Y	X- \bar{X}	Y- \bar{Y}	CP	(X- \bar{X}) ²	\hat{Y}	Y- \hat{Y}	(Y- \hat{Y}) ²
	4	1	-1	-6	6	1	4	-3	9
	7	16	2	9	18	4	13	3	9
	3	4	-2	-3	6	4	1	3	9
	5	7	0	0	0	0	7	0	0
	6	7	1	0	0	1	10	-3	9
Sum	25	35	0	0	30	10		0	36 = SS _{error}
Mean	5	7							

$b = 30/10 = 3$
 $a = 7 - 3*5 = -8$
 $\hat{Y} = 3X - 8$ (Insert X values to get \hat{Y} values.)
 $s_{est} = \sqrt{SS_{error}/df} = \sqrt{36/3} = \sqrt{12} = 3.46$

(4) **c** When a correlation is close to ± 1.0 , then the standard error of the estimate (SEE) will be **small**. When the correlation is close to 0, then SEE will be **large**.

Practice Question Solutions: Set 14

(1)

	Obs	Exp	Obs	Exp	Obs	Exp	Marg
	10	17.5	25	24.5	35	28	70
	15	7.5	10	10.5	5	12	30
Marg	25		35		40		100

Ho: Pets & allergies are independent; Ha: Pets & allergies are not independent.

$$\chi^2_{crit} (df = 2, \alpha = .05) = 5.99$$

$$\chi^2_{obs} = (f_o - f_e)^2 / f_e = 7.5^2 / 17.5 + 7.5^2 / 7.5 + .5^2 / 24.5 + .5^2 / 10.5 + 7^2 / 28 + 7^2 / 12 = 16.58$$

Observed > critical value, so reject null hypothesis; support previous findings that more pets related to fewer allergies

(2)

	Obs	Exp	Obs	Exp	Marg
	Bush		Kerry		
Male	10	8	5	7	15
Female	6	8	9	7	15
Marg	16		14		30

Ho: Gender & voting are independent; Ha: gender & voting are not independent.

$$\chi^2_{crit} (df = 1, \alpha = .01) = 6.64$$

$$\chi^2_{obs} = (f_o - f_e)^2 / f_e = 2^2 / 8 + 2^2 / 7 + 2^2 / 8 + 2^2 / 7 = 2.14$$

Observed < critical value, so fail to reject null hypothesis; evidence does not support a relationship between gender and voting

Practice Question Solutions: Set 15

- (1) Population standard deviation is unknown, so use t-test
 $\mu = \bar{X} \pm t_{crit} s_{\bar{x}}$, find value in t table for $\frac{1}{2}(1.00-CI)$, which is % in each tail (.05 1-tailed in a & .10 1-tailed in b)

(a) 90% CI: $200 \pm 1.711 (20/\sqrt{25}) = 200 \pm 6.8 = 193.2$ to 206.8

(b) 80% CI: $200 \pm 1.318 (20/\sqrt{25}) = 200 \pm 5.3 = 194.7$ to 205.3

(c) As confidence level increases, so does the critical test value, and so the confidence interval becomes **wider** (confidence-accuracy trade-off).

- (2) (a) 95% CI: $200 \pm 1.98 (20/\sqrt{64}) = 200 \pm 4.95 = 194.05$ to 204.95
 Use higher df, 120

(c) 95% CI: $200 \pm 2.021 (20/\sqrt{36}) = 200 \pm 6.74 = 193.26$ to 206.74
 Use higher df, 40

(c) You get a better estimate (narrower interval, less error) with a larger sample size.

- (3) $\mu_A - \mu_B = \bar{X}_A - \bar{X}_B \pm t_{crit} (s_{\bar{X}_a - \bar{X}_b})$, find value in t table for $\frac{1}{2}(1.00-CI)$

(a) point estimate = 5

(b) t_{crit} (2-tailed, .025, df = 23) = 2.069

$s^2_{pooled} = (2030+1420)/(14+9) = 150$

$SE = \sqrt{150/15 + 150/10} = \sqrt{25} = 5$

95% CI: $-12 \pm 2.069 (5) = -12 \pm 10.345 = -22.3$ to -1.7

(d) Since the population mean ($\mu = 0$) for the null hypothesis falls outside 95% CI, he can reject the null hypothesis and conclude there is a difference between the groups.

