

INTRODUCTION TO THE STUDY OF STATISTICS

How much coffee, on average, do Clemson University students drink every day? Does a specific medicine cure patients who have lung cancer caused by smoking? Does the drug Remdesivir help patients diagnosed with COVID-19? Are there any relationships between the size a vehicle and the chance of having an automobile accident? There are thousands and thousands of similar vital questions and challenges that arise in life. It might be relatively easy to answer such questions for a small experimental group, but could we simply generalize that answer to larger groups? Probably not! Generalizing results from a small group of subjects or individuals to a larger group is not wise unless some conditions hold, and circumstances are considered. Our plan and goal in this introductory statistics course, is to introduce these conditions and circumstances so that we can answer some exciting real-world questions.

MAJOR THEMES OF STAT 3090: INTRODUCTORY BUSINESS STATISTICS

- **Basic Data Analysis:** A discussion on gathering, displaying, and summarizing data
- **Probability:** A study of the laws of chance
- **Statistical Inference:** The science of drawing statistical conclusions from specific data using a knowledge of probability.

SECTION 1: IDENTIFYING AND GATHERING DATA

(HAWKES 1.1, 1.3, 2.1, 2.3, 8.4)

One of the most important parts of a statistical study is data. Data are the evidence in all studies and without them, it is almost impossible to answer statistical questions regarding the study. For that matter, it is as important to learn how to classify, collect and use data in our study. In this section we learn the sources of the data, how to classify them, and finally how to collect them. Once we have them, we can turn our data into usable information.

1.1 POPULATION VS SAMPLE

In statistics, our ultimate goal is to uncover truths about a population. A population is the set of all subjects or elements about which we are interested in making decisions (inferences). A **frame** or **data frame** is a list of the members of the population.

A sample is a subset of the population which is used to gain insight about the population. Samples are used to represent the larger population of interest.

A sample is chosen because the population of interest is almost always too large to survey, and typically it would be too expensive to collect data from the entire population.

One exception to this is a census, in which we collect data from every subject in a population.

Only random samples are likely to be useful in statistics, as they give us the best chance of obtaining a sample that is representative of the population. We will have a discussion on gathering random samples in this section. We will save a more in-depth study of probability for later in the course.

Example 1.1: Automated Workers

According to a Gallup poll conducted in April 2017, 26% of U.S. workers say that it is at least somewhat likely that their job will be eliminated by new technology, automation, artificial intelligence or robots within the next 20 years. Results for this Gallup poll were based on telephone interviews of a random sample of 1,127 workers, aged 18 and older, living in all 50 U.S. states and the District of Columbia.

Identify the **population** and the **sample** in the following study.

Population:

All workers aged 18 and older living in all 50 U.S. states and the District of Columbia

Sample:

The 1,127 workers who were polled.

Example 1.2: abcnews.com users

The August 23, 1999 issue of the Tampa Tribune reported on a study involving data volunteered by 17,251 users of the abcnews.com web site. Users of the site were asked to respond to questions including whether they used the Internet to escape problems, tried unsuccessfully to cut back their internet use, and found themselves preoccupied with the Internet even when not using it. Almost 6% of those responding confessed to some sort of addiction to the Internet.

Identify the **population** and the **sample** in the following study.

Population:

All abc news website users.

Sample:

The 17,251 volunteers responded to the study.

1.2 STATISTICS – TURNING DATA INTO INFORMATION

Now that we understand the concept of a population and the importance of samples, we will divide statistical techniques into two main categories: *descriptive* statistics and *inferential* statistics.

Descriptive statistics extract observable information and characteristics from populations and samples. This is the topic that we will discuss in the first part of our course.

Inferential statistics focuses on making decisions, generalizations and inferences about populations using information gathered from samples. We will discuss this topic in the latter part of our course. To reach an understanding of inferential statistics, an understanding of descriptive statistics is needed, which is why we will study descriptive statistics before inferential statistics.

We will begin our discussion of how to turn data into information by comparing the definitions and uses of descriptive and inferential statistics. We will then consider two important terms used in summarizing data that you will use throughout the course.

1.2.1 DESCRIPTIVE VS INFERENCE STATISTICS

Descriptive statistics is analysis of data that simply describes the sample or population of interest. Most of the time, we will want to analyze observed measurements from a *sample* of data. Descriptive statistics can allow us to answer questions such as the following¹:

- What is a **typical value** for the measurements?
- How much **variation** do the measurements possess?
- What is the **shape** or **distribution** of the measurements?
- Are there any **extreme values** in the measurements? If so, what does that tell us?
- What is the **relative position** of a particular measurement in the group of data?
- What kind of **relationship** exists, if any, when there are two variables for which we have measurements? How strong is this relationship?

If we only have measurements from two observations, then comprehending the data is straightforward. However, if we have measurements from many observations, then inspecting individual observations provides little useful knowledge. To comprehend a large set of data, we use descriptive statistics to organize, analyze, and present these data.

Inferential statistics, on the other hand, is an estimate or prediction or some other generalization about a population based on information contained in a sample. While it would be preferable to have measurements from the entire population, this is, in most cases, infeasible either because all the data are not obtainable or are too costly to gather. The goal of inferential statistics is to make reasonable guesses about population characteristics using sample data.

1.2.2 PARAMETER VS STATISTIC

A parameter is a numerical descriptive measure or numerical summary of a population. It is usually unknown since the population is too large to have a full collection of data, but it is often inferred from the sample data. Since parameters are descriptions of the population, a population can have many parameters.

A statistic is a numerical descriptive measure of a sample. It is calculated from the data in the sample, and thus is a known value.

¹ Pages 9-10 of *Discovering Business Statistics* by Quinton J. Nottingham and James S. Hawkes.

Example 1.3: Automated Workers (continued)

According to a Gallup poll conducted in April 2017, 26% of U.S. workers say that it is at least somewhat likely that their job will be eliminated by new technology, automation, artificial intelligence or robots within the next 20 years. Results for this Gallup poll were based on telephone interviews of a random sample of 1,127 workers, aged 18 and older, living in all 50 U.S. states and the District of Columbia.

Identify a **statistic** and define the **parameter** of interest associated with it.

Statistic:

26% of U.S. workers in the sample who believe it is somewhat likely that their job will be eliminated by technology in the next 20 years

Parameter:

The true proportion of U.S. workers who believe it is somewhat likely that their job will be eliminated by technology in the next 20 years

Example 1.4: abcnews.com users (continued)

The August 23, 1999 issue of the Tampa Tribune reported on a study involving data volunteered by 17,251 users of the abcnews.com web site. Users of the site were asked to respond to questions including whether they used the Internet to escape problems, tried unsuccessfully to cut back their internet use, and found themselves preoccupied with the Internet even when not using it. Almost 6% of those responding confessed to some sort of addiction to the Internet.

Identify a **statistic** and define the **parameter** of interest associated with it.

Statistic:

6% of the users responding confessed to some sort of addiction to the internet.

Parameter:

The true proportion of the users of abcnews.com web site that are addicted to the internet.

1.2.3 THE BIG IDEA

There is an unknown population parameter about which we wish to know something. We can estimate it or answer a question about it by using the information gathered from the sample and summarized by a statistic.

Example 1.5: Automated Workers (continued)

According to a Gallup poll conducted in April 2017, 26% of U.S. workers say that it is at least somewhat likely that their job will be eliminated by new technology, automation, artificial intelligence or robots within the next 20 years. Results for this Gallup poll were based on telephone interviews of a random sample of 1,127 workers, aged 18 and older, living in all 50 U.S. states and the District of Columbia.

A. What part of the study involves **descriptive** statistics?

Gathering responses from the 1,127 U.S. workers in the sample and summarizing the statistic of 26% with the characteristic of interest.

B. How could the descriptive statistic(s) identified in part A be used as **inferential** statistics?

Using the statistic of 26% to make a generalization about the population of all U.S. workers by approximating the parameter of interest, the true percentage of all U.S. workers who believe their jobs will be eliminated by new technology.

Example 1.6: abcnews.com users (continued)

The August 23, 1999 issue of the Tampa Tribune reported on a study involving data volunteered by 17,251 users of the abcnews.com web site. Users of the site were asked to respond to questions including whether they used the Internet to escape problems, tried unsuccessfully to cut back their internet use, and found themselves preoccupied with the Internet even when not using it. Almost 6% of those responding confessed to some sort of addiction to the Internet.

- A. What part of the study involves **descriptive** statistics?

Gathering responses from the 17,251 abcnews.com users and summarizing the statistic of 6% with the characteristic of interest is the part of the study that involves descriptive statistics.

- B. How could the descriptive statistic(s) identified in part A be used as **inferential** statistics?

We can use the statistic of 6% to make a generalization about the population of all abc news web site users. We can use it to approximate the parameter of interest, the true percentage of all abc news website users who have some sort of addicted to the internet.

MY NOTES:

1.3 CLASSIFYING DATA

As we discussed before, the goal of statistics is to reveal information about the population. Most of the time, populations are too large and we cannot feasibly gather data from every individual or subject in a population. Therefore, we select samples from the populations and gather data only from the sample. Sample data is simply data taken from the corresponding population we want to study. In this section, we learn how to classify data, and how to gather data for our samples.

1.3.1 WHAT ARE DATA

Though we tend to use the words *data* and *information* interchangeably in normal conversation, we need to think of them as different things when we are thinking about statistics.

Data are the raw values we gather from a sample before we perform any act on them.

Information is a meaningful summary of the values gathered in the data.

For example, a list of scores that each student earned on a statistics test last semester is data. If we were to summarize the scores by computing the mean (average) score or producing a table showing how many students earned each letter grade, we have turned the data into information.

Variables are the characteristic(s) recorded about each individual from a sample or population. These recorded characteristics are our data.

Example 1.7: Amazon Orders

There are 6 variables (characteristics) about an individual Amazon order listed below.

Order Number	Name	State	Unit Price	Zip Code	Gift?
23432	Meg H	SC	10.99	29672	Y
83566	April T	SC	135.8	29630	N
31078	Thomas I	GA	5.03	30037	N
40092	Ali G	VA	83.56	22907	N
93563	Johnny D	CA	30.80	92178	Y
:	:	:	:	:	:

- A. What values will the variable **Order Number** have? *Numbers associated with the order*
- B. What values will the variable **Name** have? *The names of customers*
- C. What values will the variable **State** have? *2-letter abbreviations for U.S. states*
- D. What values will the variable **Unit Price** have? *Dollar values for the amount spent*
- E. What values will the variable **Zip Code** have? *5-digit zip codes*
- F. What values will the variable **Gift** have? *Yes or No responses*

1.3.2 CLASSIFYING VARIABLES

Let's study some properties of variables that we just defined in our last example. These properties are critical in our upcoming discussions on random variables in the probability section. In the table of the Example 1.7 Amazon Orders, we observed several variables like name, state, unit price etc. Not all these variables are of the same type. In fact, we classify any variable into one the two main types: qualitative and quantitative.

A variable is qualitative (or categorical) if each observation belongs to one of a set of categories. It might consist of labels or descriptions but does not have naturally-occurring numerical values.

A variable is quantitative if observations take on numerical values that represent magnitudes or quantities of the variable (generally something we can count or measure).

Example 1.8: Amazon Orders (continued)

All of the following variables are qualitative except for Unit Price.

Order Number	Name	State	Unit Price	Zip Code	Gift?
23432	Meg H	SC	10.99	29672	Y

How can we tell?

Order Number and Zip Code, even though they are numbers, are simply categories. You would not add Order Numbers together or take an average of Zip Codes and get something meaningful. However, Unit Price is a measurement of value and is thus a quantitative variable.

Quantitative variables can be further classified into two types:

Discrete data are quantitative data that are restricted to a set of values with gaps (such as 0, 1, 2, 3, 4, ...). We say these values are **countable**.

Continuous data are quantitative data that can take on any value within some interval (such as measurements).

Example 1.9: Discrete and Continuous Variables

Determine whether each of the following variables is **discrete** or **continuous**.

- A. The number of doctors who wash their hands between patient visits. *Discrete*
- B. Amount of water consumed by the average American each day. *Continuous*
- C. The time it takes to react to a stimulus. *Continuous*
- D. The number of South Carolina voters who favor a particular candidate. *Discrete*

1.3.3 LEVELS OF MEASUREMENT

Both qualitative and quantitative data can be further classified into **levels of measurement**.

For **qualitative** variables...

- Nominal data represent whether a variable possesses some characteristic or falls into a certain category.
- Ordinal data represent categories that have some associated order.

For **quantitative** variables...

- Interval data are data that can be ordered and for which the arithmetic difference is meaningful. For this type of data, there is no meaningful zero point (i.e. a zero value does **not** mean the absence of the value, or “none”). It makes sense to consider the difference between two values, but not to divide one value by another.
- Ratio data are data that can also be ordered and for which the arithmetic difference is meaningful. This type of data has a meaningful zero point (i.e. zero means “none”), and data can be compared as a ratio (i.e. it makes sense to talk about one value being a multiple of another).

Example 1.10: Levels of Measurement

Determine the **level of measurement** for each of the following variables.

- A. SAT score *Interval (the minimum SAT score is 400)*
- B. Rating from 1 to 5 of quality of service in a restaurant *Ordinal*
- C. Make and model of a vehicle *Nominal*
- D. Number of students with a business major *Ratio*

1.4 DATA COLLECTION

While variable classification is important, another vital piece to a statistical study is how the information is collected. Here we first discuss the techniques that researchers and statisticians use to collect data, then we learn two important types of studies and the purpose of each type.

1.4.1 SAMPLING TECHNIQUES

The goal sampling is to choose a sample representative of the entire population.

1. Simple Random Sample: This is a sample of size n selected from a population in such a way that each individual in the sample has an equally likely chance of being chosen. In addition, the sample is selected so that each possible *sample* of size n has an equally likely chance of being selected. This is the sampling method we will use in this course.

Steps to obtain a simple random sample:

- (1) Obtain a list of observations (a data frame) for the entire population. This can be a list of phone numbers, names, addresses, etc.
- (2) Label each observation with a number.
- (3) Obtain random numbers, which may be generated using computer programs, calculator functions, by using a random number table, or by selecting numbers from a hat.
- (4) The first n unique numbers that are selected are the numbers of the observations to select. No duplicates may be selected; if a duplicate number is obtained, skip it and obtain a new random number.

Example 1.11: Leash Law

In a small town, the legislature is interested in passing a new leash law for dog owners. They decide to poll the community to investigate the potential for the new law to pass.

A. How might you obtain a **simple random sample** of households?

Use the tax records for the town as a data frame. Randomly select households from the tax records to participate in the poll.

B. What is the **variable** of interest?

Whether or not the household would vote for the new law.

C. Is the variable **qualitative** or **quantitative**? *It is Qualitative*

D. What **values** will the variable of interest have? *Yes, No, or Undecided*

There are additional methods for obtaining random samples which do not result in a simple random sample. These methods are commonly used in experimental design. If these methods are utilized, the inference procedures that are used in this course may not be appropriate.

2. Stratified Random Sampling : To obtain this type of sample, we divide the population of interest into non-overlapping groups called **strata** and then randomly select from each stratum.

- **Advantage:** The sample is representative of each stratum within a population.
- **Caution:** It is better to choose the strata so that the elements in each stratum are as much alike as possible.

Example 1.12: Leash Law (continued)

In a small town, the legislature is interested in passing a new leash law for dog owners. They decide to poll the community to investigate the potential for the new law to pass.

How could you take a **stratified random sample** in this situation?

Divide the town into subgroups by age, like 18 or below, 19-29, 30-39, 40-49, 50-59, and 60 or above. Using the tax records as the data frame, select a random sample from each of the eight sections. (Answers may vary.)

3. Cluster Sampling: This is another type of random sampling obtained by dividing the population into non-overlapping groups called **clusters** and then randomly selecting an entire cluster to be a part of the sample.
- **Advantage:** Easier, cheaper alternative to simple random sampling and stratified sampling.
 - **Caution:** The best results would happen if the elements within the clusters are *not* alike and in the ideal case, each cluster is representative of the population on a smaller scaler. Otherwise, the sample may not be representative of the population.

Example 1.13: Leash Law (continued)

In a small town, the legislature is interested in passing a new leash law for dog owners. They decide to poll the community to investigate the potential for the new law to pass.

How could you take a **cluster sample** in this situation?

Divide the town into eight non-overlapping sections. Randomly select two of the sections and poll every household in those selected sections. (Answers may vary.)

4. Systematic Sampling: This type of sample is obtained by randomly choosing the i^{th} element from a list of units in the population, then choosing every k^{th} unit afterwards. In other words, we would choose the i^{th} element, then the $(i + k)^{th}$ element, then the $(i + 2k)^{th}$ element, etc. until the desired sample size is chosen.
- **Advantage:** This is a simple method of sampling.
 - **Caution:** Depending on the layout of your data, systematic sampling may not be representative of the population.

Example 1.14: Leash Law (continued)

In a small town, the legislature is interested in passing a new leash law for dog owners. They decide to poll the community to investigate the potential for the new law to pass.

How could you take a **systematic sample** in this situation?

Use the tax records for the town as a data frame. Select a random starting point on the list of taxpayers and include that household in the sample. Then select the household of every tenth taxpayer on the list after that to include in the sample.

5. Convenience Sampling: To obtain this type of sample, we simply choosing subjects for a study that are convenient for the researcher. **We should *never* use convenience sampling!** It may be convenient, but it is not likely going to produce a sample that is representative of the population.

Example 1.15: Leash Law (continued)

How could you take a **convenience sample** in the situation above?

Select the first six customers at a nearby convenience store to poll, or place an ad in the local newspaper asking for volunteers to respond to a poll.

1.4.2 OBSERVATIONAL AND EXPERIMENTAL STUDIES

Statistical studies may involve evaluating one variable about the population of interest or may involve studying the effect of one or more variables called explanatory variable(s) on another called the response variable.

1. Observational studies come about by measuring “what is.” They might consider only one variable or they might observe the response of one variable to changes in another. There is no attempt to manipulate the subjects being studied in observational studies. Instead, we only observe individuals and measure variables of interest. In this type of study, cause and effect between two or more variables cannot be determined, but an association may be present.
2. Experimental studies measure the response of one variable to changes in another variable. However, the researcher attempts to control the environment of the experiment so that the effect of one variable on another can be isolated and measured. The researcher typically randomly assigns the subjects of the study to various levels of the explanatory variable(s) and measures the effect on the response variable.

Example 1.16: Flu Vaccines

Determine whether each of the following describes an **observational study** or an **experiment**.

- A. To determine the effectiveness of this season's flu vaccine, medical researchers recruit 100 participants for a study. At the start of the flu season, the researchers randomly select 50 of the participants to receive the flu vaccine and 50 to not receive it. At the end of flu season, the proportion of participants in each treatment group who contracted the flu is compared.

Designed Experiment – the researcher influenced whether or not the participant received the vaccine.

- B. To determine the effectiveness of this flu season's vaccine at the end of the season, medical researchers compare the proportion of individuals who contracted the flu among a group of 50 randomly chosen individuals who chose to receive the flu vaccine and 50 randomly chosen individuals who chose not to receive the vaccine.

Observational Study – the researcher did not influence whether or not a participant received the vaccine.

- C. What is the **response** variable and what is the **explanatory** variable for the experiment?

Explanatory variable: *Status of receiving the flu vaccine (Yes / No)*

Response variable: *Whether or not patient contracted the flu (Yes / No)*

BASIC PRINCIPLES OF EXPERIMENTAL DESIGN

(1) Control – A comparison of several treatments in the same environment is the simplest form of control. A control group (no treatment) may be utilized.

(2) Replication – There will be natural variability among experimental units, so we must replicate our results (use many subjects).

(3) Randomization – Subjects should be randomly assigned to treatments. Randomization is utilized to control the effects of variables not under study. It also allows cause and effect conclusions to be made.

Example 1.17: Flu Vaccines (continued)

To determine the effectiveness of this season's flu vaccine, medical researchers recruit 100 participants for a study. At the start of the flu season, the researchers randomly select 50 of the participants to receive the flu vaccine and 50 to not receive it. At the end of flu season, the proportion of participants in each treatment group who contracted the flu is compared.

Identify the **three components** of this designed experiment.

Control: There are two treatments – receiving the vaccine and not receiving the vaccine.

Replication: There are 100 participants in the study.

Randomization: Participants are randomly assigned to the two treatments.

In many experiments, it is useful to have a baseline with no treatment. A placebo is a fake treatment that can be used to study this effect. Placebos are used because, when studying humans, some can respond to any treatment – even if it is fake! – simply by knowing they are receiving *some* kind of treatment.

Since this “placebo effect” is prevalent in medical studies, sometimes a double blind study is used to counteract this effect. In this type of study, the subjects and study administrators are not told whether the subjects are a part of the control group or the treatment group.

Example 1.18: Flu Vaccines (continued)

To determine the effectiveness of this season's flu vaccine, medical researchers recruit 100 participants for a study. At the start of the flu season, the researchers randomly select 50 of the participants to receive the flu vaccine and 50 to not receive it. At the end of flu season, the proportion of participants in each treatment group who contracted the flu is compared.

How could we use a **placebo** and a **double blind** study in this experiment?

A placebo could be administered to those in the no vaccine group where the placebo is administered in the same way as the actual vaccine.

We could double blind the study so that neither the person administering the treatment nor the participant knows if the real vaccine or the placebo was being administered.

A confounding variable is one that was not controlled or accounted for by the researcher and thus damages the integrity of the study. It is also sometimes called a “lurking variable.”

Example 1.19: Flu Vaccines (continued)

Consider the flu vaccine observational study.

To determine the effectiveness of this flu season’s vaccine at the end of the season, medical researchers compare the proportion of individuals who contracted the flu among a group of 50 randomly chosen individuals who chose to receive the flu vaccine and 50 randomly chosen individuals who chose not to receive the vaccine.

What are some possible **confounding variables**?

Age of participant, overall health of the participant (answers may vary)

MORE ABOUT EXPERIMENTAL DESIGN AND SAMPLING

The methods of approximation and decision-making which you will learn later in this course are dependent upon a simple random sample being taken from a population and/or treatments being randomly assigned to subjects. Whether a random sample is taken or a random assignment is performed will determine what conclusions can be made about the inference procedure.

In an observational study, a random sample must be taken. If the study is experimental, it is ideal to have both a random sample and random allocation, but valid conclusions may be drawn with only random allocation.

	Random Assignment	No Random Assignment	
Random Sampling	<i>Causation generalized to the entire population</i>	<i>Association generalized to entire population</i>	<i>Inference generalized to entire population</i>
No Random Sampling	<i>Causation generalized only to those similar to those in the study</i>	<i>Association generalized only to those in the study</i>	<i>May not make inference to the entire population</i>
	<i>Causation</i>	<i>Association</i>	

MY NOTES:

PRACTICE EXERCISES FOR SECTION 1

1. Americans' perceptions of the fairness of their taxes have varied over the years, tending to be lower in years when a Democratic president was in office and higher in years when there was a Republican president. In April 2017, the Gallup News Service conducted a survey of 1,019 American adults aged 18 years or older. The results of the survey indicated that 61% of American adults believe that their federal income tax is fair. *Source: gallup.com*

- (a) What was the objective of this survey?

The objective is to approximate the proportion of American adults aged 18 or older who believe their federal income tax is fair.

- (b) Describe the population of interest.

American adults aged 18 or older

- (c) Describe the sample.

1,109 American adults aged 18 or older

- (d) Is 61% a parameter or a statistic?

Statistic – it was calculated using the sample.

- (e) Was the goal of the Gallup organization to perform descriptive or inferential statistics? Explain your reasoning.

Inferential statistics – the goal was to approximate a parameter.

2. A manufacturer of computer chips claims that less than 9% of its products are defective. When quality control inspectors draw 2,000 chips from a large production, they find 7.75% to be defective.

(a) Describe the population of interest.

All computer chips manufactured by this company.

(b) Identify the parameter of interest.

The true percentage of chips that are defective. (Note that the manufacturer claims it is 9%, but we do not know what the true percentage is!)

(c) Identify the statistic from the sample in this scenario.

The 7.75% of the 2,000 chips sampled that were defective.

(d) What part of the study involves descriptive statistics?

Summarizing the sample statistic from the 2000 chips.

(e) What part of the study involves inferential statistics?

Using the information from the sample to make a generalization to the population of all chips.

(f) Is the ultimate goal of this scenario to perform inferential or descriptive statistics? Explain your reasoning.