Problem Solutions: Study Guide for Exam 4 and Final Exam

Worksheet: Finding critical values in distribution tables

r, 1-tailed, $\alpha = 0.05$, $df = 28$; $r = .306$

r, 1-tailed, $\alpha = 0.01$, $df = 28$; $r = .423$

χ^2 , $\alpha = 0.05$, $df = 2$; $\chi^2 = 5.99$

r, 2-tailed, $\alpha = 0.05$, $df = 28$; $r = .361$

r, 2-tailed, $\alpha = 0.01$, $df = 28$; $r = .463$

χ^2 , $\alpha = 0.01$, $df = 2$; $\chi^2 = 9.21$

Sample problems

1. $H_0: \bar{X} \geq \mu$; $H_A: \bar{X} < \mu$; z_{crit} (1-tail, .05) = -1.64; (relevant formulas below)

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \quad z_{\bar{X}} = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \quad d = \frac{\bar{X} - \mu}{\sigma}$$

For 30 students: $SE = 100/\sqrt{30} = 18.25$; $z = (480-500)/18.25 = -1.096$. Since the observed z is less extreme than the critical z , fail to reject null hypothesis and conclude that evidence does not support the class being below the population average on the SAT.

For 100 students: $SE = 100/\sqrt{100} = 10$; $z = (480-500)/10 = -2$. Since this observed z is more extreme than the critical z , reject the null hypothesis and conclude that the classes are below the population average on the SAT. The effect size is $20/100$, which is .2, which is small.

$$\mu = \bar{X} \pm z_{crit}(\sigma_{\bar{X}}) = 20 \pm 1.96(10) = 20 \pm 19.6 = .4 \text{ to } 39.6$$

Remember that 95% CI means 5% error total, so z_{crit} is for .025, 1-tailed. The null hypothesis mean of 0 is just outside the CI. A 2-tailed hypothesis test with $z_{obs} = -2$ would just barely exceed $z_{crit} = 1.96$.

Extra credit: The denominator gets smaller with 100 students, and so the z -value gets larger. Since a 20-point difference is small, given that 1 SD = 100, it will only be significant with a large n . For $n = 20$, the standard error is 18.25, almost as big the difference found. For $n = 100$, SE has been reduced to 10, so the difference is now twice as large.

2. $H_0: \bar{X} \geq \mu$; $H_A: \bar{X} < \mu$; t_{crit} (1-tail, .05, 29df) = -1.699.

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = 5/\sqrt{30} = 0.91 \quad t = \frac{\bar{X} - \mu}{s_{\bar{X}}} = 98-100/0.91 = -2.198 \quad d = \frac{\bar{X} - \mu}{s} = \frac{2}{5} = .4$$

Since the observed t is more extreme than the critical t , reject the null hypothesis and conclude that the class is below the population average on this test. The effect size is moderate.

$$\mu = \bar{X} \pm t_{crit}(s_{\bar{X}}) = 2 \pm 2.045(.91) = 2 \pm 1.86 = .14 \text{ to } 3.86$$

Remember that 95% CI means 5% error, so t_{crit} is for .025, 1-tailed. The null hypothesis mean of 0 is again just outside the CI. A 2-tailed hypothesis test with $t_{obs} = -2.198$ would just barely exceed $t_{crit} = 2.045$.

Extra credit: As the effect size shows, the 2 points is 40% of its SD, while the 20 points about is only 20% of its SD.

3. $H_0: \bar{D} \leq 0; H_A: \bar{D} > 0; t_{crit} (1\text{-tail}, .05, 29df) = 1.699$

$$s_{\bar{D}} = \frac{s_D}{\sqrt{n_D}} = 1/\sqrt{30} = 0.1825; \quad t_{\bar{D}} = \frac{\bar{D} - \mu_{\bar{D}}}{s_{\bar{D}}} = 2-0/0.1825 = 10.95 \quad d = \frac{\bar{D}}{s_D} \quad 2/1 = 1$$

Since the observed t is more extreme than the critical t, reject the null hypothesis and conclude that the class is studying more after the midterm. The effect size is very large.

$$95\% \text{ CI } \mu_{\bar{D}} = \bar{D} \pm (t_{crit})(s_{\bar{D}}) = 2 \pm 2.045(.1825) = 2 \pm .37 = 1.63 \text{ to } 2.37$$

\bar{D} of H_0 (0) does not fall within confidence interval (which will always the case when the same α is used for both tests and H_0 is rejected); instead \bar{D} is the mean of the second population predicted by H_A .