Inferential statistics – the goal was to approximate the true population proportion of defective chips, not just to describe the proportion of defective chips in the sample.

3. A personnel director is interested in determining how effective a new mathematics course will be in improving the math computation skills of her company's employees. The director randomly selected twenty employees and determined the average math computation score both before and after instruction in the course.

(a) Identify the population of interest.

All company employees

(b) What characteristic of the population is being measured?

Average math computation score before and after the course

(c) Is this an example of inferential or descriptive statistics?

Inferential statistics – the goal is to approximate the scores for all company employees from the sample of twenty employees.

4. A statistics professor wants to know the average score on the Fall 2020 Final Exam for the forty-five students who took her class.

(a) Are the students in the professor's class the population or the sample in this scenario? How do you know?

Population – all the individuals of interest are included in the class of students.

(b) What characteristic of the students is being measured?

The average Fall 2020 final exam score.

(c) Is this an example of inferential or descriptive statistics? Explain your reasoning.

Descriptive statistics – the professor is only interested in the students in her class and are not interested in generalizing to a larger population.

5. Identify the following variables as qualitative/quantitative, discrete/continuous/neither, and nominal/ordinal/interval/ratio.

- (a) A researcher records whether a sample of Clemson students attend the football games or not.

Qualitative, Neither, Nominal

- (b) A bar owner records the number of margaritas a sample of patrons orders per day.

Quantitative, Discrete, Ratio

- (c) A medical student records the body temperature of a random sample of patients.

Quantitative, Continuous, Interval

- (d) A dorm RA records whether each of his residents drinks or drank while underage.

Qualitative, Neither, Nominal

- (e) A housing officer at a university records a student's class status (freshman, sophomore, junior, or senior).

Qualitative, Neither, Ordinal

- (f) An instructor records the responses to a satisfaction evaluation (Very Satisfied, Satisfied, Neutral, Slightly Unsatisfied, Very Unsatisfied).

Qualitative, Neither, Ordinal

- (g) A student records the amount of sleep per night she gets during a semester.

Quantitative, Continuous, Ratio

6. Because of concerns about employee stress, a large company is conducting a study to compare two programs (tai chi or yoga) that may help employees reduce their stress levels. Tai chi is a 1200-year-old practice, originating in China, that consists of slow, fluid movements. Yoga is a practice, originating in India, that consists of breathing exercises and movements designed to stretch and relax muscles. The company has assembled a group of volunteer employees to participate in the study during the first half of their lunch hour each day for a 10-week period. Each volunteer will be assigned at random to one of the two programs. Volunteers will have their stress levels measured just before the beginning of the program and ten weeks later at the completion of it.

- (a) Identify the three components of experimental design in this study.

Control: There are two treatments being compared.

Replication: More than just a few employees will be involved in the study.

Randomization: Employees are randomly assigned to either the yoga or tai chi group.

- (b) Someone proposes that a control group be included in the design as well. The stress level would be measured for each volunteer assigned to the control group at the start of the study and again ten weeks later. What additional information, if any, would this provide about the effectiveness of the program?

A control group that does not participate in the yoga or tai chi program will provide a baseline for comparison.

- (c) A group of volunteers who work together ask to be assigned to the same program so that they might participate in the program together. Describe a few confounding variables that could affect the results of the study if employees are allowed to choose their treatment group along with their friends.

Volunteers who work together may be under the same amount of stress due to an upcoming deadline or other work factors. Friends may also have similar interests that affect stress levels, such as tending to exercise or socialize to reduce stress.

7. The dentists in a dental clinic would like to determine if there is a difference between the number of new cavities in people who eat an apple a day and in people who eat less than one apple a week. They are going to conduct a study with 50 people in each group. Fifty clinic patients who report that they routinely eat an apple a day and 50 who report that they eat less than one apple a week will be identified. The dentists will examine the patients and their records to determine the number of new cavities the patients have had over the past two years. They will then compare the number of new cavities in the two groups.

(a) Is this an observational study or a designed experiment? Explain.

Observational study – no treatment is imposed upon the subjects.

(b) List a potential confounding variable.

More health-conscious individuals may be more prone to eat apples and have healthier teeth because they take better care of themselves.

8. London scientists conducted a study to determine if chocolate can trigger migraines. Twelve migraine-prone subjects were given a peppermint-laced chocolate candy and eight migraine-prone subjects were given a peppermint-laced placebo made of carob, peppermint, and vegetable fat. Five subjects from the group given chocolate developed a migraine within one day. No one from the group given the placebo developed a migraine in that time period.

(a) Identify the variables in this study (explanatory, response, confounding).

Explanatory: Consumption of chocolate versus placebo

Response: Number of migraine headaches within one day

Confounding: Peppermint

(b) Is this a designed experiment or an observational study?

Designed experiment

9. Which of the following sampling methods produce a random sample from a class of 36 students? If the sample is random, state what type of sampling was used.

- (a) Select the first six students to enter a room.

Convenience sampling – not random, may not be representative of the class

- (b) Suppose that the class has 18 boys and 18 girls. Select a sample of six students by using a random number table to choose one of the 18 boys, then one of the 18 girls, then a boy, then a girl, and so on, until you have chosen six students.

Stratified sampling – random, population divided by gender

- (c) Assign each student a number from 1 to 36. The girls get the numbers 1–18 and the boys get the numbers 19–36. Choose six numbers at random.

Simple random sampling

- (d) Suppose that the classroom has six rows of chairs with six chairs in each row. Assign the rows the digits 1 through 6. Throw a die and place all the students in the row corresponding to the number on the die in the sample.

Cluster sampling – each row is a cluster

SECTION 2: DISPLAYING AND SUMMARIZING DATA

(HAWKES 3.1 – 3.5, 4.1 – 4.3, 4.6)

In Section 1, we learned what data are, as well as how to classify and collect data. To better communicate to others the information contained in our data, we need clear and concise ways to present and describe data. In other words, we need to turn our data into useful information. We will refer to the presentation of data as a [data distribution](#). A distribution gives the values that a variable can take and how often it takes on those values.

Graphical displays and **numerical summaries** are good ways to achieve this goal. Graphical displays can be thought of as “pictures” that allow us to visualize the data that comes from a population or sample. A numerical summary takes our data and distills it down into one number representing a specific characteristic of the data.

2.1 QUALITATIVE DATA - GRAPHICAL DISPLAYS

Previously we considered variables (or characteristics) about our population or sample. We divided our variables into two broad categories: **qualitative (categorical) data** and **quantitative (numerical) data**. How we display and summarize our data will depend on the type of data with which we are working.

We will focus on qualitative data first.

Graphical displays for qualitative data include:

- Frequency / Relative Frequency Distributions
- Pie Charts
- Bar Charts
- Stacked Bar Charts

Instructions for generating all of the graphical displays discussed in this section using JMP can be found in the file **JMP Instructions for STAT 3090**.

2.1.1 FREQUENCY DISTRIBUTIONS

Perhaps the simplest way to present a distribution of qualitative data is by using a table. A [frequency distribution](#) summarizes data in a table by *category* (class) and the *number of observations* (frequency) in each category.

Example 2.1: Olympic Medal Winners

This table shows the gold medal winners in hockey in the Winter Olympics since 1920. Use the data to create a **frequency distribution**.

Year	Winner		Year	Winner
1920	Canada		1980	U.S.A.
1924	Canada		1984	Soviet Union
1928	Canada		1988	Soviet Union
1932	Canada		1992	Unified Team
1936	Great Britain		1994	Sweden
1948	Canada		1998	Czech Republic
1952	Canada		2002	Canada
1956	Soviet Union		2006	Sweden
1960	U.S.A.		2010	Canada
1964	Soviet Union		2014	Canada
1968	Soviet Union		2018	Russia
1972	Soviet Union		2022	Finland
1976	Soviet Union			

WINNER	FREQUENCY
Canada	9
Soviet Union	7
U.S.A.	2
Great Britain	1
Sweden	2
Czech Republic	1
Unified Team	1
Russia	1
Finland	1
Total	25

2.1.2 RELATIVE FREQUENCY DISTRIBUTIONS

Like a frequency distribution, a [relative frequency distribution](#) also summarizes data by category, but it uses the [relative frequency](#) of an event or observation, which is the proportion of times an event or observation occurs out of the total number.

Example 2.2: Olympic Medal Winners (continued)

Revisit the data from Example 2.1. Create a **relative frequency distribution**.

WINNER	RELATIVE FREQUENCY
Canada	9/25
Soviet Union	7/25
U.S.A.	2/25
Great Britain	1/25
Sweden	2/25
Czech Republic	1/25
Unified Team	1/25
Russia	1/25
Finland	1/25
Total	25/25

What proportion of gold medals were won by Canada?

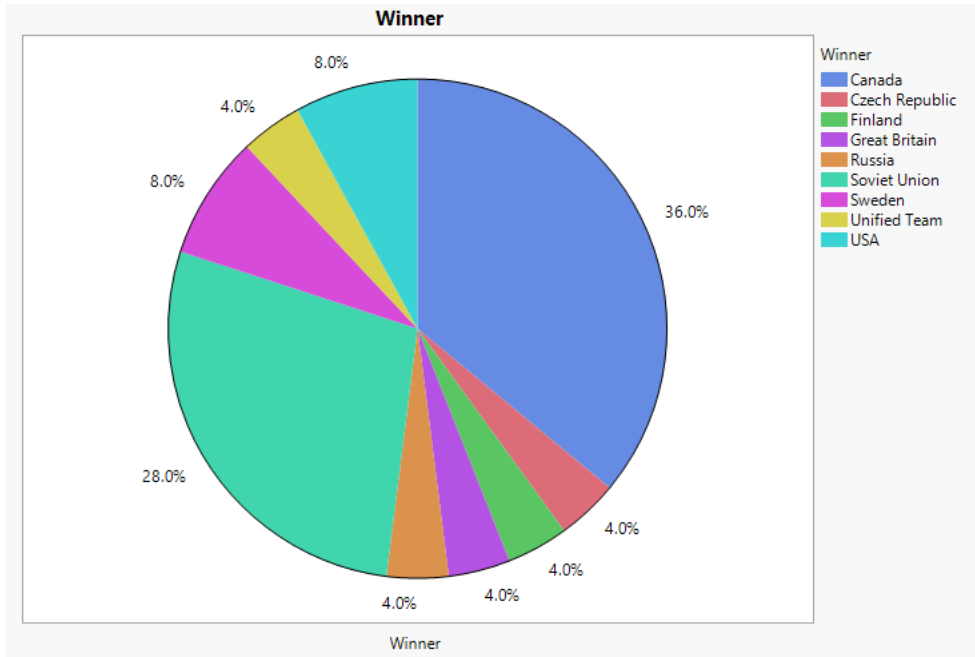
9/25=36% of gold medals were won by Canada.

With large datasets, it is useful to be able to create frequency and relative frequency distributions using technology. Instructions for creating these distributions in JMP can be found in the file **JMP Instructions for STAT 3090**.

We can use the information gleaned from a frequency or relative frequency distribution to create other graphical displays of our data.

2.1.3 PIE CHARTS

A [*pie chart*](#) provides the relative frequencies of each category represented by the proportion of a circle that corresponds to the relative frequency.



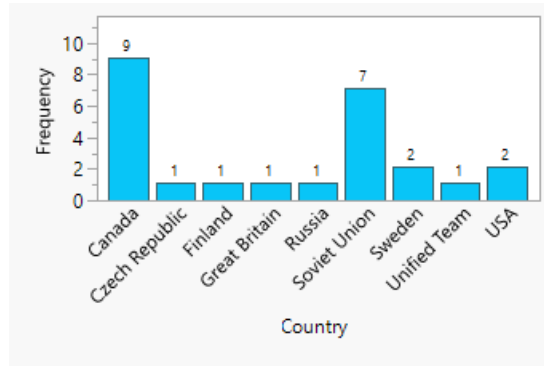
Instructions for creating pie charts in JMP can be found in the **JMP Instructions for STAT 3090**.

2.1.4 BAR CHARTS

A [bar chart](#) displays the categories using the height of a bar to represent the frequency or relative frequency of each category.

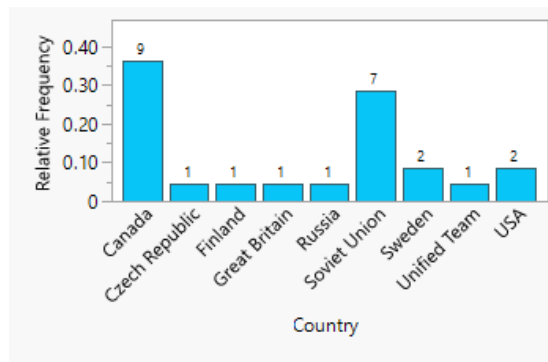
Bar Chart with Frequencies

MEN'S OLYMPIC HOCKEY WINNERS



Bar Chart with Relative Frequencies

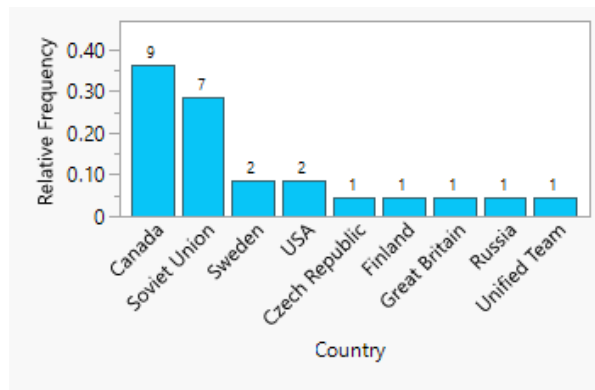
MEN'S OLYMPIC HOCKEY WINNERS



Note that the axis labels are important to distinguish between types of bar charts. When creating a bar chart, you should always include axis labels, a graph title, and units when appropriate.

It is often useful to order categories for your bar chart by their (relative) frequency. A [Pareto chart](#) is a bar chart that orders the bars from highest to lowest (relative) frequency.

OLYMPIC HOCKEY WINNERS

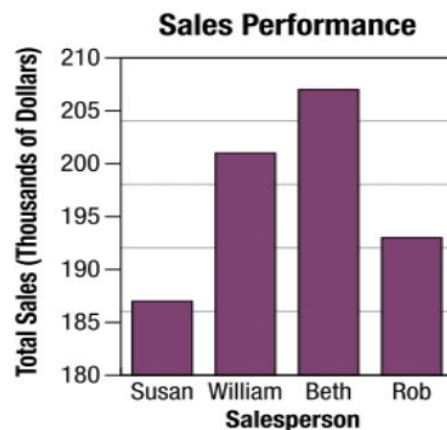
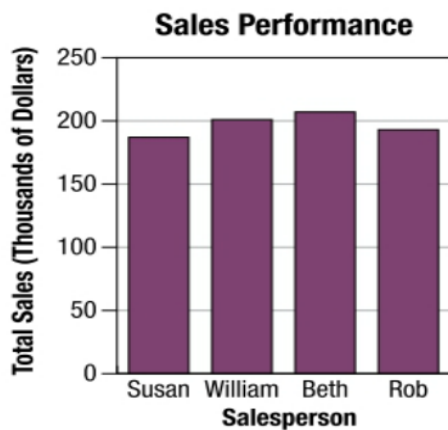


Example 2.3: Using Accurate Bar Charts

Consider the sales performance of the following salespersons.¹ Then examine the two bar charts below that were constructed for this data.

What is the difference between the graphs?

SALESPERSON	TOTAL SALES (THOUSAND DOLLARS)
Susan	187
William	201
Beth	207
Rob	193



The vertical axis of the second bar chart does not begin at zero, which can make the perceived differences in sales to be greater than they actually are.

¹ Example 2.3 taken from *Discovering Business Statistics* by Nottingham and Hawkes.

2.2 QUALITATIVE DATA - NUMERICAL SUMMARIES

So far, we have learned how to classify qualitative data and use tables or graphs to present a visual representation of them. We can further turn our data into useful information by using [numerical summaries](#). Recall that a numerical summary takes our data and distills it down into one number representing a specific characteristic of the data. It provides us with a “measurement” of our data.

Numerical summaries for qualitative data include:

- Proportions
- Odds (not covered in STAT 3090)
- Odds Ratio (not covered in STAT 3090)

Numerical summaries can be broken down into two types:

- Numerical summaries or measures that apply to *populations* are called [parameters](#).
- Numerical summaries or measures that apply to *sample data* are called [statistics](#).

2.2.1 PROPORTIONS

The population proportion, denoted by a lowercase p is the fraction of a population which possesses some characteristic of interest.

Formula 2.1: Population Proportion

For a population of size N , the **population proportion** for a characteristic of interest is:

$$p = \frac{x}{N}$$

where x is the number of observations from the population with the desired characteristic.

Since the population is usually too large to attempt to collect all of the data needed to calculate numerical summaries, we instead estimate those values using sample data. The sample proportion $\hat{p}$ is the fraction of the sample that possesses some characteristic of interest.

Formula 2.2: Sample Proportion

For a sample of size n , the **sample proportion** for a characteristic of interest is:

$$\hat{p} = \frac{x}{n}$$

where x is the number of observations in the sample that possess the characteristic.

Example 2.4: Coffee Preferences

Suppose a researcher is interested in whether Starbucks or Dunkin Donuts is preferred among all college students in the United States. A random sample of 5,830 students from universities across the country revealed that 3,673 of the students preferred Starbucks over Dunkin Donuts. What is the **proportion** of students from the sample who prefer Starbucks over Dunkin Donuts? Use appropriate notation.

$$\hat{p} = \frac{x}{n} = \frac{3673}{5830} = 0.63$$

MY NOTES:

2.3 QUANTITATIVE DATA - GRAPHICAL DISPLAYS

We have learned graphical displays and numerical summaries for qualitative data, now let's move to our second type of data - quantitative data.

Graphical displays for quantitative data include:

- Frequency / Relative Frequency / Cumulative Relative Frequency Distributions
- Histograms
- Stem and Leaf Displays
- Dot Plots
- Time Series Data (not covered in STAT 3090)

2.3.1 FREQUENCY / RELATIVE FREQUENCY / CUMULATIVE RELATIVE FREQUENCY DISTRIBUTIONS

When we use a frequency or relative frequency table to display quantitative data, we need to divide our data into categories known as bins. With qualitative data, selecting the categories for displaying the data in a table is relatively easy. However, when the data are quantitative, this is not particularly obvious. One should consider the range of values in the dataset along with the number of [*categories \(bins\)*](#) desired when constructing a frequency or relative frequency distribution. Be sure that the number of categories and bin width appropriately represent the dataset. The bin width must be [*the same*](#) for all of the data.

When quantitative data is divided into bins, we need to make sure each data point can only belong to one bin. One way to denote the endpoints of a bin for quantitative data is to use the square bracket “[” to indicate that the value closest to it is included in the bin and to use a round bracket “)” to indicate that the value closest to it is not included in the bin.

Example 2.5: Poverty Level by State

The following table gives the percentage of people living below the poverty level for each state and the District of Columbia in the year 2015 (ordered sequentially for convenience).

8.2	9.7	10.2	10.2	10.3	10.5	10.6	10.8	11.0
11.1	11.2	11.3	11.5	11.5	12.1	12.2	12.2	12.4
12.6	13.0	13.2	13.4	13.6	13.7	13.9	14.5	14.6
14.7	14.8	14.8	15.1	15.3	15.4	15.4	15.7	
15.8	15.9	16.1	16.4	16.6	16.7	17.0	17.3	
17.4	17.9	18.5	18.5	19.1	19.6	20.4	22.0	

The data ranges from roughly 8% to roughly 22%. You may choose any bin width that reasonably divides your data. A bin width of 2% seems reasonable here.

We may summarize the data using a frequency distribution, relative frequency distribution, and cumulative relative frequency distribution.

Percent Living Below the Poverty Level	Frequency (Number of States)	Relative Frequency	Cumulative Relative Frequency
[8, 10)	2	$\frac{2}{51} \approx 0.039$	$\frac{2}{51}$
[10, 12)	12	$\frac{12}{51} \approx 0.235$	$\frac{2 + 12}{51} = \frac{14}{51}$
[12, 14)	11	$\frac{11}{51} \approx 0.216$	$\frac{14 + 11}{51} = \frac{25}{51}$
[14, 16)	12	$\frac{12}{51} \approx 0.235$	$\frac{25 + 12}{51} = \frac{37}{51}$
[16, 18)	8	$\frac{8}{51} \approx 0.157$	$\frac{37 + 8}{51} = \frac{45}{51}$
[18, 20)	4	$\frac{4}{51} \approx 0.078$	$\frac{45 + 4}{51} = \frac{49}{51}$
[20, 22)	1	$\frac{1}{51} \approx 0.020$	$\frac{49 + 1}{51} = \frac{50}{51}$
[22, 24]	1	$\frac{1}{51} \approx 0.020$	$\frac{50 + 1}{51} = \frac{51}{51}$
Total	51	1	

Note, the *cumulative relative frequency* calculates how much relative frequency has been accumulated until the current bin. In other words, it is the *sum of relative frequencies* from the current and all previous bins.

Example 2.6: Using the Cumulative Relative Frequency Distribution

Use the cumulative relative frequency distribution you created in Example 2.5 to answer the following questions.

- A. What proportion of states have poverty rates less than 16%?

$$\frac{37}{51} \approx 0.725$$

- B. What proportion of states have poverty rates of at least (greater than or equal to) 14%?

$$1 - \frac{25}{51} = \frac{26}{51} \approx 0.510$$

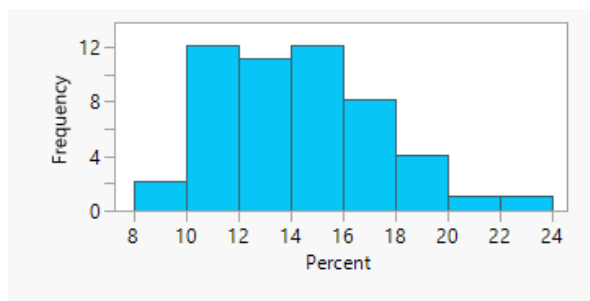
2.3.2 HISTOGRAMS

The frequency table and relative frequency table that you created in the previous example may be displayed graphically using a histogram.

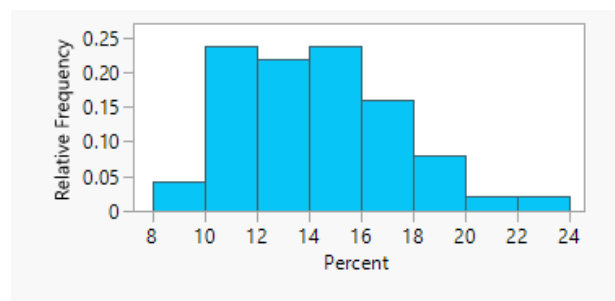
A **histogram** is a graph that uses a bar to represent the bins of the data. The height of the bars represents the frequencies or relative frequencies of the values of the variable. It is a useful graphical tool for large quantitative datasets. It is important to notice that with a histogram you know the frequency or relative frequency of a particular bin of data, but you do not know the individual data values.

PERCENT BELOW POVERTY LEVEL

Histogram Using Frequency



Histogram Using Relative Frequency



2.3.3 STEM AND LEAF DIAGRAMS

Another useful tool for displaying quantitative data is a **stem and leaf diagram**. Stem and leaf plots or diagrams are appropriate for relatively small discrete datasets. The stem and leaf diagram is similar to a histogram which can divide our data into bins (called stems here); however, unlike a histogram, values of the original dataset are apparent in the display. A stem and leaf diagram can be constructed using the following steps:

Step 1: Sort the data from low to high.

Step 2: Split the values into the stem and the leaf.

- For example, for the value 117, the stem is 11 (all digits left of the units place) and the leaf is 7 (the units place).
- For discrete decimal values, you could also designate the stem as all digits left of the decimal place and the leaf as the tenths place. For example, for the value 8.2, the stem is 8 and the leaf is 2.

Step 3: List the stems from lowest to highest. There should be no gaps in stems.

Step 4: Order the leaves from lowest to highest and place next to each stem. There could be some gaps in leaves if there's no data in the range of some stems.

Step 5: Include a key to illustrate how to read the data.