Extra credit: As in the problems above, the critical t is not the same for a 1-tailed .05 hypothesis test and a 95% CI mean estimate, which has a total of 5% error or .025, 1-tail. The critical t for a 1-tailed .05 hypothesis test is the same as that for 90% CI, which has a total of 10% error or .05 in each tail.

4. $H_0: \bar{X}_1 - \bar{X}_2 = 0; H_A: \bar{X}_1 - \bar{X}_2 \neq 0; t_{crit} (2\text{-tailed}, .05, 28df) = \pm 2.048.$

$$s = \sqrt{\frac{SS}{n-1}} = \sqrt{170/14} = 3.48 \quad s = \sqrt{\frac{SS}{n-1}} = \sqrt{200/14} = 3.78$$

$$s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2} = (170+200)/(14+14) = 13.21 \quad s_{(\bar{X}_1 - \bar{X}_2)} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = \sqrt{(13.21+13.21)/15} = 1.33$$

$$t_{obs} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{(\bar{X}_1 - \bar{X}_2)}} = ((80-78) - 0)/1.33 = 1.5$$

Since the observed t is less extreme than the critical t, fail to reject the null hypothesis and conclude that the class the evidence does not support a difference between sections on the exam.

Extra credit: The 2-point mean difference for a mathematics test was with $s = 5$, while here it is with pooled $s = 3.63$. The effect size is larger here. However, this is a less powerful design. The pooled SE is much larger (1.33 vs. .91), thereby reducing the size of the observed t (1.5 vs. 2.198), and the 2-tailed hypothesis test increases the critical test value (2.048 vs. 1.699).

5. $H_0: r \leq 0; H_A: r > 0; r_{crit} (1\text{-tail}, .05, 28df) = 0.306.$

$$r = \frac{SP}{\sqrt{SS_X SS_Y}} = 25/\sqrt{(30*185)} = 25/74.5 = 0.336$$

Since the observed r is more extreme than the critical r , reject the null hypothesis and conclude that studying is positively related to exam scores. The amount of variance is r^2 , which is 11%.

$$b = \frac{SP}{SS_x} = 25/30 = 0.833 \quad a = \bar{Y} - b\bar{X} = 79 - .833(5) = 74.835$$

$$\hat{Y} = (X)(\text{slope}) + (\text{intercept}) = bX + a = .833X + 74.835$$

$$s_{est} = \sqrt{\frac{SS_{error}}{df}} = \sqrt{270/28} = 3.11$$

$$\text{If } X = 0, \hat{Y} = .833(0) + 74.835 = 74.835$$

$$\text{If } X = 3, \hat{Y} = .833(3) + 74.835 = 77.334$$

If $X = 5$, $\hat{Y} = .833(5) + 74.835 = 79$ (mean of X predicts mean of Y ; variant of formula for intercept)

$$\text{If } X = 10, \hat{Y} = .833(10) + 74.835 = 83.165$$

Extra credit: This is because of regression to the mean. While the X mean always predicts the Y mean, other X scores predict Y scores closer to the Y mean, and the discrepancy gets larger as X scores get farther from the mean

6. H_0 : Grades are independent of graduating as a psych major;
 H_A : Grades are not independent of graduating as a psych major;
 χ^2 crit (.05, 4df) = 9.49.

Psych Graduate	Grades										Marginal
	A		B		C		D		F/W		
	Obs	Exp	Obs	Exp	Obs	Exp	Obs	Exp	Obs	Exp	
Yes	20	10.42	30	22.92	20	25	5	12.5	0	4.17	75
No	5	14.58	25	32.08	40	35	25	17.5	10	5.83	105
Marginal	25		55		60		30		10		180

Expected: YesA: $75 \cdot 25 / 180$; YesB: $75 \cdot 55 / 180$; YesC: $75 \cdot 60 / 180$; YesD: $75 \cdot 30 / 180$;
 YesF: $75 \cdot 10 / 180$; NoA: $105 \cdot 25 / 180$; NoB: $105 \cdot 55 / 180$; NoC: $105 \cdot 60 / 180$; NoD:
 $105 \cdot 30 / 180$; NoF: $105 \cdot 10 / 180$

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 8.81 + 2.19 + 1 + 4.5 + 4.17 + 6.29 + 1.56 + 0.71 + 3.21 + 2.98 = 35.42$$

Since this observed χ^2 is more extreme than the critical χ^2 , reject the null hypothesis and conclude that grades in statistics are related to graduating with a psychology major.