Example 2.7: Poverty Level by State (continued)

Consider the data presented in Example 2.5, which are already sorted least to greatest for convenience. Use the data to create a **stem and leaf diagram**. Include a title and key.

```

08 | 2
09 | 7
10 | 2 2 3 5 6 8
11 | 0 1 2 3 5 5
12 | 1 2 2 4 6
13 | 0 2 4 6 7 9
14 | 5 6 7 8 8
15 | 1 3 4 4 7 8 9
16 | 1 4 6 7
17 | 0 3 4 9
18 | 5 5
19 | 1 6
20 | 4
21 |
22 | 0
  
```

PERCENTAGE OF POPULATION LIVING BELOW POVERTY

Key: 8 | 2 = 8.2%

8.2	9.7	10.2	10.2	10.3	10.5	10.6	10.8	11.0
11.1	11.2	11.3	11.5	11.5	12.1	12.2	12.2	12.4
12.6	13.0	13.2	13.4	13.6	13.7	13.9	14.5	14.6
14.7	14.8	14.8	15.1	15.3	15.4	15.4	15.7	
15.8	15.9	16.1	16.4	16.6	16.7	17.0	17.3	
17.4	17.9	18.5	18.5	19.1	19.6	20.4	22.0	

Note that in JMP, the order of the stems is reversed. An example of stem and leaf diagram generated by JMP for the data in Example 2.5 is:

Stem	Leaf	Count
22	0	1
21		
20	4	1
19	16	2
18	55	2
17	0349	4
16	1467	4
15	1344789	7
14	56788	5
13	024679	6
12	12246	5
11	012355	6
10	223568	6
9	7	1
8	2	1

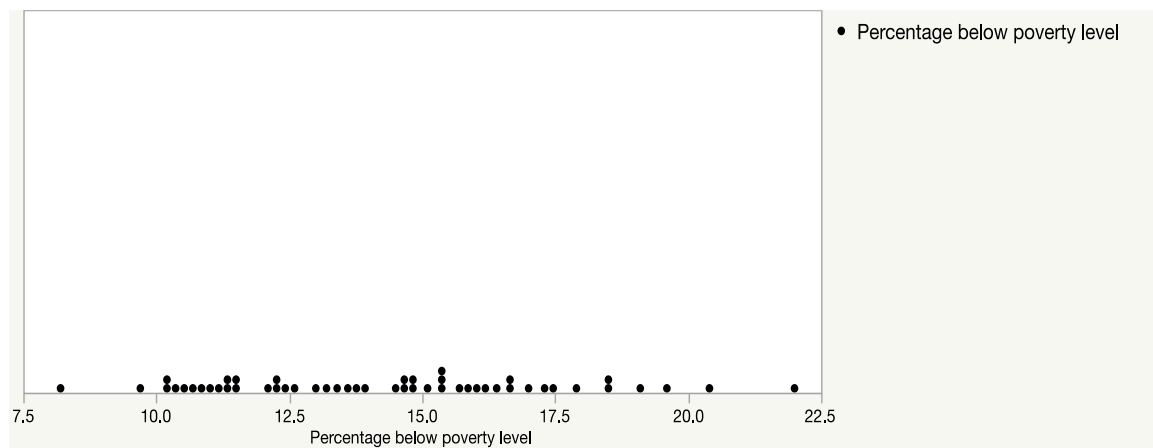
8|2 represents 8.2

Instructions for creating a stem and leaf diagram can be found in the **JMP Instructions for STAT 3090**.

2.3.4 DOT PLOTS

A **dot plot** shows a dot for each observation, placed above each value on the number line for that observation. Like the stem and leaf plot, it portrays each observation; in other words, you could recreate the dataset solely on the information contained in the plot.

The following shows a dot plot for a distribution of the poverty data from Example 2.5.



When you make a graphical display of quantitative data, you should look at the graph and take note of the shape of the distribution of the data. Is it approximately symmetric, i.e., is the left side a mirror image of the right side? Does the frequency or relative frequency of the data seem to taper off to the right or left? Are there major gaps where there is no data? We will look more closely at the shapes of distributions in Section 3.

2.4 QUANTITATIVE DATA - NUMERICAL SUMMARIES

Similar to qualitative data, we can further turn quantitative data into meaningful information by using numerical summaries or measures.

Numerical summaries for quantitative data include:

- Measures for central tendency: mean, median, mode
- Measures for spread: range, variance, standard deviation, interquartile range

Each of these measures can be applied to populations and samples.

Note, the way in which we calculate the measures for a population may be slightly different than the way in which we calculate the corresponding measure for a sample.

2.4.1 MEASURES OF CENTRAL TENDENCY

There are three numerical summaries that provide a measure for the central or typical value of a quantitative variable: mean median and mode. The mean and the median are the most commonly used measures of center. The mode is also a measure of center, but it is not used as often.

Mean, median, and mode all have the same units as the observations in the population or sample to which they refer.

Mean

The mean (or average) is the *sum* of the observations divided by the *number* of observations. It can be thought of as a balancing point of the data.

SYMBOLS FOR MEAN		
μ	Population Mean	“mu”
$\bar{x}$	Sample Mean	“x-bar”

Formula 2.3: Sample Mean

For a sample of n observations on a variable x , the **mean** is denoted by $\bar{x}$ and has the formula:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

where x_i is the value of the i^{th} observation of x and the Greek letter Σ (capital sigma) indicates the operation of adding up all the possible values of x . In this example we have used the variable x to represent the data values. You can use any letter or phrase appropriate for the problem. It is typical to place a bar above the letter or phrase used as the name of your variable to indicate a sample mean.

Formula 2.4: Population Mean

For a population of N observations on a variable x , the **mean** is denoted by μ or μ_x and has the formula:

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

where x_i is the value of the i^{th} observation of x . It is common to use a capital letter N as the notation to indicate the total number of observations in the population and to use a lower case n to indicate the total number of observations in a sample.

Median

The median is the middle observation when the observations are listed in ascending order. If n is the number of observations in a dataset, then...

- the median is the middle value if n is odd,
- the median is the average of the middle two values if n is even,
- the median of a population and a sample is found by the same method.

Mode

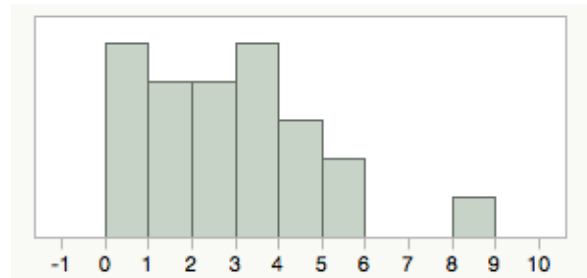
The mode is the observation that occurs the most often in either a sample or a population.

- There can be more than one mode in a dataset.
- There can also be no mode if all data values occur only once.
- Mode is not used as often as mean and median to describe the central or typical value in a data set.

Example 2.8: Number of Defects

The following data represents the number of defects per shipment of a random sample of shipments of computer parts produced by a technology company. Let x represent the number of defects per shipment.

3	0	2	0	1	3	5	2
5	1	3	0	0	1	3	2
4	3	1	8	4	2	4	0



- A. Find the **mean** number of defects per shipment.

$$\bar{x} = \frac{3 + 0 + 2 + \dots + 2 + 4 + 0}{24} = \frac{57}{24} = 2.375 \text{ defects}$$

- B. Find the **median** number of defects per shipment.

Sort the data: 0 0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 3 3 4 4 4 5 5 8

$$\text{Median} = \frac{2 + 2}{2} = 2$$

Since there are 24 values in the dataset, we find the mean of the 12th and 13th values (the middle two values) to find the median.

- C. Find the **mode** of the number of defects per shipment.

The most frequently occurring values are 0 and 3. Therefore, the modes for this data set are 0 and 3 defects.

The data from Example 2.8 may be summarized the frequency distribution to the right. We may find the mean of these summarized data by using the weighted mean. This measure of center uses the same symbol as the ordinary sample or population mean; it is simply another way to calculate the same value given data in a different form.

Number of Defects (v_j)	Frequency (w_j)
0	5
1	4
2	4
3	5
4	3
5	2
8	1

Formula 2.5: Weighted Sample Mean – Frequency Table

For a sample of n observations on a variable x , where there are k unique values ($1 \leq k \leq n$) the **weighted sample mean** is denoted by $\bar{x}$ and has the formula:

$$\bar{x} = \frac{v_1 w_1 + v_2 w_2 + \cdots + v_k w_k}{w_1 + w_2 + \cdots + w_k} = \frac{\sum_{j=1}^k v_j w_j}{\sum_{j=1}^k w_j} = \frac{\sum_{j=1}^k v_j w_j}{n}$$

where v_j is the j^{th} unique value of the observations of x and w_j is the corresponding weight or frequency.

Example 2.9: Number of Defects (continued)

- A. Verify that the formula for weighted mean results in the same mean found in Example 2.8.

$$\bar{x} = \frac{(0 \times 5) + (1 \times 4) + (2 \times 4) + (3 \times 5) + (4 \times 3) + (5 \times 2) + (8 \times 1)}{5 + 4 + 4 + 5 + 3 + 2 + 1} = \frac{57}{24} = 2.375 \text{ defects}$$

- B. How would one use the frequency table to find the median?

Count the total frequency (24 observations), then find the row that contains the 12th and 13th values and take the average. The median is 2.

Example 2.10: Number of Defects (continued)

The table to the right is the **relative frequency** distribution for the previous example (verify that each relative frequency is the *frequency* divided by the *total frequency*).

How would the calculation for the mean change for this relative frequency distribution?

Number of Defects	Relative Frequency
0	0.2083
1	0.1667
2	0.1667
3	0.2083
4	0.125
5	0.0833
8	0.04167

The calculation is similar, but rather than dividing by the total frequency at the end, each relative frequency is already divided by the total number of observations, so only multiplication and addition are necessary.

$$\bar{x} = 0(0.2083) + 1(0.1667) + 2(0.1667) + 3(0.2083) + 4(0.125) + 5(0.0833) + 6(0.04167) = 2.375 \text{ defects}$$

This idea gives us a general formula to calculate the weighted sample mean from a relative frequency table.

Formula 2.6: Weighted Sample Mean – Relative Frequency Table

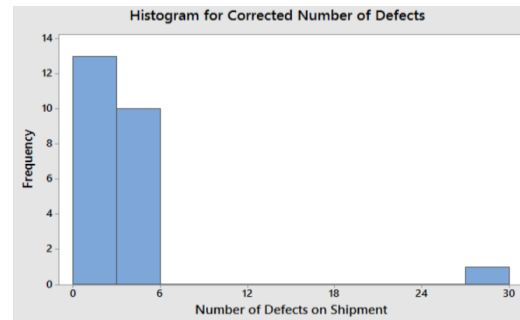
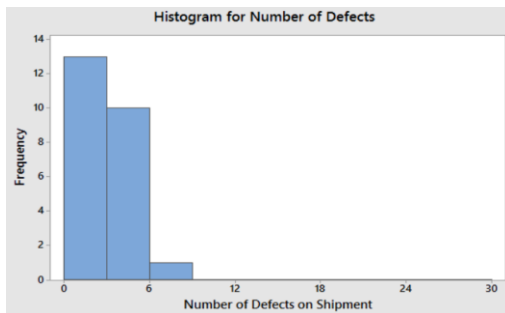
For a sample of n observations on a variable x , where there are k unique values ($1 \leq k \leq n$) the **weighted sample mean** is denoted by $\bar{x}$ and has the formula:

$$\bar{x} = \sum_{j=1}^k v_j w_j$$

where v_j is the j^{th} unique value of the observations of x and w_j is the corresponding **relative frequency** weight.

Example 2.11: Number of Defects (continued)

Suppose the data for the number of defects in the shipment were incorrect and that there were 28 defects instead of 8 defects (the dataset is the same except for the last observation). Which measure of center (mean or median) will change? Which measure is a better measure of the location of the number of defects in a *typical* shipment in this situation?



Number of Defects	Frequency
0	5
1	4
2	4
3	5
4	3
5	2
28	1

Sample Mean: $\bar{x} = 3.21$

This is about a 35% increase in the mean when only 6 values in the dataset are larger than the new mean.

The median stays the same: Median = 2

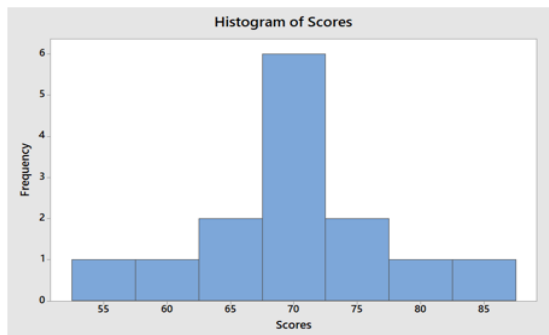
*The **median** best describes the typical number of defects since it ignores such mistake.*

2.4.2 MEASURES OF SPREAD

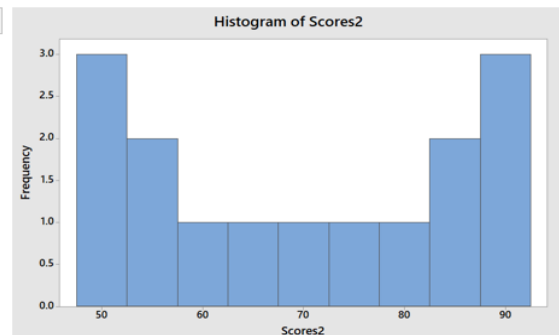
The measure of center is not enough to adequately describe a distribution because it does not describe the dispersion of the data, which is referred to as **variability** or **spread** of the distribution.

Example 2.12: Variability of a Dataset

The following distributions of test scores have the same mean score of 70; however, the spreads of the distributions are different. Which distribution has larger overall variability?



2



The second distribution has larger overall variability.

We will primarily use four different measures to describe the variability in a dataset: range, variance, standard deviation, and interquartile range. The interquartile range will be introduced later in this section.

Range

The range is the difference between the largest and smallest values in your dataset. It always has the same units as the original dataset (ounces, dollars, meters, etc.).

Formula 2.7: Range

$$\text{Range} = \text{Maximum} - \text{Minimum}$$

Variance

The variance is a numerical measurement of the average squared deviation, or average squared distance, from the mean.

SYMBOLS FOR VARIANCE		
σ^2	Population Variance	“sigma”
s^2	Sample Variance	“s-squared”

Formula 2.8: Sample Variance

For a sample of n observations on a variable x with mean $\bar{x}$, the **sample variance** has the formula:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

where x_i is the value of the i^{th} observation of x .

The units for variance are squared. For example, if the variable x represented dollars, the units for variance would be “dollars squared.”

Because it is rare to have data from an entire population, we will rarely calculate a population variance and standard deviation. Their formulas are included below for reference.

Formula 2.9: Population Variance

For a population with N observations on variable x with mean μ , the **population variance** has the formula:

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

Pay attention to the denominators when calculating variance from sample or population. The sample variance uses $n - 1$ while the population variance uses N . This difference is relevant to the degrees of freedom or unbiased estimator. The detailed explanation won't be covered in STAT 3090. Feel free to explore it on your own.

Standard Deviation

The standard deviation is the measure of the average deviation or distance from the mean. Its units are the same as the original dataset (for example, ounces, dollars).

Formula 2.10: Sample Standard Deviation

For a sample of n observations on variable x , the **sample standard deviation** has the formula:

$$s = \sqrt{s^2}$$

where s^2 is the sample variance as calculated in Formula 2.8.

Formula 2.11: Population Standard Deviation

For a population with N observations on variable x , the **population standard deviation** has the formula:

$$\sigma = \sqrt{\sigma^2}$$

where σ^2 is the population variance as calculated in Formula 2.9.

Example 2.13: Potato Chips

Six snack size bags of Lay's potato chips contain the following number of chips per bag:

11 13 13 15 18 20

A. Find the **mean** of the sample.

$$\text{Mean: } \bar{x} = \frac{11 + 13 + 13 + 15 + 18 + 20}{6} = 15$$

B. Find **standard deviation** of the sample.

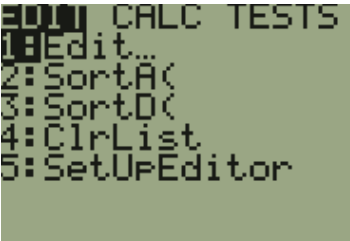
x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
11	$11 - 15 = -4$	16
13	$13 - 15 = -2$	4
13	$13 - 15 = -2$	4
15	$15 - 15 = 0$	0
18	$18 - 15 = 3$	9
20	$20 - 15 = 5$	25

$$s^2 = \frac{16 + 4 + 4 + 0 + 9 + 25}{6 - 1} = 11.6 \text{ chips}^2, \quad s = \sqrt{11.6} \approx 3.41 \text{ chips}$$

Calculator Tutorial 2.1: Mean and Standard Deviation

You can find the **sample mean** and **sample standard deviation** using a TI-84 calculator. This can be a great way to check your calculations and verify that your work is correct.

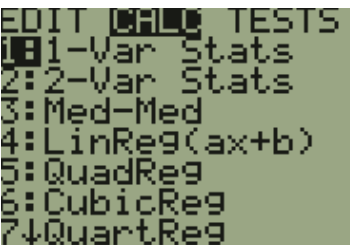
- (1) Click the STAT button on your calculator. The first item 1: EDIT... on the EDIT menu will already be selected. Click ENTER.



- (2) Enter your data into L1. Click 2ND + MODE to exit the list.



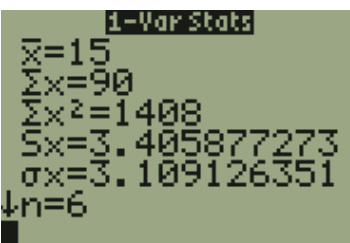
- (3) Click the STAT button again. Click the right arrow key once to navigate to the CALC menu. The first item 1: 1-VAR STATS will already be selected. Click ENTER.



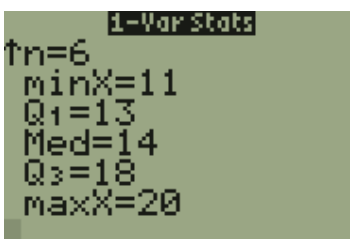
- (4) Make sure that L1 appears in the LIST option. (If it does not, click 2ND + 1 to enter L1.) The space next to FREQLIST should be blank. Use the down arrow key to Select CALCULATE and click ENTER.



- (5) This will bring up descriptive statistics for the sample you entered in L1. The **sample mean** is denoted $\bar{x}$. The **sample standard deviation** is denoted S_x .



The calculator labels the value of the sample standard deviation as S_x and labels the value of the population standard deviation as σ_x . You can use the down arrow key to navigate to additional descriptive statistics.



MY NOTES:

2.5 APPLICATIONS OF QUANTITATIVE NUMERICAL SUMMARIES

So far, we have learned several numerical summaries for quantitative data – mean, median, mode, range, variance, and standard deviation. Can we use these numerical summaries to find out more information from the data? The answer is yes.

2.5.1 Z-SCORE

A z-score is a standardized measure of relative position, with respect to the mean and variability (as measured by the standard deviation) of a dataset.

Formula 2.12: Z-Score Formula

For a given observation x from a population with mean μ and standard deviation σ , the **z-score** for the observation is:

$$z = \frac{\text{observation} - \text{mean}}{\text{standard deviation}} = \frac{x - \mu}{\sigma}$$

We can see from the formula, the z-score tells us how many standard deviations a data point is below or above the mean, therefore, it describes where a particular data point falls in relation to the entire dataset, i.e., the **relative position** of a data point in the entire dataset.

Note, for an observation with a **negative** z-score, we say that the observation falls z standard deviations *below* the mean; for an observation with a **positive** z-score, we say that the observation falls z standard deviations *above* the mean.

Example 2.14: Test Scores

A statistics student scored 75 on the first exam of the semester and 82 on the second exam. The mean and standard deviation of scores for the class are given in the table on the right. On which exam did the student perform better relative to the rest of the class? Compare the student's relative standing using **z-scores**.

	Mean	Standard Deviation
Exam 1	74	10
Exam 2	85	7

$$z_1 = \frac{75-74}{10} = 0.1$$

$$z_2 = \frac{82-85}{7} = -0.43$$

The student's score on Exam 1 is relatively higher to the rest of the class.

2.5.2 CHEBYSHEV'S THEOREM

Chebyshev's Theorem can be used with the standard deviation to tell us *at least* what proportion of the data falls within k standard deviations of the mean. While this theorem can't provide useful information on the fraction of data within 1 standard deviation of the mean, it is useful when applied to data that fall within a larger range.

Formula 2.13: Chebyshev's Theorem

The proportion of any dataset lying within k standard deviation of the mean is *at least*:

$$\left(1 - \frac{1}{k^2}\right), \quad \text{for } k > 1$$

Example 2.15: Applying Chebyshev's Rule

- A. For any dataset, at least what percentage of data should lie within 2 standard deviations of the mean?

For $k = 2$: $1 - \frac{1}{2^2} = \frac{3}{4} \Rightarrow 75\%$

At least 75% of data should lie within 2 standard deviations of the mean.

- B. At least what percentage of data should lie within 3 standard deviations of the mean?

For $k = 3$: $1 - \frac{1}{3^2} = \frac{8}{9} \Rightarrow 88.89\%$

At least 88.89% of data should lie within 3 standard deviations of the mean.

- C. The average score on a pre-employment test is 26 with standard deviation 5. State the range in which at least 84% of the data will reside using Chebyshev's Theorem.

Find k :

$$0.84 = 1 - \frac{1}{k^2}$$

$$0.16 = \frac{1}{k^2}$$

$$k^2 = \frac{1}{0.16}$$

$$k = 2.5 \quad (k \text{ must be positive})$$

Thus, at least 84% of the data should lie within 2.5 standard deviations.

$$26 - 2.5(5) = 13.5 \text{ is the lower limit}$$

$$26 + 2.5(5) = 38.5 \text{ is the upper limit}$$

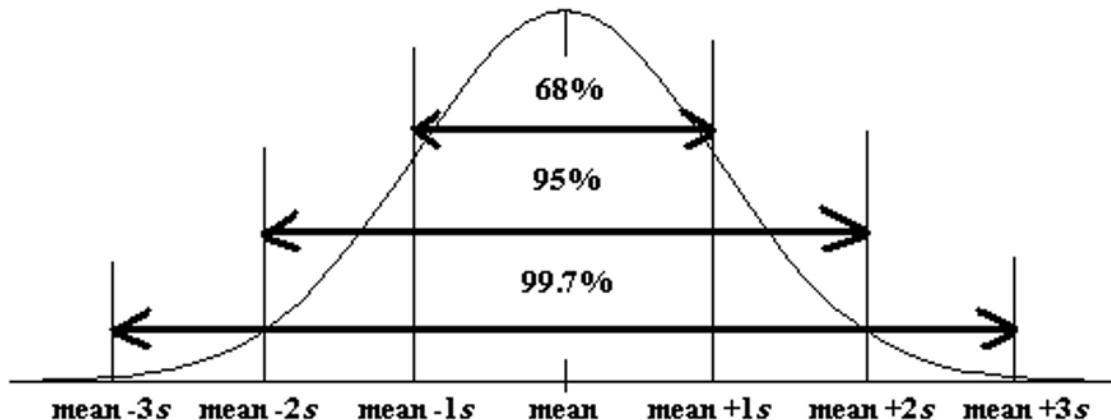
That is, at least 84% of the data should lie between 13.5 and 38.5.

2.5.3 THE EMPIRICAL RULE

Chebyshev's Theorem can only tell us the minimum amount of data within k standard deviations of the mean and k needs to be larger than 1. Can we do better? Depending on the shape of our data when we construct a histogram or other graphical display, we may be able to get a more accurate value for the percentage of data that is within 1, 2, or 3 standard deviations of the mean. **The Empirical Rule** states that if the shape of our data distribution is approximately symmetric and is what is often called bell-shaped or mound shaped, then:

1. Approximately **68%** of all observations fall within **one** standard deviation of the mean.
2. Approximately **95%** of all observations fall within **two** standard deviations of the mean.
3. Approximately **99.7%** of all observations fall within **three** standard deviations of the mean.

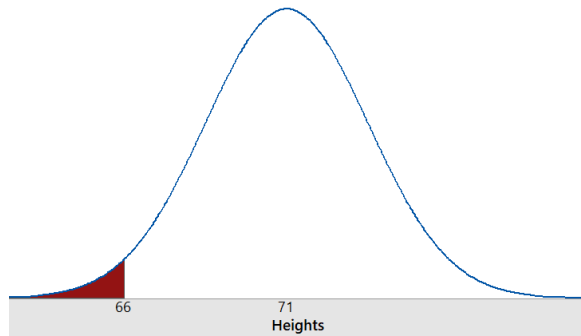
When considering using the Empirical Rule, draw a smooth curve over the graph of your data. If it is similar to the shape below, then you can use the Empirical Rule.



Example 2.16: Heights of Students

The heights of male students at Clemson have a histogram that is approximately bell-shaped with a mean of 71 inches and standard deviation of 2.5 inches. For each of the following questions, sketch a bell-shaped curve to represent the histogram. Shade the appropriate section under the curve, then use the Empirical Rule to solve.

- A. What percent of the male students are shorter than 66 inches?

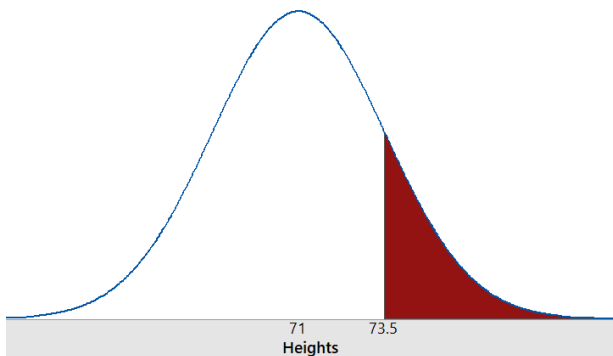


66 is two standard deviations below the mean height of 71 in.

95% of heights are within 2 std. deviations.
5% of heights are outside 2 std. deviations.

So, 2.5% of heights are represented in the lower tail. 2.5% of male students are shorter than 66 in.

- B. What percent are taller than 73.5 inches?

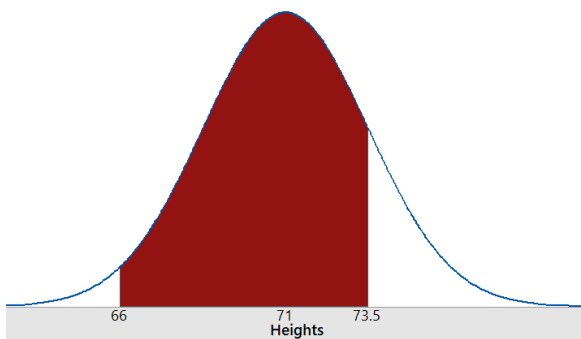


73.5 is one standard deviation above the mean height of 71 in.

50% of heights are greater than 71 in. and 34% of heights are between 71 and 73.5 in.

$50 - 34 = 16\%$ of males at Clemson are taller than 73.5 inches.

- C. What percent are between 66 and 73.5 inches?



68% of heights are between 68.5 and 73.5 in.

$\frac{95-68}{2} = 13.5\%$ of heights are between 66 and 68.5 in.

$68 + 13.5 = 81.5\%$ of males at Clemson are between 66 and 73.5 inches.

2.5.4 CHEBYSHEV'S THEOREM VS THE EMPIRICAL RULE

When should we use Chebyshev's theorem? When should we use the Empirical rule? The key point is to observe the [shape](#) of the data distribution. If the shape of distribution is not given or observable, we must use [Chebyshev's theorem](#). The theorem is derived from Chebyshev's inequality, which is a general truth applied to any distribution of data. Chebyshev's theorem is usually conservative, meaning it is only giving us a [lower bound](#) on the percentage of data. On the other hand, if we have the additional knowledge that the data distribution is [bell-shaped](#), we can apply a more accurate method. The Empirical rule is designed for this.

In Example 2.15 part A, without knowing any information about the shape of the data distribution, we have to use Chebyshev's theorem. We can only say that [at least](#) 75% of the data will lie within 2 standard deviations of the mean. Let's consider another scenario, if we know the data distribution is bell-shaped, such as in Example 2.16, then the Empirical rule will be applicable and will provide a more accurate estimation - approximately 95% of data should lie within 2 standard deviations of the mean. In this case, Chebyshev's theorem is still correct, since "at least 75%" is within "approximately 95%", but it's less accurate than the Empirical rule.

In conclusion, if the data distribution does not have a bell or mound shape with one peak, we must use Chebyshev's theorem to secure the correctness of our estimation; if the data distribution is bell-shaped, we prefer the Empirical rule to improve the accuracy of our estimation.

2.6 BOX PLOTS

Along with frequency/relative frequency tables, histograms, stem and leaf plots, and dot plots, [box plots](#) are another important method for graphically demonstrating the data. A box plot is a graphical display of the central tendency, the spread, sometimes the skewness, and the existence of outliers in the data. Box plots have the following advantages:

- They are easy to construct,
- They provide a convenient way to handle outliers,
- The construction of the plot is not subjective (like a histogram),
- They are useful for comparing medium to large datasets.

(Box plots should not be used for datasets smaller than size $n = 10$.)

To construct a box plot we need to discuss another numerical summary, the P^{th} percentile. Let's get familiar with it first.

2.6.1 PERCENTILE RANK AND P^{th} PERCENTILE

Recall the z-score which is a measure of relative position. We can also measure the relative position of a data point using its [percentile rank](#). The percentile rank of a particular data point tells us the percentage of values in dataset that is [at or below](#) this data point's value.

On the other hand, the value x whose percentile rank is $P\%$ is called the [\$P^{th}\$ percentil](#). For a set of data with n observations, the value x , as the P^{th} percentile of data, means:

- approximately P percent ($P\%$) of the data is less than or equal to x , and
- approximately $(100 - P)$ percent of the data is greater than x .

Consider the following examples:

- Suppose the 90^{th} percentile of the test scores is 85. It means approximately 90% of the test scores are less than or equal to 85. Approximately 10% of the test scores are greater than 85. Equivalently, the percentile rank of 85 is 90%.
- A baby's head circumference of 50 cm is the 30^{th} percentile if approximately 30% of the baby's head circumferences are less than or equal to 50 cm and approximately 70% are greater than 50 cm. Equivalently, the percentile rank of 50 cm is 30%.

Formula 2.14: Finding the P^{th} Percentile

To find the P^{th} percentile of a dataset with n observations:

- (1) Form an ordered array by placing the data in order from smallest to largest.
- (2) Calculate the $\ell = n \left(\frac{P}{100} \right)$, where ℓ is an approximate location index for the P^{th} percentile.
- (3) If ℓ is not an integer, then round ℓ up to the next greatest integer and find the data value at the ℓ^{th} position of the ordered array. (For example, if $\ell = 7.1$, then round ℓ up to 8. The value at the 8th position in the ordered array will be the P^{th} percentile.)

If ℓ is an integer value, then take the average of data values at the ℓ^{th} location and the $(\ell + 1)^{th}$ location. (For example, if $\ell = 7$, then take the average of values at the 7th and 8th positions. The average will be the P^{th} percentile.)

Note that there is more than one method of calculating percentiles. For example, JMP uses a slightly different calculation to find percentiles, as do many calculators. For this class, you should **always** use the formula presented above *unless specifically stated to do otherwise*.

Instructions on how to calculate percentiles using JMP can be found in the **JMP Instructions for STAT 3090**.

2.6.2 FIVE-NUMBER SUMMARY

The 25th, 50th, and 75th percentiles are known as quartiles and are often of particular interest when describing a dataset. These values are denoted Q_1 , Q_2 , and Q_3 . Note that the 50th percentile Q_2 , is equal to the median. Another common way to summarize a data set is to use a five-number summary which includes: **Minimum, Q_1 , Median, Q_3 , Maximum.**

Example 2.17: Number of Defects (continued)

The following represents the number of defects per shipment of a random selection of computer parts produced by a technology company. The values have been arranged in ascending order for convenience.

0	0	0	0	0	1	1	1
1	2	2	2	2	3	3	3
3	3	4	4	4	5	5	8

- A. Find and interpret the 30th percentile.

Find ℓ : $\ell = 24(0.30) = 7.2$. This is not an integer, so round up to 8.

The 8th value in the dataset is 1. Thus, 1 is the 30th percentile in this data set.

Interpretation: Approximately 30% of the shipments have 1 or fewer defect.

- B. Find the five-number summary.

Min: 0

Q_1 : $\ell_1 = 24(0.25) = 6$. This is an integer, so find the mean of the 6th and 7th values. $Q_1 = \frac{1+1}{2} = 1$

Q_2 (Median): $\ell_2 = 24(0.5) = 12$. This is an integer, so find the mean of the 12th and 13th values. Q_2 (Median) = $\frac{2+2}{2} = 2$

Q_3 : $\ell_3 = 24(0.75) = 18$. This is an integer, so find the mean of the 18th and 19th values. $Q_3 = \frac{3+4}{2} = 3.5$

Max: 8

- C. Determine the percentile rank for 4 defects. Include all of the data points having 4 defects.

$\frac{21}{24}$ or 87.5% of the values are at or below 4 defects.

2.6.3 INTERQUARTILE RANGE (IQR)

The **interquartile range**, abbreviated IQR, is a measure of spread (variability). It describes the range of the middle fifty percent of the data.

Formula 2.15: Interquartile Range

$$IQR = Q_3 - Q_1$$

Example 2.18: Number of Defects (continued)

What is the **IQR** value for the number of defects data in Example 2.17?

$$Q_1 = 1 \text{ and } Q_3 = 3.5$$

$$IQR = Q_3 - Q_1 = 3.5 - 1 = 2.5 \text{ defects}$$

2.6.4 FENCES AND OUTLIERS

An **outlier** is an observation that is unusually large or small compared to the other values in the dataset. Outliers are typically attributed to one of the following causes...

- The measurement is observed, recorded, or entered into the computer incorrectly.
- The measurement comes from a different population.
- The measurement is correct but represents a rare (chance) event.

The determination of outliers by eyeball estimation is not always reliable. Box plots provide us a more scientific way for this. Using the values of Q_1 , Q_3 and IQR , we will calculate outlier fences and identify data points that are outside of these fences as outliers.

Formula 2.16: Fences and Outliers

For a dataset with quartiles Q_1 and Q_3 and $IQR = Q_3 - Q_1$, the calculations for the **outlier fences** needed to identify outliers are as follows:

$$\text{Lower Fence} = Q_1 - 1.5(IQR)$$

$$\text{Upper Fence} = Q_3 + 1.5(IQR)$$

Data points below the lower fence are considered low outliers. Data points above the upper fence are considered high outliers.

2.6.5 CONSTRUCTING A BOX PLOT

- (1) Find the **five number summary** (Min, Q_1 , Median, Q_3 , Max).
- (2) Draw a number line (axis) that includes the range of your data.
- (3) Draw a box from Q_1 to Q_3 .
- (4) Draw a line inside the box at the Median.
- (5) Calculate the **outlier fences** (see Formula 2.16) and identify any low or high outliers.
- (6) Draw whiskers that extend to the largest and smallest data values **inside** the outlier fences.
- (7) Mark any outliers with an asterisk, star, or point.

Note that outlier fences are **not** directly drawn on the box plot, and the whiskers are **not** drawn to the fence values! The fence values are simply for **reference** when constructing your box plot.

Example 2.19: Constructing a Box Plot

The five-number summary for the number of defects data in Example 2.17 is:

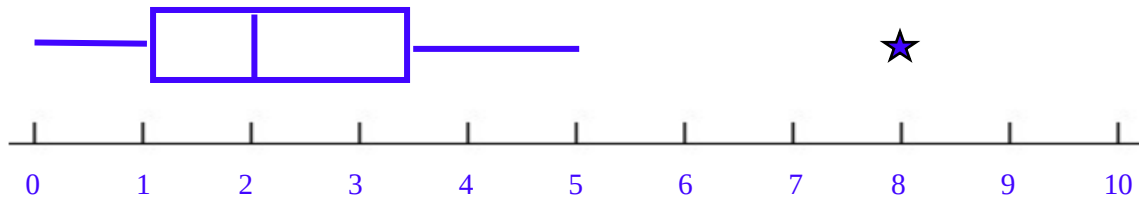
Min: 0 Q_1 : 1 Med: 2 Q_3 : 3.5 Max: 8

Sketch a box plot for the number of defects data.

Fences:

Lower Fence: $Q_1 - 1.5(IQR) = 1 - 1.5(2.5) = -2.75$ (No data points are less than -2.75 , so there's no low outliers. The smallest value inside lower fence is just the minimum, which is 0. Extend left whisker to 0.)

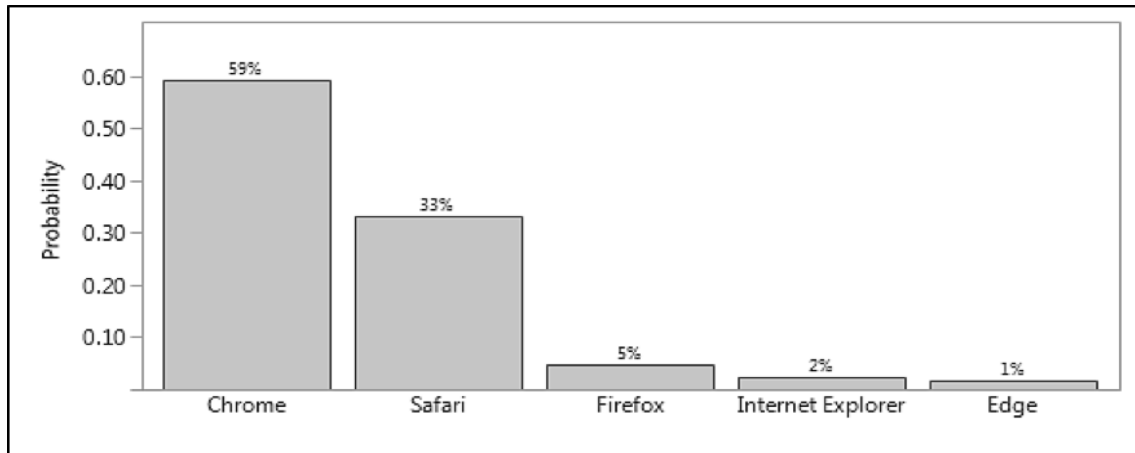
Upper Fence: $Q_3 + 1.5(IQR) = 3.5 + 1.5(2.5) = 7.25$ ($8 > 7.25$, so it is a high outlier. The largest value inside upper fence is 5. Extend right whisker to 5, and mark 8 as an outlier.)



MY NOTES:

PRACTICE EXERCISES FOR SECTION 2

1. A survey asked students to report the web browser they primarily use. The relative frequency bar graph below summarizes the 639 student responses.



- (a) How many **more** respondents primarily use Chrome than primarily use Firefox?

Approximately 345 respondents.

- (b) What proportion of respondents use a browser **other than** Chrome?

$$1 - 0.59 = 0.41$$

2. Suppose the following relative frequency distribution represents the proportion of majors for STAT 3090 students.

Major	Relative Frequency
Business	0.429
Finance	0.179
Computer Science	0.146
Marketing	0.211
Other	0.035

Which of the following graphs is most appropriate for this data and why?

- A. Since the variable Major is quantitative, we can use a histogram.
 B. *Since the variable Major is qualitative, we can use a pie chart.*
 C. Since the variable Major is qualitative, we can use a stem and leaf plot.
 D. Since the variable Major is quantitative, we can use a dot plot.

3. Based on the data provided by:

[https://countrymeters.info/en/United States of America \(USA\)](https://countrymeters.info/en/United_States_of_America_(USA)) in 2022 the population of the United States was about 336.20 million, of which 165.97 million are male and 170.23 million are female.

- (a) What was the proportion of females in United States in 2022? Is this a population proportion or sample proportion? Include the appropriate symbol in your answer.

It's a population proportion.

$$p = \frac{x}{N} = \frac{170.23}{336.20} \approx 0.506$$

- (b) Clemson government randomly collected a sample of size 112 composed of 51 males and 61 females. What is the proportion of females in these 112 people? Is this a population proportion or sample proportion? Include the appropriate symbol in your answer.

It's a sample proportion.

$$\hat{p} = \frac{x}{n} = \frac{61}{112} \approx 0.545$$

4. At the beginning of a semester, Jason asked 200 Clemson undergraduate students how many years they have attended classes at Clemson. Use this information to complete the table:

Years	Frequency	Relative Frequency	Cumulative Relative Frequency
(0, 1.5]	200(.2)=40	20%	20%
(1.5, 2.5]	60	30%	50%
(2.5, 3.5]	80	40%	90%
(3.5, ∞]	20	10%	100%

5. Suppose we have a **cumulative** relative frequency table for a data set:

Data	Cumulative Relative Frequency
(12, 15)	0.12
[15, 18)	0.45
[18, 21)	0.62
[21, 24)	0.84
[24, 27)	1

- (a) What is the proportion of data which are less than 18?

0.45

- (b) What is the proportion of data which are greater than or equal to 21?

$1 - 0.62 = 0.38$

6. The following data represent the number of days required to collect insurance payments for a random sample of customers of a local dentist. Create a stem and leaf plot for these data. Remember to include a title and a key for your stem and leaf diagram.

Number of Days to Collect Payment

22 24 28 29 30
 31 32 33 33 34
 35 35 35 36 36
 37 38 38 39 41
 43 45 47 48 55
 56 56 60 65 66

NUMBER OF DAYS TO COLLECT PAYMENT

Key: 2 | 2 = 22 days

2 | 2 4 8 9
 3 | 0 1 2 3 3 4 5 5 5 6 6 7 8 8 9
 4 | 1 3 5 7 8
 5 | 5 6 6
 6 | 0 5 6

7. Wanda works in the Botanical Gardens and is studying how to best care for the rose bushes. One of the ways she evaluates how the bushes are doing is by observing how many flowers a single bush produces. The number of roses she finds on each of 20 bushes are given below.

9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 6, 5, 4, 10, 9, 6, 7, 4

- (a) Find the mean number of roses per bush.

$$\bar{x} = \frac{9+2+5+\dots+7+4}{20} = \frac{132}{20} = 6.6 \text{ roses}$$

- (b) Find the median of the dataset.

Sort the data: 2 3 4 4 4 4 5 5 6 6 7 7 7 8 9 9 9 10 11 12

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

- (c) Find the mode of the dataset.

$$\text{Mode} = 4 \text{ roses}$$

- (d) Find the variance of the dataset.

$$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1} = \frac{1018 - \frac{(132)^2}{20}}{19} = 7.7263 \text{ roses}^2$$

- (e) Find the standard deviation of the dataset.

$$s = \sqrt{7.7263} \approx 2.8 \text{ roses}$$

8. Nutritionists measured the sugar content (in grams) for 32 drinks at All In Coffee Shop. The drinks had a mean of 18 g and a standard deviation of 5 g. A Grande Mocha Cappuccino at All In contains 14 g of sugar.

(a) Calculate the standardized score (z-score) for the Grande Mocha Cappuccino.

$$z = \frac{14 - 18}{5} = -0.8$$

(b) Interpret the z-score.

The sugar content in a Grande Mocha Cappuccino is 0.8 standard deviation below the mean sugar content in drinks at All In Coffee Shop.

- (c) A different coffee shop called Creek Coffee serves drinks containing a mean of 16 g of sugar with a standard deviation of 6 g. A Grande Mocha Cappuccino here contains 13 g of sugar. Calculate the z-score for this drink at Creek Coffee.

$$z = \frac{13 - 16}{6} = -0.5$$

- (d) At which coffee shop does the drink have less sugar relative to the other drinks sold at its respective location?

The drink has relatively less sugar at All In Coffee Shop.

9. The lifespans of tigers in a particular zoo have a normal (bell-shaped) distribution. The average tiger lives 17 years with a standard deviation of 2.5 years. Use the Empirical Rule to estimate the proportion of tigers at this zoo that live 14.5 years or longer.

*The Empirical Rule says that 68% of the data lies between 14.5 and 19.5 years, so 34% of the data lies between 14.5 and 17 years, and 50% lies above 17 years.
 $34\% + 50\% = 84\%$ of tigers at this zoo live 14.5 years or longer.*

10. Mr. Darcy's nineteenth century English literature class of 25 students had an average of 85 on their last test. Miss Bennet's class of 22 students had an average of 87 on the same test. What is the average of the two classes combined?

- A. 85.94
- B. 86.00
- C. 91.80
- D. 80.78

11. A sample of sociology books has a mean cost of \$70 and a standard deviation of \$10. Use **Chebyshev's** theorem to answer the following questions.

- (a) At least what percentage of sociology books cost between \$50 and \$90?

50 is 2 standard deviations below mean of 70 (equivalently, 90 is 2 standard deviations above mean of 70), so $k = 2$. Therefore, the percentage is:

$$1 - \frac{1}{k^2} = 1 - \frac{1}{2^2} = 0.75 = 75\%$$

At least 75% of sociology books cost between \$50 and \$90.

- (b) State the range of price in which at least 88.89% of the data will reside.

Solving $1 - \frac{1}{k^2} = 0.8889$ we get $k = 3$.

Thus, the lower limit is $70 - 3(10) = 40$ and upper limit is $70 + 3(10) = 100$.

Therefore, at least 88.89% of the data will reside in the range of \$40 to \$100.

12. The following dataset represents the lengths (in cm) of fish caught in a pond at a campground in an afternoon.

9 11 20 10 12 17 15 5 16 15 28 6

- (a) Find and interpret the 80th percentile of the data.

Sort the data: 5 6 8 10 11 12 15 15 16 17 20 28

$\ell = n \left(\frac{P}{100} \right) = 12(0.80) = 9.6$. Not an integer. Round up to 10.

Find the data in the 10th position: 17 cm. So, 17 cm is the 80th percentile of the data.

Interpretation: Approximately 80% of the fish caught this afternoon were 17 cm or shorter.

- (b) Write the five number summary for the dataset.

Median: $\frac{12+15}{2} = 13.5$ cm

$Q_1: \ell_1 = 12(0.25) = 3$. Average the 3rd and 4th values. $Q_1 = \frac{8+10}{2} = 9$ cm

$Q_3: \ell_2 = 12(0.75) = 9$. Average the 9th and 10th values. $Q_3 = \frac{16+17}{2} = 16.5$ cm

Five Number Summary:

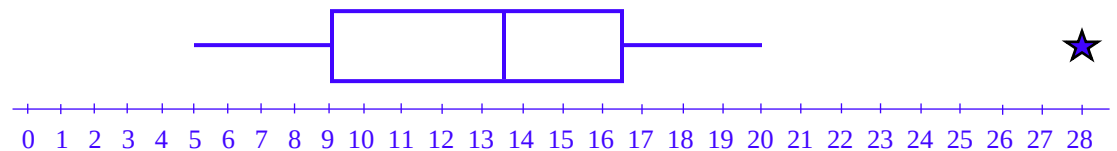
Min: 5 cm Q_1 : 9 cm Med: 13.5 cm Q_3 : 16.5 cm Max: 28 cm

- (c) Construct a box plot for the dataset.

Fences:

Lower: $Q_1 - 1.5(\text{IQR}) = 9 - 1.5(7.5) = -2.25$ (No data less than -2.25 .)

Upper: $Q_3 + 1.5(\text{IQR}) = 16.5 + 1.5(7.5) = 27.75$ (28 is an outlier.)



SECTION 3: DESCRIBING QUANTITATIVE DATA DISTRIBUTIONS

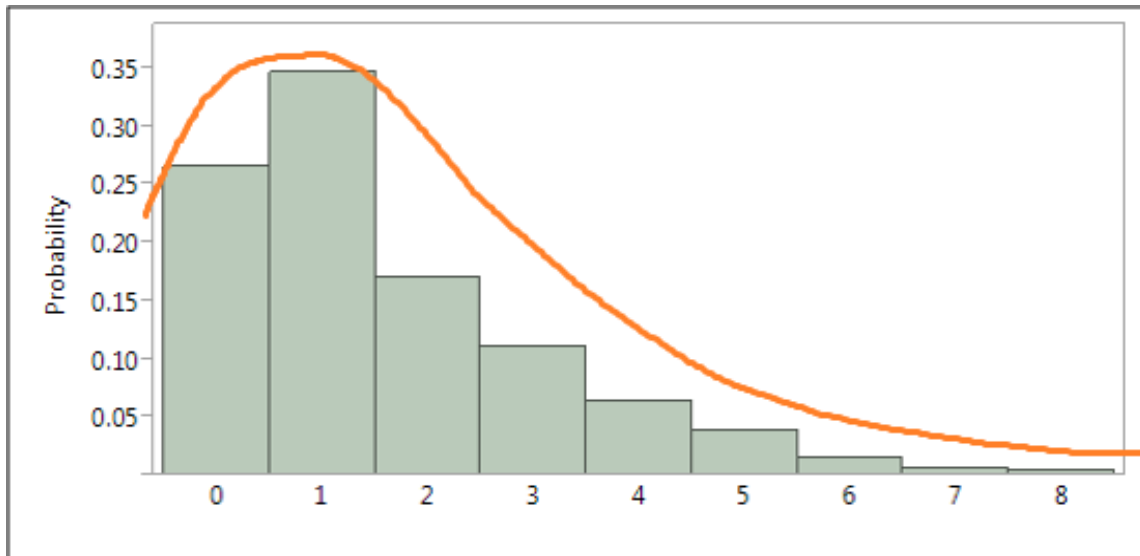
(HAWKES 3.4, 4.1, 4.2, 4.3)

When describing a distribution of quantitative data, include descriptions of the following elements...