VII. Summary of Formulas

Univariate Statistics	For a population	For a sample
Mean	$\mu = \Sigma \frac{X}{N}$	$\bar{X} = \Sigma \frac{X}{n}$
Sum of squares	$SS = \Sigma(X - \mu)^2$	$SS = \Sigma(X - \bar{X})^2$
Variance	$\sigma^2 = \frac{SS}{N}$	$s^2 = \frac{SS}{n-1}$
Standard deviation	$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{SS}{N}}$	$s = \sqrt{s^2} = \sqrt{\frac{SS}{n-1}}$
z-score	$z = \frac{X - \mu}{\sigma}$	$z = \frac{X - \bar{X}}{s}$

Bivariate Statistics

Sum of the products	$SP = \Sigma(X - \bar{X})(Y - \bar{Y})$
Pearson's correlation coefficient	$r = \frac{SP}{\sqrt{SS_x SS_y}} \quad \frac{\Sigma z_x z_y}{n-1}$
Degrees of freedom	$df = n-2$
Regression line	$\hat{Y} = (X)(\text{slope}) + (\text{intercept}) = Xb + a = bX + a$
Slope	$b = \frac{SP}{SS_x}$
Intercept	$a = \bar{Y} - b\bar{X}$
Total squared error	$SS = (Y - \hat{Y})^2$
Standard error of estimate	$s_{est} = \sqrt{\frac{SS_{error}}{df}}$

Hypothesis Testing and Parameter Estimation

z-test

Standard error (σ known)	$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
	$z_{\bar{X}} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$ <p style="text-align: right;">z-observed</p>
Effect size	$d = \frac{\bar{X} - \mu}{\sigma}$
Parameter estimate (σ known)	$\mu = \bar{X} \pm z_{crit}(\sigma_{\bar{X}})$

One-Sample t-test

Degrees of freedom	$n - 1$
Standard error (σ unknown)	$s_{\bar{X}} = \frac{s}{\sqrt{n}}$
One-sample t-observed	$t = \frac{\bar{X} - \mu}{s_{\bar{X}}}$
Effect size	$d = \frac{\bar{X} - \mu}{s}$
Parameter estimate (σ unknown)	$\mu = \bar{X} \pm t_{crit}(s_{\bar{X}})$

Related-Samples t-test

Degrees of freedom	$n_D - 1$
Mean of differences	$\bar{D} = \frac{\sum D}{n}$
Sum of squares of differences	$SS_D = \sum (D - \bar{D})^2$
Standard deviation of differences	$s_D = \sqrt{\frac{SS_D}{n_D - 1}}$

Standard error of differences $s_{\bar{D}} = \frac{s_D}{\sqrt{n_D}}$

Related-Samples observed-t $t_{\bar{D}} = \frac{\bar{D} - \mu_{\bar{D}}}{s_{\bar{D}}}$

Effect size $d = \frac{\bar{D}}{s_D}$

Parameter estimate (related samples) $\mu_D = \bar{D} \pm t_{crit}(s_{\bar{D}})$

Independent-Samples t-test

Degrees of freedom $df_1 = (n_1 - 1), df_2 = (n_2 - 1)$
 $df_{total} = df_1 + df_2 = n_1 + n_2 - 2$

Pooled variance of independent samples $s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2}$ $s_p^2 = \frac{s_1^2 + s_2^2}{2}$
 (2nd averaging formulas if $n_1 = n_2$)

Standard error of independent samples $s_{(\bar{X}_1 - \bar{X}_2)} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$ $s_{(\bar{X}_1 - \bar{X}_2)} = \sqrt{\frac{2s_p^2}{n}} = s_p \sqrt{\frac{2}{n}}$
 (2nd formula if $n_1 = n_2$)

Independent-Samples observed t $t_{obs} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{(\bar{X}_1 - \bar{X}_2)}}$

Effect size $d = \frac{\bar{X}_2 - \bar{X}_1}{s_p}$

Parameter (independent samples) $\mu_1 - \mu_2 = \bar{X}_1 - \bar{X}_2 \pm t_{crit}(s_{(X_1 - X_2)})$

Chi-Square test

Estimated cell frequencies $f_e = \frac{f_{column} f_{row}}{n}$ OR $f_e = \frac{f_{row}}{n} * f_{column}$

Observed chi-square $\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$

Degrees of freedom $df = (\#columns - 1) * (\#rows - 1)$