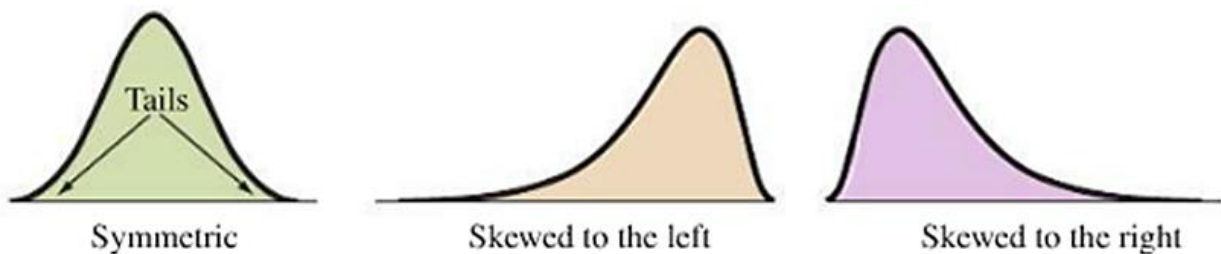
- Shape (type of symmetry or skewness)
- Center (mean, median, or mode)
- Spread (range, interquartile range, variance, or standard deviation)
- Unusual (any outliers or gaps in your data)

3.1 SHAPE

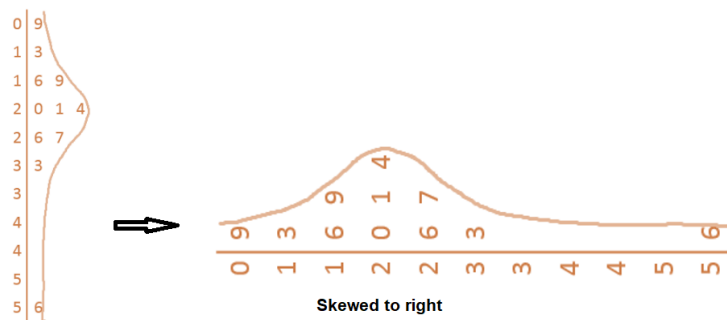
To determine the shape of a distribution, we need to mention whether the distribution is symmetric or skewed. If it is symmetric, state whether it is unimodal, multimodal, or uniform, etc. If it is skewed, state whether it is left-skewed or right skewed. It is often helpful to outline the histogram, stem and leaf plot or dot plot of our data with a smooth curve such as is pictured below.



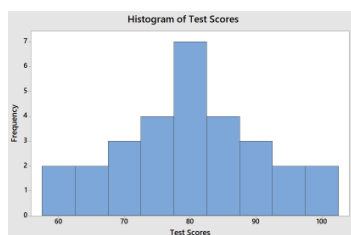
- A distribution is symmetric if the right and left sides of the graph are approximate mirror images.
- A distribution is skewed left if the left side of the graph extends much further out than the right.
- A distribution is skewed right if the right side of the graph extends much further out than the left.



You can also tell the direction of skewness from vertically displayed graphs. You will need to rotate them to a horizontal layout which puts the horizontal axis in increasing order:

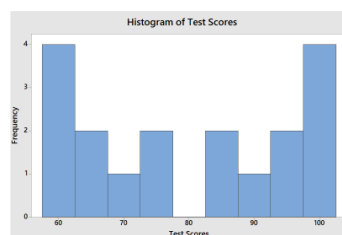


Simply saying that a distribution is “symmetric” to describe its shape is *not* sufficient! If possible, you should always include a description of the *type* of symmetry present.



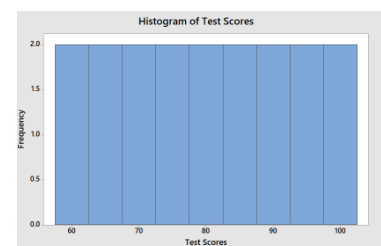
Unimodal

One peak



Multimodal

More than one peak



Uniform

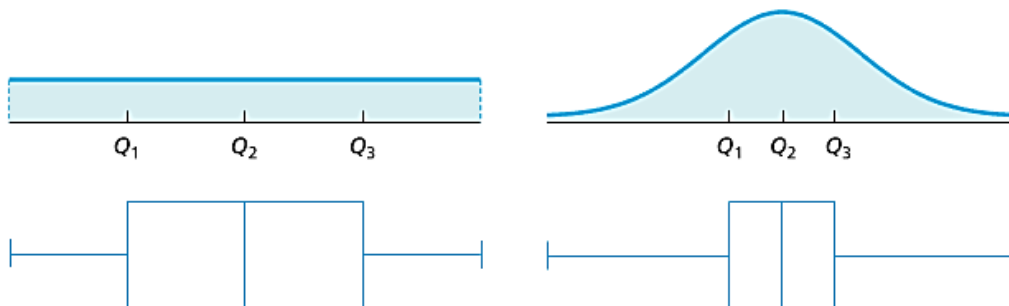
no outstanding peak

You can also tell the shape of a distribution from a box plot. The following are suggested guidelines for describing the shape of a distribution when you have only a box plot,

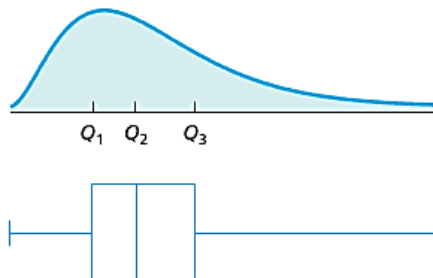
- For distributions that are approximately *symmetric*, the median will be near the *center* of the box. The left and right whiskers will likely have *similar* lengths.
- For distributions that are *skewed right*, the median tends to be slightly *left* of the center of the box. The right whisker tends to be longer than the left, or there may be high outliers.
- For distributions that are *skewed left*, the median tends to be slightly *right* of the center of the box. The left whisker tends to be longer than the right, or there may be low outliers.

Some examples of distributions and their corresponding box plots are pictured below.

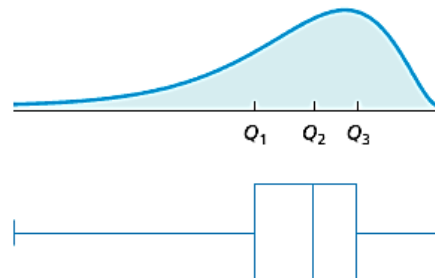
Symmetric Distributions



Right-Skewed Distribution



Left-Skewed Distribution



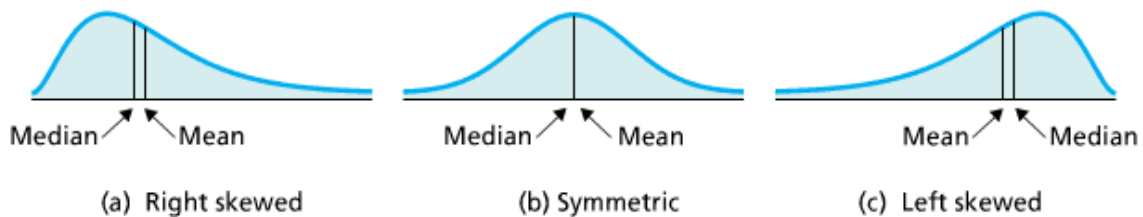
3.2 CENTER

We know that mean and median are both measures of center, however, we need to carefully choose from them to describe the center of a distribution especially when it is skewed. A [resistant measure](#) is more reasonable to report as the measure of center for a skewed distribution.

3.2.1 RESISTANT MEASURES VS NON-RESISTANT MEASURES

Resistant measures are statistics that are not affected by [outliers](#). Recall outliers are observations that fall outside the overall pattern of data and are typically considered to be “extreme” values. In a dataset with these outliers, the mean will be pulled in the direction of the extreme values. The median, however, is not affected, so median is a resistant measure while mean is non-resistant.

What is the relationship between the mean and median in a dataset that falls in a symmetric distribution? In a skewed distribution?



Generally, if the shape of the distribution is...

- Right-skewed: $\text{Mean} > \text{Median}$
- Approximately symmetric: $\text{Mean} \approx \text{Median}$
- Left-skewed: $\text{Mean} < \text{Median}$

3.2.2 WHICH MEASURE OF CENTER IS APPROPRIATE?

Non-resistant measures are usually easy to calculate. However, if the distribution is skewed with a tail because of the existence of extreme values or outliers, using a non-resistant measure to describe the center will counter our understanding of “the most typical” value of the distribution. Thus, if the distribution is:

- Symmetric: Mean and median are approximately equal, so report [mean](#) as the measure of center.
- Skewed: Mean is pulled in the direction of skewness, so report [median](#) as the measure of center.

Since we will use the same idea to describe the spread of a distribution later, let's see what measures are resistant (and you think about why).

	Resistant	Non-resistant
Center	<i>median</i>	<i>mean</i>
Spread	<i>IQR</i>	<i>range, variance, standard deviation</i>

3.3 SPREAD

Similar to describing the center, when the distribution is skewed, we choose resistant measure to report for spread, which is the [*IQR*](#). This will minimize the effects caused by extreme values or outliers. When the distribution is symmetric, we prefer to use the non-resistant measures to report for spread, e.g., [*range, variance, and standard deviation*](#).

3.4 UNUSUAL VALUES

Report the locations of gaps and outliers if there are any present. Gaps tend to split the data into obvious clusters. They are locations where there are no data values within the interior of the data set.

In a box plot, report the locations of points marked by asterisks as outliers. Gaps in data do not appear in box plots. In a histogram, stem and leaf plot, or dot plot report the locations of outliers by eyeball estimation.

3.5 COMPARING TWO OR MORE DISTRIBUTIONS

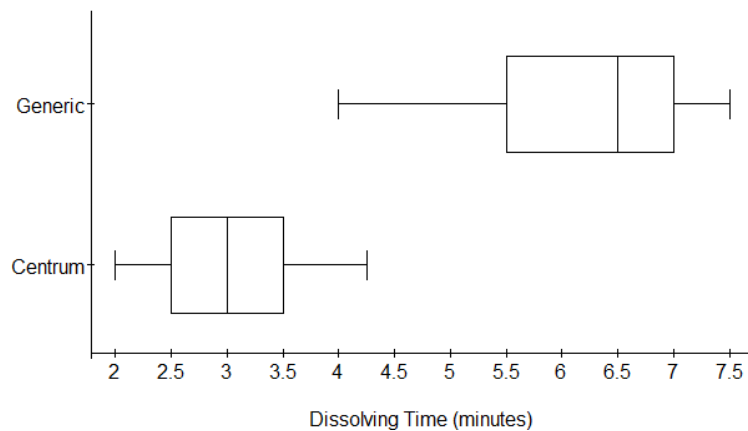
When comparing two or more distributions, we need to compare the shapes, centers, spreads, and unusual values as we learned before. You should...

- Provide a comparison of the [*symmetry*](#) and/or the [*skewness*](#) for each distribution.
- Provide a comparison of the measures of [*center*](#) for the distributions. Discuss which is larger/smaller. **Always** use the same measures of center for all the distributions. Otherwise, the comparison is not valid. If at least one of the distributions is strongly skewed or has outliers, it is usually best to compare the [*medians*](#). Otherwise, it is standard to compare means.
- Provide a comparison of the measures of [*spread*](#) for each distribution. Discuss which is larger/smaller. Similar to your comparison of the centers, **always** use the same measures of spread for all the distributions. If at least one of the distributions is strongly skewed or has outliers, it is usually best to compare the [*IQR values*](#). Otherwise, use the standard deviations.
- Note unusual values (gaps and outliers) when appropriate.

Drawing side-by-side plots on the same scale allows several distributions to be easily compared.

Example 3.1: Centrum Vitamins

Do Centrum vitamins dissolve faster than the corresponding generic brand? To find out, a student randomly selected 30 vitamins of each brand. The student used vinegar as a proxy for stomach acid and measured the time (in minutes) it took for a vitamin to completely dissolve. Using the side-by-side box plots shown below, compare the distribution of dissolving time for the Centrum and generic brand vitamins.



Shape: The distribution of dissolving times is approximately symmetric for Centrum vitamins but skewed left for the generic brand.

Center: Centrum vitamins typically dissolve faster, with a median of 3 minutes compared to the median for the generic brand of 6.5 minutes.

Spread: Centrum vitamins have more consistency (less variability) in their dissolving times, with the IQR for Centrum of 1 min being smaller than the IQR of 1.5 min for the generic brand.

MY NOTES:

PRACTICE EXERCISES FOR SECTION 3

1. Suppose the distribution of heights (in feet) of waterfalls in a state is:

0	8
1	0 0 0 2 2 2 5 5 5 5 5 5 8 8
2	0 0 0 5 5 8
3	0
4	0 0
5	0 0 0
6	
7	0 5
8	0 0
9	
10	0
11	
12	0 5
13	
14	0
15	0

key: 1|2 = 12 feet

- (a) Which of the following can best describe the shape of this distribution?

- A. It's left-skewed.
- B. It's right-skewed.*
- C. It's symmetric

- (b) Which of the following is most likely true for this distribution?

- A. mean > median*
- B. mean < median
- C. mean = median

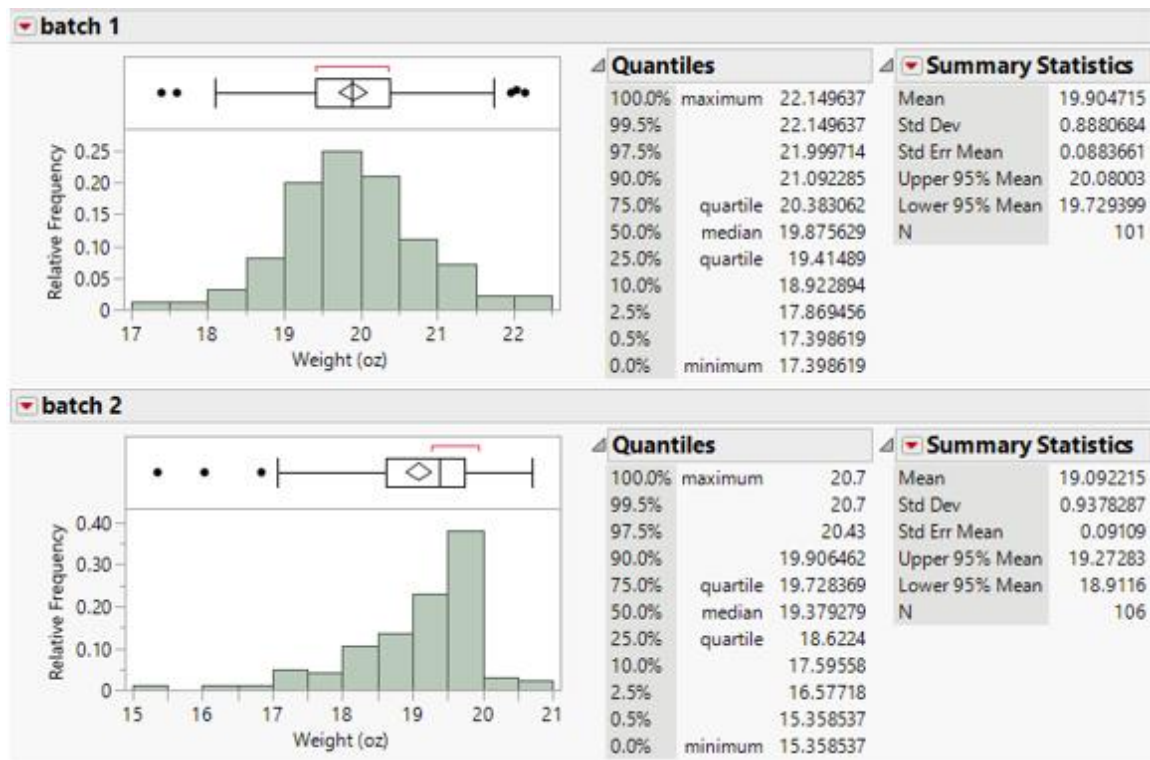
2. Suppose we have a data set with a skewed distribution. Which of the following is the best measure to use to report the center of the distribution?

- A. mean
- B. median*
- C. IQR
- D. It depends on whether it's left- or right-skewed.

3. Suppose we have a data set with a skewed distribution. Which of the following is the best measure to use to report the spread of the distribution?

- A. variance
- B. median
- C. IQR*
- D. It depends on whether it's left- or right-skewed.

4. Consider a brand of snack produced in 2 batches. The distributions of weights and summary statistics are presented below:



- (a) Compare their shapes.

Batch 1 is approximately symmetric and unimodal, while batch 2 is left-skewed.

- (b) Compare their centers.

Batch 1 is centered at the median of 19.88 oz which is larger than the median of batch 2 at 19.38 oz.

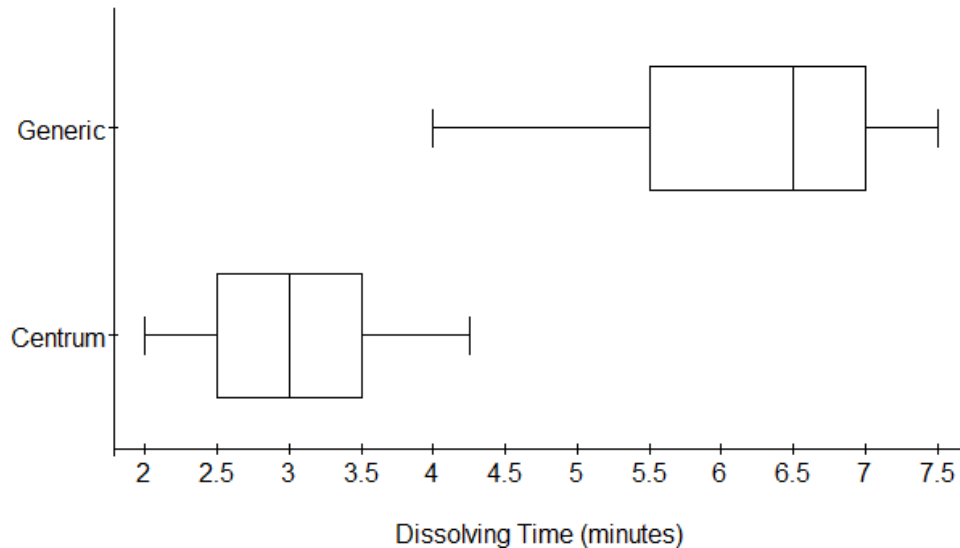
- (c) Compare the spreads.

The IQR of batch 1 is 0.97 oz which is smaller than the IQR of batch 2 at 1.11 oz.

- (d) Compare the outliers.

Batch 1 has 2 low outliers at around 17.5 oz and 3 high outliers at around 22 oz. Batch 2 has 3 low outliers in the range of 15 oz to 17 oz.

5. Do Centrum vitamins dissolve faster than the corresponding generic brand? To find out, a student randomly selected 30 vitamins of each brand. The student used vinegar as a proxy for stomach acid and measured the time (in minutes) it took for a vitamin to completely dissolve. Using the side-by-side boxplots shown below, compare the distribution of dissolving time for the Centrum and generic brand vitamins.



Which of the following statements is true?

- A. All “Generic” dissolved slower than “Centrum”.
- B. More than 50% of “Generic” took 7 minutes or longer to dissolve.
- C. Some “Generic” dissolved faster than the median of dissolving time of “Centrum”.
- D. “Centrum” has a smaller IQR compared to “Generic”.

SECTION 4: INTRODUCTION TO PROBABILITY

(HAWKES 5.1, 5.2, 5.3, 5.4)

This section introduces **probability** – the way we quantify uncertainty. Probability is concerned with measuring the chances of the possible outcomes for **random phenomena**, the many things in your life for which the outcome is uncertain. For example:

- You could use probability to find the chance you'll win the lottery,
- You could use the rules of probability to find the likelihood that an employer's drug test correctly detects whether or not you've used drugs.

The ideas in this section set the foundation for how we will use probability later in the course to make inferences about a population based on sample data.

We begin our study of probability by looking at some simple examples. The principles that we learn will then be applied to more complex situations.

In probability, an experiment is an act or process of observation with uncertain results that can be repeated. The result of any single trial of the experiment is not known ahead of time. Some examples are:

- Rolling dice, flipping a coin, or spinning a wheel.
- Observing whether a patient develops side effects when taking a new drug.

Each of the previous examples meets the requirements for a random experiment:

- (1) There is one distinct outcome for each trial.
- (2) The outcome is uncertain.
- (3) The set of all outcomes can be specified and is called the **sample space**.

4.1 PROBABILITY AS RELATIVE FREQUENCY

In an experiment, the probability of an outcome is the proportion of times that outcome would occur in a **long run** of observations. This proportion is the relative frequency of the outcome. Therefore, **probabilities are ALWAYS between 0 and 1**. Probabilities may also be reported as a percentage.

With a small number of observations, outcomes of a probability experiment may look different than expected. For instance, if you were to flip a coin 4 times, you may see 4 heads in a row. If you were to flip a coin 1000 times, the proportion of heads will be near 0.50.

With randomized experiments, the proportion of times that something happens is highly random and *variable in the short run*, but very *predictable in the long run*.

In a probability experiment, each possible outcome is called a sample point, which is the most basic result of an experiment.

The set of all sample points (outcomes) for an experiment is called the sample space. We denote this space by $S = \{ \text{list of outcomes} \}$.

Example 4.1: Sample Spaces

List or describe the **sample space** for each scenario.

A. Toss a coin.

$$S = \{ \text{heads, tails} \} \text{ or } S = \{ H, T \}$$

B. When investing in a stock, check the price change since yesterday.

$$S = \{ \text{stock value increases, stock value decreases, stock value stays the same} \}$$

C. Select an individual at random and observe their shoe size.

$$S = \{ \dots 3, 3.5, 4, 4.5, 5, 5.5, 6, \dots \}$$

D. Shoot 5 free throws and record the number of free throws made.

$$S = \{ 0, 1, 2, 3, 4, 5 \}$$

4.2 PROBABILITY EVENTS AND RULES

Sometimes, the result of an experiment can occur in more than one way. An event is a specific collection of sample points. There are two types:

- (1) A **simple** event is a single sample point.
- (2) A **compound** event is a collection of two or more sample points.

Events are typically denoted with capital letters.

Example 4.2: Events

Determine if each event is simple or compound.

- A. Observe the up face on a coin. Let the event A = the coin comes up heads.

$A = \{H\}$
Simple Event

- B. Draw two marbles at random from a bag of black and red marbles. Let the event B = both marbles are red.

$B = \{RR\}$
Simple Event

- C. Let the event C = at least one marble is black.

$C = \{RB, BB\}$ - if the order of the outcomes does not matter
or
 $C = \{RB, BR, BB\}$ - if outcomes record the sequence of draws
Compound Event – there is more than one way for this to occur:

- D. Invest a stock, Let the event D = the price of the stock does not decrease.

$D = \{\text{price stays the same, price increases}\}$
Compound Event – the stock price can either stay the same or increase.

4.2.1 ASSIGNING PROBABILITIES TO EVENTS

Associated with **each** outcome (i.e., each simple event) in the sample space of a probability experiment is a number, called the outcome's **probability**. The probabilities for an experiment must meet two criteria. Suppose we have a probability experiment with n outcomes (there are n sample points). We will call each of these simple events E_i where $i = 1 \dots n$. Let the associated probability for event E_i be $P(E_i)$.

- (1) All sample point probabilities must lie between 0 and 1:

$$(0 \leq P(E_i) \leq 1)$$

- (2) The probabilities of all sample points within a sample space must sum to 1:

$$\sum_{i=1}^n P(E_i) = 1$$

This property should make sense once we recall the relative frequency definition of probabilities in which the probabilities are the proportions of the occurrences over all possible observations. Now, suppose event A is compound event comprised of several simple events. Then the probability of event A is the sum of probabilities of the simple events:

$$P(A) = \sum_{i \in A} P(E_i)$$

If all n outcomes of an experiment are equally likely, meaning that each simple outcome has a probability equal to $\frac{1}{n}$, then:

$$P(A) = \frac{\text{number of outcomes in } A}{n}$$

You can write probabilities as fractions or decimals. Typically, probabilities in decimal form are rounded to 4 decimal places if the directions for a problem do not specify the number of decimal places to use.

Example 4.3: Medal Counts

Consider the Olympic medal counts in the table to the right.

- A. What is the probability a randomly selected medal was won by the United States?

$$P(U) = \frac{121}{974} = 0.1242$$

- B. What is the probability a randomly selected medal was won by China?

$$P(C) = \frac{70}{974} = 0.0719$$

Country	Count
United States (U)	121
China (C)	70
Great Britain (GB)	67
Russia (R)	56
Other (O)	660
Total	974

4.2.2 COMPLEMENTARY EVENTS

The complement of an event A is the event that A does *not* occur; in other words, it is the event consisting of all sample points that are not in event A .

Notation: A^c

Example 4.4: Complementary Events

Define the complement of each event.

A. G = gold medal

G^c = not gold (silver or bronze)

B. U = won by the U.S.

U^c = not won by the U.S.

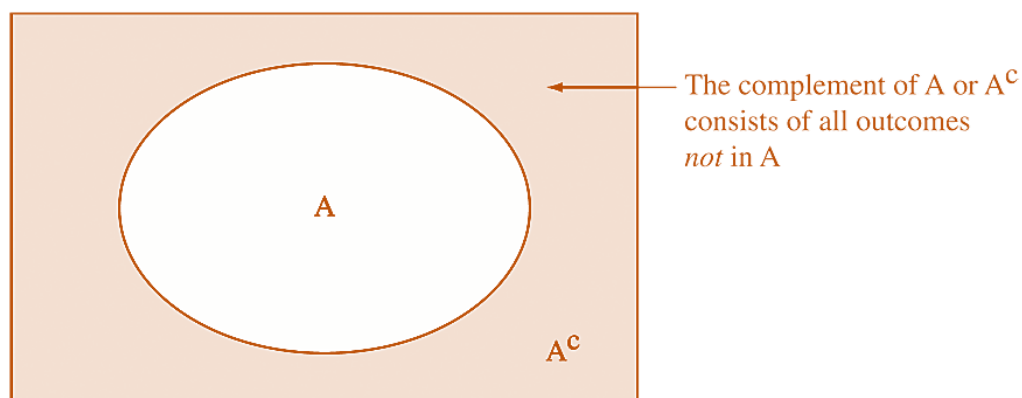
We can find the probability of event using the **complement rule**.

Formula 4.1: The Complement Rule

If A^c is the complement of A , then the probability of A^c is:

$$P(A^c) = 1 - P(A)$$

The idea of the complement of an event can be illustrated using a **Venn diagram** – a diagram that uses a box to represent the sample space and the areas inside to represent certain events.



Example 4.5: Medal Counts (continued)

- A. What is the probability that a randomly selected medal was won by the United States?

$$P(U) = \frac{121}{974} = 0.1242$$

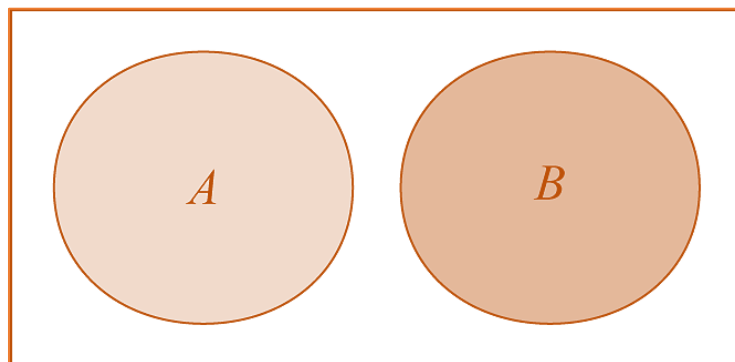
Country	Count
United States (U)	121
China (C)	70
Great Britain (GB)	67
Russia (R)	56
Other (O)	660
Total	974

- B. What is the probability that a randomly selected medal was won by a country other than the United States?

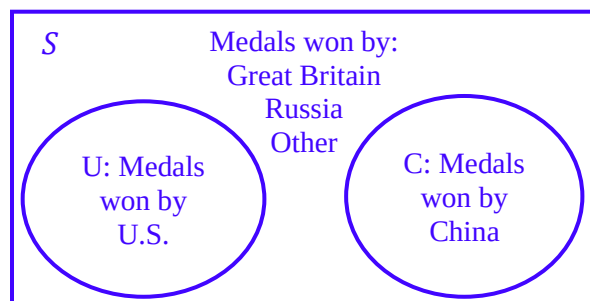
$$P(U^c) = 1 - 0.1242 = 0.8758$$

4.2.3 MUTUALLY EXCLUSIVE EVENTS

Two events A and B are **mutually exclusive** (or disjoint) if they do not have any common outcomes.

**Example 4.6: Medal Counts (continued)**

Are the events U and C mutually exclusive? (Hint: Draw a Venn diagram.)



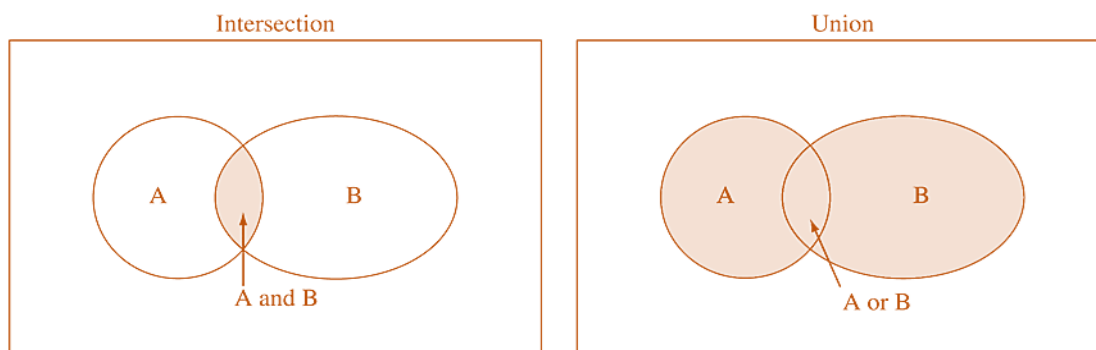
Since there is no overlap between events U and C , they are mutually exclusive.

$$P(U \text{ and } C) = 0$$

4.2.4 INTERSECTIONS AND UNIONS OF EVENTS

The **intersection** of two events A and B consists of the outcomes in both A and B .

The **union** of two events A and B consists of outcomes that are in A or B . In probability, this means that A occurs, or B occurs, or both occur.



Example 4.7: Medal Counts (continued)

A. What is the **union** of the events U and C ?

$U \cup C = U \text{ or } C$
(Medals won by U.S. OR won by China)

B. What is the **intersection** of events U and C ?

$U \cap C = U \text{ and } C = \{\}$
(No medals can be won by *both* countries)

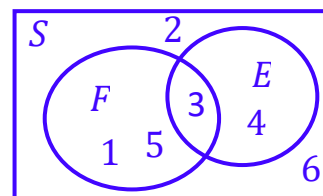
Country	Count
United States (U)	121
China (C)	70
Great Britain (GB)	67
Russia (R)	56
Other (O)	660
Total	974

Example 4.8: Die Rolls

Consider a probability experiment that consists of rolling a single fair die. Let E denote the event a 3 or 4 is observed and F denote the event an odd number is observed.

A. List the outcomes in E and F . Draw a Venn Diagram.

$E = \{3, 4\}$ $F = \{1, 3, 5\}$



B. What is the intersection of E and F ? What is the probability of the intersection of E and F ?

$E \cap F = E \text{ and } F = \{3\}$ $P(E \cap F) = P(E \text{ and } F) = \frac{1}{6}$

4.2.5 THE ADDITION RULE

We can calculate the probability of the union of two events by using the **addition rule**.

Formula 4.2: The Addition Rule

Let A and B be two events. Then the probability of A or B is:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Note that if A and B are mutually exclusive, the addition rule becomes: $P(A \text{ or } B) = P(A) + P(B)$

Why is this the case? *Because $P(A \text{ and } B) = 0$ for mutually exclusive events.*

Example 4.9: More Medal Counts

- A. What is the probability a randomly selected medal is gold?

$$P(G) = \frac{307}{974} = 0.3152$$

- B. What is the probability a randomly selected medal was won by Russia?

$$P(R) = \frac{56}{974} = 0.0575$$

	Gold (G)	Silver (S)	Bronze (B)	Total
United States (U)	46	37	38	121
China (C)	26	28	26	70
Great Britain (GB)	27	23	17	67
Russia (R)	19	18	19	56
Other (O)	189	211	260	660
Total	307	307	360	974

- C. What is the probability that a randomly selected medal was gold **or** won by Russia?

$$P(G \cup R) = P(G) + P(R) - P(G \cap R) = \frac{307}{974} + \frac{56}{974} - \frac{19}{974} = \frac{344}{974} = 0.3532$$

- D. What is the probability that a randomly selected medal was won by the United States **or** Russia? Are these events mutually exclusive?

U and R are mutually exclusive (a participant can't win a medal for both countries).

$$P(U \cup R) = P(U) + P(R) - P(U \cap R) = \frac{121}{974} + \frac{56}{974} - 0 = \frac{177}{974}$$

MY NOTES:

4.2.6 CONDITIONAL PROBABILITY

Conditional probability is the probability of an event B occurring given that event A has occurred.

Formula 4.3: Conditional Probability

The **conditional probability** of B given A is denoted:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Example 4.10: More Medal Counts (continued)

	Gold (G)	Silver (S)	Bronze (B)	Total
United States (U)	46	37	38	121
China (C)	26	28	26	70
Great Britain (GB)	27	23	17	67
Russia (R)	19	18	19	56
Other (O)	189	211	260	660
Total	307	307	360	974

- A. Given that a randomly selected medal is gold, what is the probability that the medal was won by the United States?

$$P(U|G) = \frac{P(U \cap G)}{P(G)} = \frac{46/974}{307/974} = \frac{46}{307} = 0.1498$$

- B. Given that a randomly selected medal was won by the United States, what is the probability that it is gold?

$$P(G|U) = \frac{P(U \cap G)}{P(U)} = \frac{46/974}{121/974} = \frac{46}{121} = 0.3802$$

- C. What is the probability that a medal won by China is bronze?

$$P(B|C) = \frac{P(B \cap C)}{P(C)} = \frac{26/974}{70/974} = \frac{26}{70} = 0.3714$$

4.2.7 INDEPENDENT EVENTS

Events A and B are said to be **independent** events if the occurrence of B does not alter the probability that A has occurred.

This means that $P(A|B) = P(A)$ and it is also true that $P(B|A) = P(B)$.

If two events are not independent, then we say they are **dependent**.

A couple of examples of independent events that you may be familiar with are:

- Coin flips
- Die tosses

Example 4.11: More Medal Counts (continued)

	Gold (G)	Silver (S)	Bronze (B)	Total
United States (U)	46	37	38	121
China (C)	26	28	26	70
Great Britain (GB)	27	23	17	67
Russia (R)	19	18	19	56
Other (O)	189	211	260	660
Total	307	307	360	974

Is $P(G|U) = P(G)$ or $P(U|G) = P(U)$? What can this tell us about the independence of U and G ?

$$P(G|U) = \frac{46}{121} \neq P(G) = \frac{307}{974}$$

Alternatively,

$$P(U|G) = \frac{46}{307} \neq P(U) = \frac{121}{974}$$

So U and G are NOT independent. (You only need to use one of these to test whether two events are independent.)

4.2.8 MULTIPLICATION RULE

In table form, it is easy to find $P(A \cap B)$. However, when we do not have data set up in a nice table, the formula for conditional probability (Formula 4.3) can be rearranged to find $P(A \cap B)$. The way we retrieve the $P(A \cap B)$ by conditional probability is called **multiplication rule**.

Formula 4.4: The Multiplication Rule

Consider two events A and B with probabilities $P(A)$ and $P(B)$, respectively. Suppose that their conditional probabilities $P(A|B)$ and $P(B|A)$ are also given. The **multiplication rule** is:

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$$

When A and B are **independent**, recall that $P(A|B) = P(A)$ and $P(B|A) = P(B)$. Therefore, the multiplication rule becomes:

$$P(A \cap B) = P(A)P(B)$$

This will be another way to test whether A and B are independent. Along with the two ways we learned in 4.4.7, we have three ways of testing for the independence of two events.

For events A and B :

Test 1: If $P(A|B) = P(A)$, then events A and B are independent.

Test 2: If $P(B|A) = P(B)$, then events A and B are independent.

Test 3: If $P(A \cap B) = P(A)P(B)$, then events A and B are independent.

You only need to use one of these tests to show two events are independent. Use the most convenient test based on the information you are given.

Example 4.12: Three Tests for Independence

Use the probabilities listed below to test the independence between events U and G :

$$P(G) = \frac{307}{974} \quad P(U) = \frac{121}{974} \quad P(G|U) = \frac{46}{121} \quad P(U|G) = \frac{46}{307}$$

Test 1: $P(G|U) = P(G)$? $P(G|U) = \frac{46}{121} \neq P(G) = \frac{307}{974}$, so they are not independent.

Test 2: $P(U|G) = P(U)$? $P(U|G) = \frac{46}{307} \neq P(U) = \frac{121}{974}$, so they are not independent.

Test 3: $P(G \cap U) = P(G)P(U)$? (Suppose you only have the 4 probabilities listed above without the entire table. You need to use the **multiplication rule** to find the probability on left-hand side.)

$$P(G \cap U) = P(G|U)P(U) = \frac{46}{121} \cdot \frac{121}{974} = 0.0472 \quad \text{and} \quad P(G)P(U) = \frac{307}{974} \cdot \frac{121}{974} = 0.0392$$

G and U are **not independent** because $P(G \cap U) \neq P(G)P(U)$.

4.3 APPLICATIONS OF PROBABILITY RULES – SEQUENTIAL EVENTS

When trying to find the probability of sequential events (events that happen one after another), the probability rules that you have learned thus far can be useful.

Example 4.13: More Medal Counts (continued)

	Gold (G)	Silver (S)	Bronze (B)	Total
United States (U)	46	37	38	121
China (C)	26	28	26	70
Great Britain (GB)	27	23	17	67
Russia (R)	19	18	19	56
Other (O)	189	211	260	660
Total	307	307	360	974

- A. If two medals are randomly selected **without** replacement, what is the probability that the first selected is gold and the second selected is silver?

G_1 = event that the first medal selected is gold

S_2 = event that the second medal selected is silver

$$P(G_1 S_2) = P(G_1 \text{ and } S_2) = P(G_1)P(S_2|G_1) = \frac{307}{974} \cdot \frac{307}{973} = 0.0995$$

Moreover, these two events are not independent. Think about why.

- B. If two medals are randomly selected without replacement, what is the probability that one is gold and one is silver?

$$\begin{aligned} P(G_1 S_2 \text{ or } S_1 G_2) &= P(G_1 \text{ and } S_2) + P(S_1 \text{ and } G_2) - P(G_1 S_2 S_1 G_2) \\ &= P(G_1)P(S_2|G_1) + P(S_1)P(G_2|S_1) - 0 \\ &= \frac{307}{974} \cdot \frac{307}{973} + \frac{307}{974} \cdot \frac{307}{973} = 0.1987 \end{aligned}$$

- C. If two medals are selected **with** replacement, what is the probability that the first one is bronze and the second one is silver?

B_1 = event that the first medal is bronze

S_2 = event that the second medal is silver

$$P(B_1 \text{ and } S_2) = P(B_1)P(S_2|B_1) = \frac{360}{974} \cdot \frac{307}{974} = P(B_1)P(S_2) = 0.1165$$

Moreover, these two events are independent. Think about why.

MY NOTES:

PRACTICE EXERCISES FOR SECTION 4

1. Consider a couple with two children.

- (a) Determine the sample space of possibilities for the genders of their two children (F = female, M = male) and the probability of each outcome. Assume that for a single birth, there is an equal chance of having a male or a female.

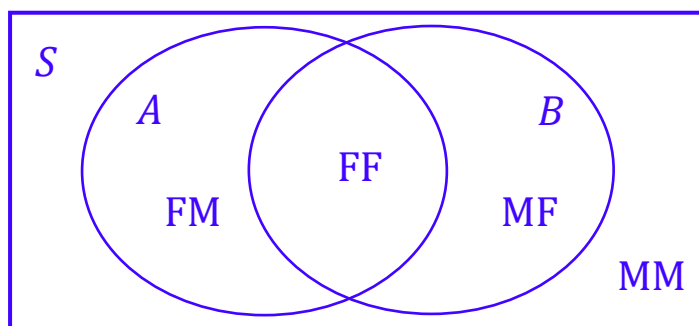
$$S = \{ FF, FM, MF, MM \}$$

$n = 4$ equally-likely outcomes – the probability of each is $1/4$.

- (b) Let A be the event their first child is a girl and B be the event their second child is a girl. Draw a Venn Diagram to illustrate A and B , including all possible outcomes from your sample space in Part (a).

$$A = \{ FF, FM \}$$

$$B = \{ FF, MF \}$$



- (c) List the outcomes in the event “ A and B .” Find $P(A \text{ and } B)$, the probability the couple has two girls.

$$A \text{ and } B = \{ FF \}$$

$$P(A \text{ and } B) = \frac{1}{4}$$

- (d) List the outcomes in the event “ A or B .” Find $P(A \text{ or } B)$, the probability the couple has at least one girl.

$$A \text{ or } B = \{ FM, FF, MF \}$$

$$P(A \text{ or } B) = \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{3}{4}$$

2. The table below displays the results of a class survey about smart phones.

	iPhone (I)	Samsung (S)	Neither (N)	TOTAL
Female (F)	28	1	0	29
Male (M)	31	2	2	35
TOTAL	59	3	2	64

(a) Find the probability of owning an iPhone.

$$P(I) = \frac{59}{64} = 0.9219$$

(b) Describe the event S^C and find its probability.

S^C = event someone does not own a Samsung
(event someone owns an iPhone or neither)

$$P(S^C) = 1 - P(S) = 1 - \frac{3}{64} = 0.9531$$

(c) Find the probability of owning neither phone and being male.

$$P(N \cap M) = \frac{2}{64} = 0.0313$$

(d) Find the probability of being female or owning an iPhone.

$$P(F \cup I) = P(F) + P(I) - P(F \cap I) = \frac{29}{64} + \frac{59}{64} - \frac{28}{64} = 0.9375$$

(e) Given that someone owns a Samsung, what is the probability that they are male?

$$P(M|S) = \frac{P(M \cap S)}{P(S)} = \frac{2/64}{3/64} = \frac{2}{3} = 0.6667$$

(f) In this example, are being female and having a Samsung phone independent?

$$P(F|S) = \frac{1}{3} \neq P(F) = \frac{29}{64} \Rightarrow F \text{ and } S \text{ are not independent.}$$

SECTION 5: DISCRETE RANDOM VARIABLES

(HAWKES SECTIONS 6.1 – 6.3 AND 6.5)

Having learned the basics of probability in the last section, the following three sections take the next step in learning how probability provides the basis for making statistical inferences about populations using samples. We do this by presenting **probability distributions** that play key roles in statistical inference.

The ideas of randomness and uncertainty were introduced in the last section. This section extends those concepts by describing a pattern of randomness for an entire set of outcomes for a random phenomenon.

Recall: What is a *variable*?

A variable is a characteristic recorded about each individual in an observational study or experiment.

- It can be **qualitative** (categorical) or **quantitative** (numerical).
- Quantitative variables can be **discrete** or **continuous**.

A random variable is a function from the set of random outcomes (the sample space S) to the set of numerical values associated with each outcome.

In other words, it is a variable that takes on **numerical values** associated with the random outcomes or events of an experiment, where one (and only one) numerical value is assigned to each outcome.

Example 5.1: Coin Tosses

Suppose you have a coin and you toss it 3 times. You record whether you see Heads or Tails and in what order.

- A. What is the sample space for tossing a coin 3 times? Use H to indicate Heads and T to indicate Tails.

$$S = \{ HHH, HHT, HTH, THH, TTH, THT, HTT, TTT \}$$

Let's attempt to define a random variable X for this set of outcomes for our probability experiment. Let X = the **number** of heads observed in 3 tosses of a coin.

- B. What numerical value will be assigned to each outcome? In other words, what numerical value does X map to each of the outcomes?

$$HHH \Rightarrow 3$$

$$HHT \Rightarrow 2$$

$$HTH \Rightarrow 2$$

$$THH \Rightarrow 2$$

$$TTH \Rightarrow 1$$

$$THT \Rightarrow 1$$

$$HTT \Rightarrow 1$$

$$TTT \Rightarrow 0$$

Notice that for each outcome, there is only one value assigned. It is OK if more than one outcome has the same numerical value. It is NOT OK if a single outcome has more than one value assigned to it.

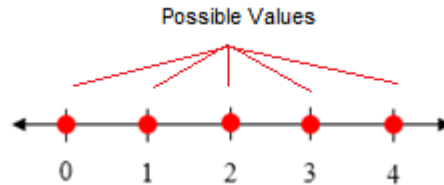
- C. List the set of values that X can be in this coin tossing experiment.

$$X = 0, 1, 2, 3$$

Since X is a characteristic recorded from an experiment, and X maps each outcome of the experiment to only one numerical value, X is a random variable.

5.1 TYPES OF RANDOM VARIABLES

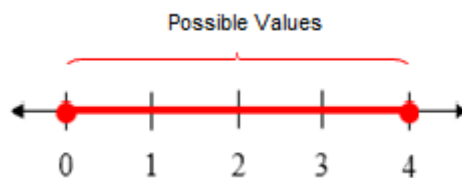
A discrete **random variable** is a random variable that has either a *finite* number of possible values or a *countable* number of values. The term **countable** refers to values that we can count, even if there are an infinite number of possibilities. In other words, this type of random variable cannot take on every possible value in an interval of numbers (there are gaps).



The following are examples of discrete random variables:

- The number of heads observed in 3 tosses of a coin: $X = 0, 1, 2, 3$ (*finite*)
- The number of football games that Clemson will win next year: $X = 0, 1, 2, \dots, 15$ (*finite*)
- The number of car accidents in a 100 mile stretch of highway: $X = 0, 1, 2, 3, 4, 5 \dots$ (*countable*)
- Suppose you are playing a game and you can lose \$1.00, you can win \$1.50 or you can break even (neither win nor lose). The random variable in this game could be the amount of money won/lost in the game. $X = -1.00, 0.00, 1.50$ (*finite*)

A continuous **random variable** has an infinite number of possible values that are *not* countable. This type of random variable can take on every possible value in an interval of numbers.



The following are examples of continuous random variables:

- The height (in feet) of a randomly selected college student: $X > 0$
- The amount of snowfall (in inches) in Clemson in the month of January: $X \geq 0$

One way to determine what type of random variable you are looking at is to ask: How do I get to the value of this random variable?

- If the random variable can only take specific values (these can be positive or negative), it is discrete.
- If you can count to get the value of a random variable, it is discrete.
- If you measure to get the value of a random variable, it is continuous.

5.2 DISCRETE PROBABILITY DISTRIBUTIONS

The **probability distribution** for a random variable describes how the probabilities are distributed over the values of the random variable. Let's investigate the probability distribution for a **discrete** random variable first. We will do the same thing for continuous random variables in the next section.

A legitimate (valid) probability distribution for a **discrete** random variable should have the following properties. You will notice that these properties are the same as what we learned in Section 4 where we discussed the probability for an event.

Properties of a Discrete Probability Distribution

Let $P(X = x)$ denote the probability that a discrete random variable X is equal to a particular value x . Then...

$$(1) \quad 0 \leq P(X = x) \leq 1$$

$$(2) \quad \sum P(X = x) = 1$$

Note, if there's no ambiguity, e.g. there's only one random variable, we can use $P(x)$ instead of $P(X = x)$.

How do we know what a probability distribution looks like? Here are some good ways. For a **discrete** random variable, its probability distribution can be completely defined by a:

- probability mass function
- cumulative distribution function
- characteristic function (not covered in STAT 3090)

5.2.1 PROBABILITY MASS FUNCTION

The probability mass function (pmf) of a discrete random variable X is given by:

$$f_X(x) = P(X = x) \text{ for all possible values of } x$$

We can see the pmf of a discrete random variable X indicates the possible values that X can take, and calculates the probability that X takes a particular value x .

Note, if there's no ambiguity, we can omit the subscript X , i.e., $f(x)$. Also, since the pmf directly indicates the probabilities in a **discrete** random variable, it's ok to use either notation $f_X(x)$ or $P(X = x)$ for the pmf in a **discrete** case. (It's not correct in a continuous case, which we will introduce later.) If you choose to use $P(X = x)$ or $P(x)$ as the notation for pmf, keep it in mind that it is a function when applied to general x 's, but a probability when we assign x a particular value.

PROBABILITY MASS FUNCTION AS A PROBABILITY DISTRIBUTION TABLE

We need to find some good representations for a pmf, which can clearly display the possible values and corresponding probabilities. Using a table is a good idea.

For a discrete random variable X , its probability distribution table lists all possible values X can take in one column (row), and the corresponding probabilities in another column (row). It looks very similar to a relative frequency table.

Example 5.2: Computer Network Interruptions

The following probability distribution table gives the number of interruptions per day (a discrete random variable denoted as X) in a large computer network and the corresponding probabilities.

X	0	1	2	3	4	5
$f_X(x)$ or $P(X = x)$	0.35	0.25	0.20	0.10	0.05	0.05

- A. Verify that this is a legitimate discrete probability distribution.

(1) $0 \leq P(X = x) \leq 1$ ✓

(2) $\sum P(X = x) = 0.35 + 0.25 + 0.20 + 0.10 + 0.05 + 0.05 = 1$ ✓

It is a legitimate discrete probability distribution.

- B. What is the probability that, on a randomly selected day, the network will experience two interruptions?

$P(X = 2) = 0.20$

- C. What is the probability that, on a randomly selected day, the network will experience at most two interruptions?

Hint: Are $P(X = 0)$, $P(X = 1)$, and $P(X = 2)$ mutually exclusive? Use the addition rule!

$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = 0.35 + 0.25 + 0.20 = 0.80$

- D. Describe the complement of the event that the network experiences at most two interruptions, then calculate the probability of this complement event

The event that the network experiences more than two interruptions. Its probability is calculated by:

$P(X > 2) = 1 - P(X \leq 2) = 1 - 0.80 = 0.20$

OR

$P(X > 2) = P(X \geq 3) = P(X = 3) + P(X = 4) + P(X = 5) = 0.10 + 0.05 + 0.05 = 0.20$

- E. Which of these two events is more likely to occur on any given day?

Event A: At least 2 interruptions

Event B: Less than 1 interruption

Since $P(A) = P(X \geq 2) = 0.20 + 0.10 + 0.05 + 0.05 = 0.40$ and

$P(B) = P(X < 1) = P(X = 0) = 0.35$,

it's more likely to have at least 2 interruptions in a day.

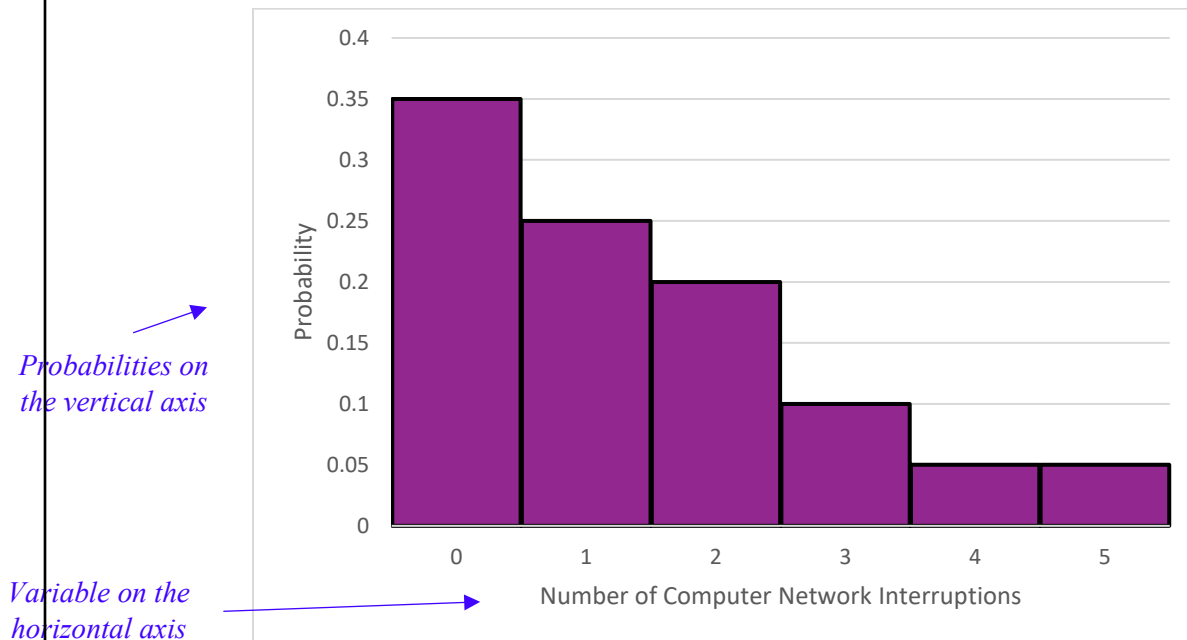
PROBABILITY MASS FUNCTION AS A PROBABILITY HISTOGRAM

The pmf can also be represented by a histogram. Similar to the relationship between relative frequency table and histogram for quantitative data we learned in Section 2, the probability distribution table can also be converted to a probability histogram.

For a discrete random variable X , its probability histogram is a histogram in which the horizontal axis corresponds to the possible values X can take, and the height of each bar represents the corresponding probability that X is equal to that value.

Example 5.3: Computer Network Interruptions (continued)

The following represents the probability histogram for the network interruptions in Example 5.2.



What is the probability that the number of interruptions is between 1 and 3 (inclusive) on a randomly selected day?

$$P(1 \leq X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3) = 0.25 + 0.20 + 0.10 = 0.55$$

PROBABILITY MASS FUNCTION AS A MATHEMATICAL FORMULA

Maybe a real function is the most straightforward way to represent the probability mass “function”, so let’s see what a pmf represented by mathematical formula looks like and how it can help us find the probabilities.

Example 5.4: Draw Cards

Suppose we have 6 cards with numbers 1 or 2 or 3 on each of them. There is one card with number 1 on it, two cards with number 2 on them, and three cards with number 3 on them.

Let X = the number on the randomly selected card.

- A. Define the probability distribution of X by a probability distribution table.

It’s easy to see X is a discrete random variable and $X = 1, 2, 3$. By basic knowledge of probability we learned in Section 4:

X	1	2	3
$f_X(x)$ or $P(X = x)$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$

- B. Find the **mathematical formula** for the pmf of X . Calculate probabilities by the formula. Compare the result with the probability distribution table.

The pmf of X is:

$$f_X(x) = \frac{x}{6}, \quad x = 1, 2, 3$$

Then the probabilities can be calculated by this pmf as:

$$P(X = 1) = f_X(1) = \frac{1}{6}$$

$$P(X = 2) = f_X(2) = \frac{2}{6}$$

$$P(X = 3) = f_X(3) = \frac{3}{6}$$

The results are the same as probability distribution table.

You can see that the pmf represented by a formula generates the same results as the table. Moreover, the pmf represented by a formula is a more mathematical and concise way to define a probability distribution, especially when there are lots of or even an infinite number of possible values, making tables or histograms hard to draw.

If a random variable follows a distribution which is “special” enough, i.e., has some specific characteristics, usually we can find a mathematical formula for its pmf. One example of such distribution is called binomial distribution, which will be introduced later in this section.

5.2.2 CUMULATIVE DISTRIBUTION FUNCTION (FOR DISCRETE RANDOM VARIABLE)

For a discrete random variable X , the function $F_X(x) = P(X \leq x)$ is called cumulative distribution function (cdf) of X . If there's no ambiguity, omit subscript X , i.e., $F(x)$.

The probability distribution of a discrete random variable X can also be defined by its cdf, since it gives the probability that X is less than or equal to a certain value x .

Similar to pmf, the cdf of a discrete random variable can also be represented by a table, histogram, and mathematical formula. We will introduce the table and histogram for a cdf in the following example.

Example 5.5: Computer Network Interruptions (continued)

The following probability distribution table gives the number of interruptions per day (a discrete random variable denoted X) in a large computer network and the corresponding probabilities

X	0	1	2	3	4	5
$f_X(x)$ or $P(X = x)$	0.35	0.25	0.20	0.10	0.05	0.05

A. Make a table for the **cdf** of X .

X	$F(X)$
0	$F(0) = P(X \leq 0) = P(0) = 0.35$
1	$F(1) = P(X \leq 1) = P(0) + P(1) = 0.60$
2	$F(2) = P(X \leq 2) = P(0) + P(1) + P(2) = 0.80$
3	$F(3) = P(X \leq 3) = P(0) + P(1) + P(2) + P(3) = 0.90$
4	$F(4) = P(X \leq 4) = P(0) + P(1) + P(2) + P(3) + P(4) = 0.95$
5	$F(5) = P(X \leq 5) = P(0) + P(1) + P(2) + P(3) + P(4) + P(5) = 1.00$

B. What is the probability that, on a randomly selected day, the network will experience at most one interruption?

We want to find $P(X \leq 1)$. Recall the definition of cdf:

$$F_X(x) = P(X \leq x)$$

Be careful, it's the " $\leq$ " symbol not "<" symbol!

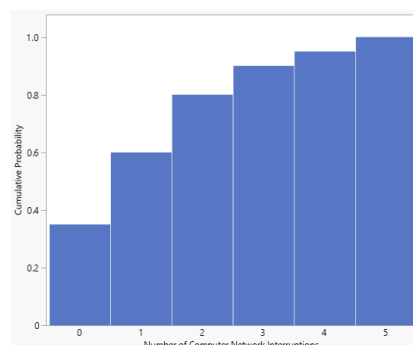
Thus, we can directly find this probability from the cdf table:

$$P(X \leq 1) = F(1) = 0.60$$

C. What is the probability that, on a randomly selected day, the network will experience at least four interruptions?

$$P(X \geq 4) = 1 - P(X < 4) = 1 - P(X \leq 3) = 1 - F(3) = 1 - 0.90 = 0.10$$

D. Draw a histogram of cdf for X .



MY NOTES:

5.3 MEAN AND STANDARD DEVIATION OF A PROBABILITY DISTRIBUTION

To describe a probability distribution more concisely, we can use any of the numerical summaries discussed in Section 2 such as the mean, median, IQR, and standard deviation.

- It is most common to use the [mean](#) and [standard deviation](#) to describe the center and spread of a probability distribution.

Anytime we are working with a probability distribution, we think of the distribution in terms of a [population](#). Recall that numerical summaries for populations are called [parameters](#); we denote the mean of a population as μ and the standard deviation of a population as σ .

Therefore, like numerical summaries of populations, numerical summaries of probability distributions are also referred to as [parameters](#).

For a random variable X , we can denote its mean by μ_X and its standard deviation by σ_X .

5.3.1 MEAN OF A DISCRETE RANDOM VARIABLE

The **mean** of a discrete random variable X , denoted μ_X , can be thought of as the mean outcome of the probability experiment if we repeated the experiment many times. The mean of a random variable X is also called the **expected value** of X , which we denote $E(X)$.

Formula 5.1: Mean (Expected Value) of a Discrete Random Variable

The **mean** or **expected value** of a discrete random variable X is given by the formula:

$$E(X) = \mu_X = \sum_x x \cdot f_X(x)$$

where x is the value of the random variable X , the sum is taken over all possible values of x , and $f_X(x) = P(X = x)$ which is the probability of observing the value x .

The expected value of a random variable is not necessarily an obtainable value for a single observation, but rather what we expect the **average** to be in a **long run** of observations.

Example 5.6: Computer Network Interruptions (continued)

Consider the probability distribution of computer network interruptions in Example 5.2.

X	0	1	2	3	4	5
$f_X(x)$ or $P(X = x)$	0.35	0.25	0.20	0.10	0.05	0.05

If we randomly select a day, how many interruptions can we expect to occur? (In other words, what is the mean or expected value of X ?) Interpret this expected value based the context of the problem.

$$\begin{aligned}\mu_X &= \sum x \cdot f_X(x) = 0(0.35) + 1(0.25) + 2(0.20) + 3(0.10) + 4(0.05) + 5(0.05) \\ &= 0 + 0.25 + 0.40 + 0.30 + 0.20 + 0.25 = 1.4 \text{ interruptions}\end{aligned}$$

Note: We do not round expected value.

Interpretation: If we observed the number of network interruptions for a large number of days (i.e. in the long run), then we would expect the average number of interruptions to be 1.4 interruptions.

5.3.2 STANDARD DEVIATION OF A DISCRETE RANDOM VARIABLE

The **standard deviation** of a discrete random variable X , denoted σ_X , measures the spread of a distribution and describes how far the random variable falls, on average, from the mean μ_X . We will need to first calculate the **variance** of X before finding the standard deviation.

Formula 5.2: Variance of a Discrete Random Variable

The **variance** of a discrete random variable X is given by the formula:

$$\sigma_X^2 = \sum_x (x - \mu_X)^2 \cdot f_X(x) = \sum_x (x^2 \cdot f_X(x)) - \mu_X^2$$

where x is the value of the random variable X , μ_X is the mean of the random variable X , $f_X(x) = P(X = x)$ is the probability of observing the value x , and the sum is taken over all possible values of x .

Formula 5.3: Standard Deviation of a Discrete Random Variable

The **standard deviation** of a random variable X is found by taking the square root of the variance of X :

$$\sigma_X = \sqrt{\sigma_X^2}$$

Example 5.7: Computer Network Interruptions (continued)

Consider the probability distribution of computer network interruptions in Example 5.2.

X	0	1	2	3	4	5
$f_X(x)$ or $P(X = x)$	0.35	0.25	0.20	0.10	0.05	0.05

What is the standard deviation of the number of computer interruptions? Recall that in Example 5.6, we found the mean to be $\mu_X = 1.4$ interruptions.

$$\begin{aligned}
 \sigma_X^2 &= \sum_x (x - \mu_X)^2 \cdot P(X = x) \\
 &= (0 - 1.4)^2(0.35) + (1 - 1.4)^2(0.25) + (2 - 1.4)^2(0.2) + (3 - 1.4)^2(0.1) \\
 &\quad + (4 - 1.4)^2(0.05) + (5 - 1.4)^2(0.05) \\
 &= 2.04 \text{ interruptions}^2
 \end{aligned}$$

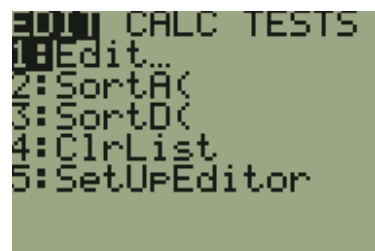
You could also find the variance using the second form of the variance formula:

$$\begin{aligned}
 \sigma_X^2 &= \sum_x (x^2 \cdot P(X = x)) - \mu_X^2 \\
 &= 0^2(0.35) + 1^2(0.25) + 2^2(0.20) + 3^2(0.10) + 4^2(0.05) + 5^2(0.05) - (1.4)^2 \\
 &= 2.04 \text{ interruptions}^2 \\
 \sigma_X &= \sqrt{2.04} = 1.43 \text{ interruptions}
 \end{aligned}$$

Calculator Tutorial 5.1: Finding Mean and Standard Deviation of a Random Variable

You can find the **mean** and **standard deviation** of a random variable using a TI-84 calculator.

- (1) Click the STAT button on your calculator. The first item 1: EDIT... on the EDIT menu will already be selected. Click ENTER.

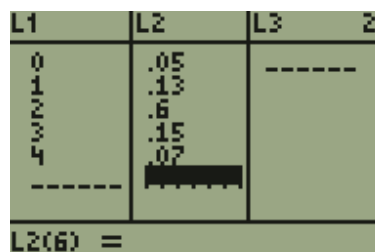


```

STAT TESTS
1: Edit...
2: SortA(
3: SortD(
4: ClrList
5: SetUpEditor

```

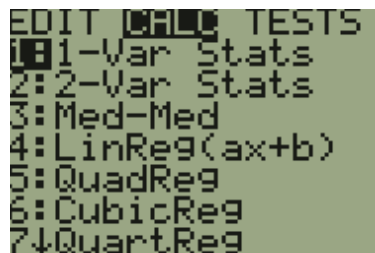
- (2) Enter the values for the random variable (e.g. X) in L1. Enter the corresponding probabilities (e.g. $P(X)$) in L2. Click 2ND + MODE to exit the list editor.



L1	L2	L3	Z
0	.05	-----	
1	.13		
2	.6		
3	.15		
4	.07		
-----	-----		

L2(6) =

- (3) Click the STAT button again. Click the right arrow key once to navigate to the CALC menu. The first item 1: 1-VAR STATS will already be selected. Click ENTER.

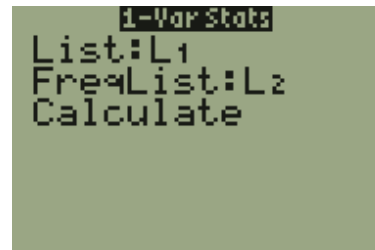


```

EDIT CALC TESTS
1: 1-Var Stats
2: 2-Var Stats
3: Med-Med
4: LinReg(ax+b)
5: QuadReg
6: CubicReg
7: QuartReg

```

- (4) For the LIST option, enter L1 (the list where the values of your random variable are stored). For the FREQLIST option, enter L2 (the list where the corresponding probabilities are stored). Using the down arrow key, navigate to CALCULATE and click ENTER.

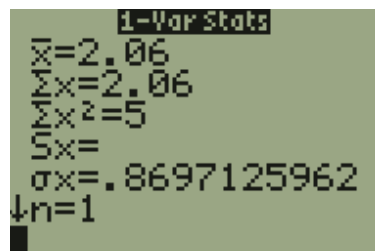


```

1-Var Stats
List: L1
FreqList: L2
Calculate

```

- (5) The **mean** or **expected value** of the random variable will be listed as $\bar{x}$. The **standard deviation** of the random variable will be listed as σ_x . Notice that there is no value for S_x since we do not have a specific sample size.



```

1-Var Stats
x̄=2.06
Σx=2.06
Σx²=5
Sx=
σx=.8697125962
n=1

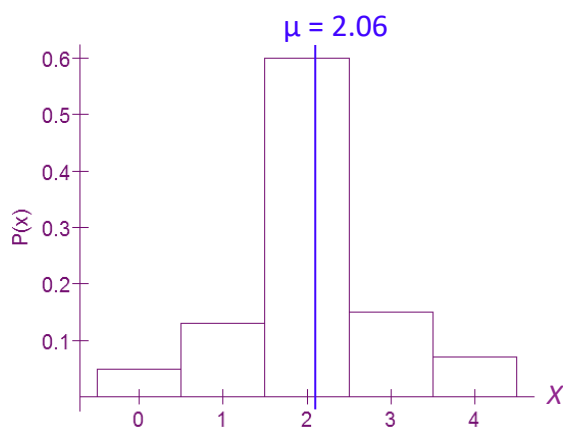
```

Example 5.8: (Practice with calculator) Ideal Number of Children

Let X denote the response of a randomly selected person to the question, “What is the ideal number of children for a family to have?” The probability distribution of X in the United States is shown below according to the results of a General Social Survey.

X	0	1	2	3	4
$f_X(x)$ or $P(X = x)$	0.05	0.13	0.60	0.15	0.07

- A. Sketch a probability histogram for the distribution of X . Label the horizontal and vertical axes.



- B. Calculate the mean of X and mark it on the horizontal axis of your probability histogram. Check your answer using the list function in your calculator.

$$\begin{aligned}\mu_X &= \sum x \cdot f_X(x) \\ &= 0(0.05) + 1(0.13) + 2(0.6) + 3(0.15) + 4(0.07) = 2.06\end{aligned}$$

- C. Interpret the mean, explaining why it does not have to be a whole number even though each possible value for the ideal number of children is a whole number.

Interpretation: If we asked a large number of persons in the U.S. the ideal number of children in a family, we would expect the mean response to be approximately 2.06 children. And the mean of a large amount of whole numbers is not necessarily to be a whole number.

- D. Calculate the standard deviation of X using the list function in your calculator.

$$\sigma_X = 0.87 \text{ children}$$

- E. Interpret the standard deviation of X .

The observations for the ideal number of children deviate from the mean by 0.87 children on average.

MY NOTES:

5.4 THE BINOMIAL DISTRIBUTION

The **binomial distribution** is a specific type of discrete probability distribution whose probability mass function (pmf) is most often represented using a mathematical formula. It applies to probability experiments composed of multiple trials with only two possible outcomes for each individual trials.

5.4.1 CHARACTERISTICS OF A BINOMIAL EXPERIMENT

An experiment is called a binomial experiment if it has the following four properties:

- (1) The experiment consists of a fixed number of n identical trials.
- (2) There are only two possible outcomes in each trial. We denote one outcome by S (for “success”) and the other by F (for “failure”).
- (3) The probability of S remains the same from trial to trial. This probability is denoted p , and the probability of F is $1 - p$.
- (4) The n trials are independent. That is, the result for one trial does not depend on the results of the other trials.

5.4.2 BINOMIAL DISTRIBUTION AND BINOMIAL RANDOM VARIABLE

In a binomial experiment with parameters n and p , n is the number of trials and p is the probability of success. If a random variable X is the count (frequency) of the number of successes in n trials, X is called a binomial random variable.

We can also say X follows binomial distribution with parameters n and p .

Using mathematical notations, $X \sim B(n, p)$.

Note that $P(\text{Success})$ has the same notation as the population proportion in that it uses the lowercase letter p . You can think of $P(\text{Success})$ as the **proportion** of the items in population that has a particular characteristic of interest. Any item in the population with that characteristic is considered a “Success” regardless of whether having that characteristic is a good thing, bad thing, or neutral.

Example 5.9: Identifying a Binomial Experiment

Determine which of the following probability experiments qualify as a **binomial experiment**. For those that are binomial experiments, define the binomial random variable X , identify the number of trials n , identify the probability of success p , and list the possible values of X . For those that are not binomial experiments, specify which characteristic(s) it violates.

- A. You observe the sex of the next 50 children born at a hospital and record the number of girls. We assume the chance of having a boy or a girl in any single birth is 50:50.

Yes, it is a binomial experiment.

Let X = number of girls born in next 50 births, $X = 0, 1, 2, \dots, 50$. $X \sim \mathcal{B}(n = 50, p = 0.5)$

- B. A couple decides to continue having children until their first girl is born. The total number of children the couple has is recorded.

Not binomial – the total number of trials (n) is not fixed.

5.4.3 PMF FOR THE BINOMIAL PROBABILITY DISTRIBUTION

If a discrete random variable X follows binomial distribution, we can use its probability mass function (pmf) represented by a mathematical formula to find the probability that X takes a particular value. We will begin with an example to assist us in understanding the pmf for the binomial probability distribution.

Example 5.10: Deriving the PMF for a Binomial Random Variable

Kyrie Irving's free throw percentage was 93.9% in the 2016 NBA Finals in which the Cleveland Cavaliers beat the Golden State Warriors. Let's assume that Kyrie's free throw percentage is fixed and his attempts are independent.

Suppose Kyrie shoots 3 free throws ($n = 3$) and let X = the number of successful free throws in these 3 throws. Then X follows a binomial distribution (and you think about why). Create a probability distribution to represent X .

X	0	1	2	3
$f_X(x)$ or $P(X = x)$	0.0002	0.0105	0.1614	0.8280

A. What is the probability that Kyrie makes exactly 0 free throws in 3 attempts?

The probability that Kyrie lands a single free throw is $P(S) = p = 0.939$

The probability that Kyrie misses a single free throw is $P(F) = 1 - p = 1 - 0.939$

There is one way this can happen: FFF

Since all attempts are independent, by multiplication rule:

$$P(X = 0) = P(F) \cdot P(F) \cdot P(F) = (1 - 0.939)^3 = 0.0002$$

$$P(X = 0) = p^0(1 - p)^3$$

B. What is the probability that Kyrie makes exactly 1 free throw in 3 attempts?

There are three ways for this to occur: SFF, FSF, FFS. Each way has an equal probability of occurring.

$$P(X = 1) = 3(0.939)(1 - 0.939)^2 = 0.0105$$

$$P(X = 1) = 3p^1(1 - p)^2$$

C. What is the probability that Kyrie makes exactly 2 free throws in 3 attempts?

There are three ways for this to occur: SSF, SFS, FSS. Each way has an equal probability of occurring.

$$P(X = 2) = 3(0.939)^2(1 - 0.939) = 0.1614$$

$$P(X = 2) = 3p^2(1 - p)^1$$

D. What is the probability that Kyrie makes exactly 3 free throws in 3 attempts?

There is one way this can happen: SSS

$$P(X = 3) = (0.939)^3 = 0.8280$$

$$P(X = 3) = p^3(1 - p)^0$$

E. Did we get a valid probability distribution?

Yes, since $0 \leq P(X = x) \leq 1$ for all x

and

$$P(X = 0) + \dots + P(X = 3) \approx 1 \text{ (the difference from one is simply round off error)}$$

In Example 5.10, we begin to see a pattern emerge. In addition to considering the probability of a single success or failure, when calculating binomial probabilities, we also must consider the number of ways something can occur, i.e., the number of outcomes in a compound event. We do this using the **binomial coefficient**.

Formula 5.4: The Binomial Coefficient

For a binomial experiment with n independent trials and a probability of success p , the **binomial coefficient** represents how many ways an event can occur. It is read as “ n choose x ” and is calculated using the formula:

$${}_nC_x = \frac{n!}{x!(n-x)!} = \binom{n}{x}$$

Recall that $n! = 1(2)(3)\dots(n-1)(n)$ and $0! = 1$.

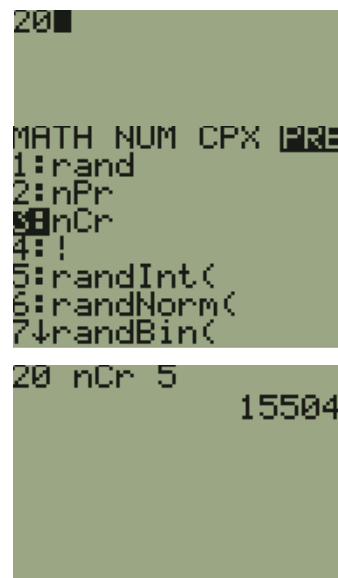
Calculator Tutorial 5.2: The Binomial Coefficient

You can calculate the **binomial coefficient** in your calculator. As an example, the following instructions show how to find ${}_{20}C_5$, read “twenty choose five.”

- (1) Type the value of n on the main calculator screen.
In this example, $n = 20$.
- (2) Click the MATH button on the calculator and use the arrow keys to navigate to the PRB menu. Then use the down arrow key to select the third item 3: nCr. Click ENTER.

You will now be back on the main calculator screen

with nCr showing after the value of n . Type the value of x and click ENTER. In this example, $x = 5$.



The binomial coefficient in the calculator tutorial is ${}_{20}C_5 = 15,504$. As you can see, this number can be quite large as your number of trials n gets larger.

The pattern we see in Example 5.10 leads us to the **pmf** for the Binomial Distribution, which allows us to calculate the probability of a certain number of successes in a binomial experiment.

Formula 5.5: The Binomial Probability Mass Function

Let X be a **binomial random variable** with n trials and probability of success p , i.e., $X \sim \mathcal{B}(n, p)$. Then the probability of observing x successes out of n trials is:

$$f_X(x) = P(X = x) = {}_nC_x p^x (1 - p)^{n-x}, \quad x = 0, 1, \dots, n$$

Recall that the values of n and p are [parameters](#). In addition to being numerical summaries about our population, we say that these parameters “completely characterize” the binomial distribution. In other words, if we have the parameters n and p for a binomial random variable, then we have everything we need to know to calculate binomial probabilities.

Example 5.11: Sneeze Etiquette

According to a study done by Nick Wilson of Otago University Wellington, the probability a randomly selected individual will cover his or her mouth when sneezing is 0.73. Suppose you sit on a bench in a mall and observe people's habits as they sneeze. What is the probability that, among 5 randomly observed individuals, exactly 3 of them cover their mouths when sneezing?

Let X = number of people who cover their mouth when sneezing in 5 randomly observed individuals.

X follows binomial distribution with $n = 5$ and $p = 0.73$: $X \sim \mathcal{B}(5, 0.73)$

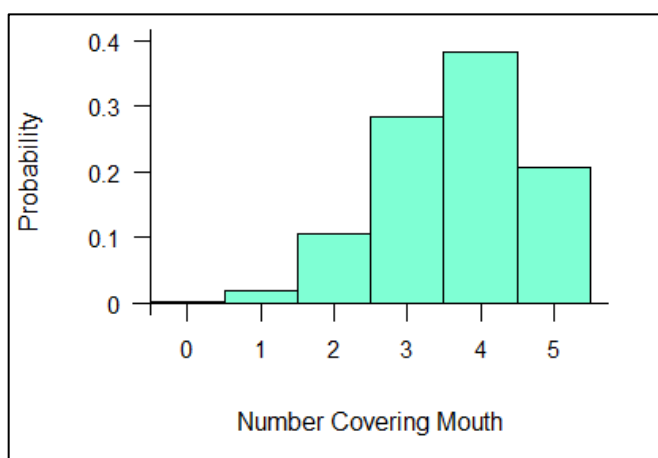
$$P(X = 3) = {}_5C_3(0.73)^3(0.27)^2 = \frac{5!}{3!2!}(0.73)^3(0.27)^2 = 0.2836$$

Recall that the binomial coefficient ${}_5C_3 = 10$ tells us that there are ten ways to have three successes and 2 failures.

5.4.3 MEAN AND STANDARD DEVIATION FOR A BINOMIAL RANDOM VARIABLE

Since a binomial random variable is just a special case of a discrete random variable, binomial probability distributions can also be in the form of a table or a graph. A table of the probability distribution and probability histogram for Example 5.11 are given below.

X	$f_X(x)$ or $P(X = x)$
0	${}_5C_0(.73)^0(.27)^5 = .0014$
1	${}_5C_1(.73)^1(.27)^4 = .0194$
2	${}_5C_2(.73)^2(.27)^3 = .1049$
3	${}_5C_3(.73)^3(.27)^2 = .2836$
4	${}_5C_4(.73)^4(.27)^1 = .3834$
5	${}_5C_5(.73)^5(.27)^0 = .2073$



Example 5.12: Sneezing Etiquette (continued)

Find the **expected number** of people who cover their mouth when sneezing in a random sample of size 5. (Remember this is simply a discrete random variable)

$$E(X) = \sum x \cdot f_X(x) = 0(0.0014) + 1(0.0194) + \cdots + 5(0.2073) = 3.65 \text{ people}$$

We have seen that, for a binomial distribution, there's a pmf with special formula to calculate the probabilities. What makes things even better is, once we know a random variable follows a binomial distribution, there are some special and simple formulas to find its mean μ and standard deviation σ based on the parameters, rather than using the formula that applies to a discrete random variable in general.

Formula 5.6: Binomial Mean

For a binomial random variable X with n trials and probability of success p , the **expected value** of X is

$$\mu_X = np$$

Formula 5.7: Binomial Standard Deviation

For a binomial random variable X with n trials and probability of success p , the **standard deviation** of X is

$$\sigma_X = \sqrt{np(1-p)}$$

Example 5.13: Sneezing Etiquette (continued)

Find the **mean** and **standard deviation** in Example 5.11 using Formulas 5.6 and 5.7. Does your expected value match what you found in Example 5.12?

$$X \sim \mathcal{B}(n = 5, p = 0.73), \quad \mu_X = np = 5(0.73) = 3.65 \text{ people}$$

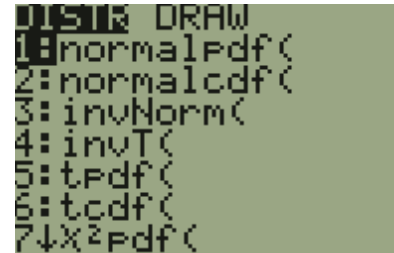
$$\sigma_X = \sqrt{np(1-p)} = \sqrt{5(0.73)(0.27)} = 0.9927 \text{ people}$$

Calculator Tutorial 5.3: Binomial Probabilities

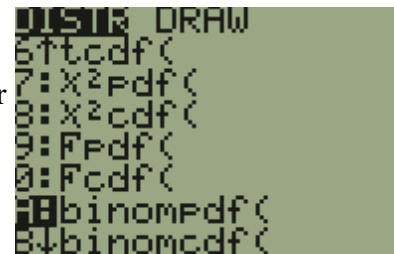
The formula for finding binomial probabilities is practical only when n is small or the event of interest consists of only a few outcomes. Otherwise, it is more practical to use technology to find binomial probabilities.

To find the probability of **exactly** x successes out of n :

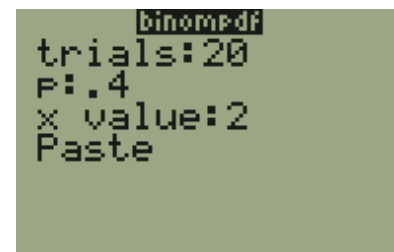
- (1) Click 2ND + VARS to navigate to the distribution menu in your calculator.



- (2) Use the down arrow key to select item A: binompdf(. Click ENTER. (It is called pdf in calculator, which is a term for continuous random variables we will learn in next section. But keep it in mind, these programs were not written by statisticians, and the correct name here is binompmf)



- (3) For trials, enter the value of n . For p , enter the probability of success p on a single trial. For x value, enter the desired number of successes for the experiment. Use the down arrow key to navigate to PASTE and click ENTER.



- (4) The binomial pmf is now in the main screen of your calculator. Click ENTER again. The value displayed is $P(X = x)$ for the x you specified.

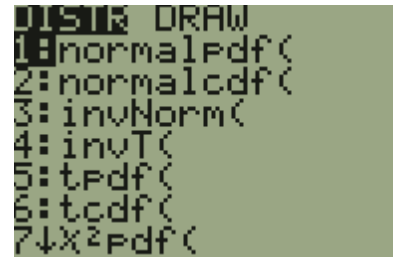


This example computes $P(X = 2)$ for a binomial random variable X .

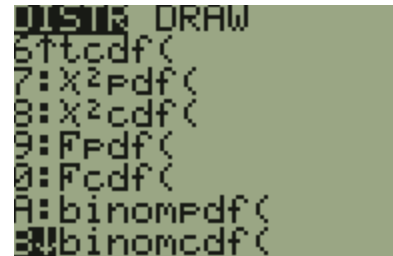
(continued on next page)

To find the probability of **at most** x successes (in other words, the probability that the number of successes is less than or equal to x)...

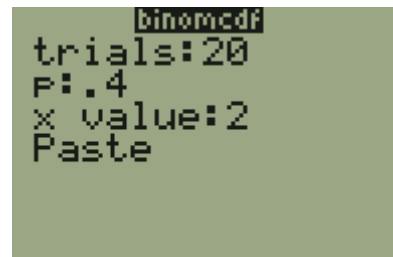
- (1) Click 2ND + VARS to navigate to the distribution menu in your calculator.



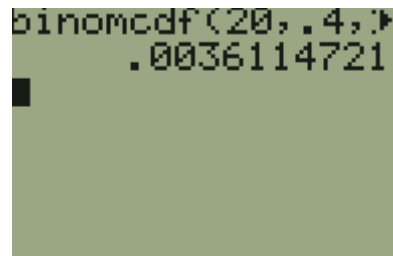
- (2) Use the down arrow key to select item B: binomcdf(. Click ENTER.



- (3) For trials, enter the value of n . For p , enter the probability of success p on a single trial. For x value, enter the desired maximum number of successes for the experiment. Use the down arrow key to navigate to PASTE and click ENTER.



- (4) The binomial cdf is now in the main screen of your calculator. Click ENTER again. The value displayed is $P(X \leq x)$ for the x you specified.



This example computes $P(X \leq 2)$ for a binomial random variable X .

Example 5.14: (Practice with Calculator) Restaurant Menu Planning

You operate a restaurant. You read that a sample survey by the National Restaurant Association shows that 40% of adults are committed to eating nutritious food when eating away from home. To help you plan your menu, you decide to conduct a sample survey in your own area. You will use random digit dialing to contact a random sample of 20 households by telephone.

Answer the following questions assuming the national result holds in your area.

- A. In this sample of size 20, what are the **expected number** and **standard deviation** of households who seek nutritious food when eating out?

Let X = number of households who seek nutritious food when eating out.

$$X \sim \mathcal{B}(n = 20, p = 0.4)$$

$$E(X) = \mu_X = np = 20(0.4) = 8 \text{ households}$$

$$\sigma_X = \sqrt{np(1-p)} = \sqrt{20(0.4)(0.6)} = 2.19 \text{ households}$$

- B. What is the probability that exactly 11 households seek nutritious foods when eating out in a sample of size 20? Fill in the formula and use the binompdf function in the calculator to compute the probability.

$$P(X = 11) = {}_{20}C_{11}(0.4)^{11}(0.6)^9 = 0.0710$$

$$\text{binompdf}(\text{trials} = 20, p = 0.4, x = 11) = 0.0710$$

- C. What is the probability that **at most** 2 seek nutritious foods when eating out in these 20 households? Fill in the formula, then use binomcdf function in the calculator to compute the probability.

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$$

$$= {}_{20}C_0(0.4)^0(0.6)^{20} + {}_{20}C_1(0.4)^1(0.6)^{19} + {}_{20}C_2(0.4)^2(0.6)^{18}$$

$$= 0 + 0.0005 + 0.0031$$

$$= 0.0036$$

Note, $P(X \leq 2) = F_X(2)$, where $F_X(x)$ is the cdf of X .

$$\text{binomcdf}(\text{trials} = 20, p = 0.4, x = 2) = 0.0036$$

- D. What is the probability that **at least** 3 seek nutritious foods when eating out in these 20 households?

$$P(X \geq 3) = 1 - P(X < 3) = 1 - P(X \leq 2) = 1 - 0.0036 = 0.9964$$

- E. What is the probability that the number of people seeking nutritious foods is between 7 and 14 (inclusive) in these 20 households?

$$P(7 \leq X \leq 14) = P(X \leq 14) - P(X \leq 6) = F_X(14) - F_X(6)$$

Note, the subtrahend is $P(X \leq 6)$, or equivalently $P(X < 7)$, in this discrete case, rather than $P(X \leq 7)$. Think about why.

By calculator:

$$F_X(14) = \text{binomcdf}(\text{trials} = 20, p = 0.4, x = 14) = 0.9984$$

$$F_X(6) = \text{binomcdf}(\text{trials} = 20, p = 0.4, x = 6) = 0.2500$$

$$\text{Therefore, } P(7 \leq X \leq 14) = 0.9984 - 0.2500 = 0.7384$$

MY NOTES:

PRACTICE EXERCISES FOR SECTION 5

1. The U.S. Census Bureau collects demographics concerning the number of people in families per household. Assume that the distribution of the number of people per household is contained in the following table, where X = number of people in families per household.

X	2	3	4	5	6	7
$P(X = x)$	0.27	0.25	0.28	0.13	0.04	0.03

- (a) Verify that the distribution above is a valid probability distribution.

$0 \leq P(X = x) \leq 1$ for all x and $\sum P(X = x) = 1$, so it is a valid probability distribution.

- (b) Calculate the expected number of people per household in the U.S.

$$\begin{aligned}\mu_X &= E(X) = \sum x \cdot f_X(x) \\ &= 2(0.27) + 3(0.25) + 4(0.28) + 5(0.13) + 6(0.04) + 7(0.03) \\ &= 3.51 \text{ people}\end{aligned}$$

- (c) Calculate the variance and standard deviation of people per household.

$$\begin{aligned}\sigma_X^2 &= [\sum x^2 \cdot f_X(x)] - \mu_X^2 \\ &= 2^2(.27) + 3^2(.25) + 4^2(.28) + 5^2(.13) + 6^2(.04) + 7^2(.03) - 3.51^2 \\ &= 1.6199 \text{ people}^2\end{aligned}$$

$$\sigma_X = \sqrt{1.6199} = 1.27 \text{ people}$$

- (d) Find the probability that a household has exactly three people.

$$P(X = 3) = 0.25$$

- (e) Find the probability that a household has at most three people.

$$P(X \leq 3) = 0.27 + 0.25 = 0.52$$

- (f) Find the probability that a household has four or more people.

$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - 0.52 = 0.48$$

2. A box contains four slips of paper marked 1, 2, 3, and 4. Two slips are selected without replacement. The possible outcomes can be listed in the following sample space:

$$S = \{(1,2), (1,3), (1,4), (2,1), (2,3), (2,4), (3,1), (3,2), (3,4), (4,1), (4,2), (4,3)\}$$

For each of the following discrete random variables, list the possible values along with the associated probabilities.

- (a) X = sum of the two numbers

X	3	4	5	6	7
$P(X = x)$	$\frac{2}{12}$	$\frac{2}{12}$	$\frac{4}{12}$	$\frac{2}{12}$	$\frac{2}{12}$

- (b) Z = number of slips selected that show an even number

Z	0	1	2
$P(Z = z)$	$\frac{2}{12}$	$\frac{8}{12}$	$\frac{2}{12}$

- (c) W = number of slips selected that show a 4

W	0	1
$P(W = w)$	$\frac{6}{12}$	$\frac{6}{12}$

3. Determine whether or not each of the following are valid probability distributions. If they are not, give the characteristic which is not satisfied by the distribution.

(a)

X	1	2	3
$P(X = x)$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{3}$

This is not a valid probability distribution because $\sum P(X = x) = \frac{4}{3} > 1$.

(b)

Y	0	1	2	3	4
$P(Y = y)$	0.232	0.441	-0.119	0.172	0.036

This is not a valid probability distribution because $P(Y = 2) = -0.119 < 0$.

4. For each of the following, determine whether the random variable described is binomial. If it is binomial, state its parameters and possible values. If it is not binomial, state why.

- (a) You have a group of 10 people, 4 with blue eyes and 6 with brown eyes. You need to form a committee of size 2 from this group. You will select 2 people at random to be on the committee. Let X = the number of people on the committee with blue eyes.

Not binomial since the trials are not independent. Once you select a person for the committee, that person cannot be selected again. Notice also the value of p , the probability of selecting a person with blue eyes, does not remain the same from the first person selected to the second person selected.

- (b) According to data collected by the National Parks Service, a certain park reports that you have a 34% chance of seeing a bear while on a hike. You interview 30 solo hikers and ask them if they saw a bear. Let B = the number of hikers who see a bear.

B is binomial with $n = 30$, $p = 0.34$, and $B = 0, 1, 2, \dots, 30$

5. You browse a menu that has 16 possible appetizers to choose from. You and your friends want to order the Sampler Platter, which allows you to choose three appetizers. How many possible combinations of appetizers are there? (Hint: Use the binomial coefficient.)

Number of ways to choose 3 appetizers out of 16: ${}_{16}C_3 = \frac{16!}{3!13!} = 560$

6. Clarinex-D is a medication whose purpose is to reduce the symptoms associated with a variety of allergies. In clinical trials of Clarinex-D, 5% of patients in the study experience insomnia as a side effect. A random sample of 25 Clarinex-D users is obtained, and the number of patients X who experience insomnia is recorded.

- (a) Find the probability that exactly 3 patients experience insomnia as a side effect.

$$P(X = 3) = {}_{25}C_3 (.05)^3 (.95)^{22} = 0.0930$$

$\text{binompdf}(\text{trials} = 25, p = 0.05, x = 3)$

- (b) Find the probability that 5 or fewer experiences insomnia as a side effect.

$$P(X \leq 5) = 0.9988$$

$\text{binomcdf}(\text{trials} = 25, p = 0.05, x = 5)$

SECTION 6: CONTINUOUS RANDOM VARIABLES

We now shift our focus to **continuous random variables**. These random variables can take on any value within an *interval* of numbers.

6.1 CONTINUOUS PROBABILITY DISTRIBUTION

The probability distribution for a continuous random variable can be completely defined by a:

- probability density function
- cumulative distribution function
- characteristic function (not covered in STAT 3090)

6.1.1 PROBABILITY DENSITY FUNCTION (PDF)

A continuous random variable does not have a probability mass function. Instead, it has a probability density function (pdf), typically denoted $f_X(x)$ or $f(x)$. The pdf is a function of the continuous random variable X . The graph of the pdf will show us the shape of the random variable's probability distribution. The area under the curve of the pdf, $f(x)$, represents the probability that our random variable falls between two values a and b . Thus, we will be using integration, (or possibly geometry) to calculate probabilities for continuous random variables. A pdf is usually represented by a mathematical formula or a continuous curve, rather than a table or histogram. The values that the continuous random variable X can take are known as the support of X .

Properties of a PDF for a Continuous Random Variable

Let $f_X(x)$ denote the **probability density function** (pdf) of a continuous random variable X . This probability distribution of X has the following properties:

- (1) $f_X(x) \geq 0$
- (2) Areas under the curve of $f_X(x)$ correspond to probabilities. The areas can be obtained by using integration, i.e.,

$$P(a \leq X \leq b) = \text{Area under curve } f_X(x) \text{ between } a \text{ and } b = \int_a^b f_X(x) dx$$

- (3) The probability associated with one particular outcome is always **zero**. That is...

$P(X = x) = 0$, which further indicates that (try this out in your calculator)

$$P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = P(a < X < b) = \int_a^b f_X(x) dx.$$

- (4) The interval containing all possible values of X , i.e., over the entire support of X , has probability equal to one, so the total area under the curve equals **one**.

$$\int_{-\infty}^{\infty} f_X(x) dx = 1$$

6.1.2 CUMULATIVE DISTRIBUTION FUNCTION (FOR CONTINUOUS RANDOM VARIABLE)

For a continuous random variable X , the function $F_X(x) = P(X \leq x)$ is called the cumulative distribution function (cdf) of X . If there's no ambiguity, omit subscript X , i.e., $F(x)$.

The cdf for a continuous random variable is derived from the pdf and has some important properties.

Properties of a CDF for a continuous random variable:

Let $F_X(x)$ denote the **cumulative distribution function** (cdf) of a continuous random variable X . The cdf of X has the following properties:

- (1) $F_X(x) = P(X \leq x) = \int_{-\infty}^x f_X(t) dt$
- (2) $F_X(b) - F_X(a) = P(a \leq X \leq b) = \int_a^b f_X(x) dx$

Both discrete and continuous random variables share the same definition of cdf, that is $F_X(x) = P(X \leq x)$. A main difference is that the cdf of a discrete distribution is obtained by summation using the pmf, and the cdf of a continuous distribution is obtained by integration of the pdf.

6.2 NUMERICAL SUMMARIES OF CONTINUOUS PROBABILITY DISTRIBUTIONS

Similar to discrete probability distributions, we can calculate the **mean**, **variance**, and **standard deviation** of continuous random variables by replacing the summation with integration

6.2.1 MEAN OF CONTINUOUS PROBABILITY DISTRIBUTIONS

Formula 6.1: Mean of a Continuous Random Variable

Let X be a continuous random variable with probability distribution function $f_X(x)$. The **mean** or **expected value** of X is given by the formula:

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x) dx$$

where x represents the value of the random variable. Integration is over the support of X , i.e., the *lower* limit of integration is the *smallest possible value* and the *upper* limit is the *largest possible value* for the random variable.

As with discrete random variables, the expected value represents what we expect to be the average value of the variable in the long run.

6.2.2 STANDARD DEVIATION OF CONTINUOUS PROBABILITY DISTRIBUTIONS

The **standard deviation** of a continuous random variable X measures the spread of a distribution and describes how far the random variable falls, on average, from the mean.

Formula 6.2: Variance of a Continuous Random Variable

The **variance** of a continuous random variable X with pdf $f_X(x)$ and mean μ_X is given by:

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 \cdot f_X(x) dx$$

The **limits of integration** will depend on the support of X .

Formula 6.3: Standard Deviation of a Continuous Random Variable

The **standard deviation** of a continuous random variable X with pdf $f_X(x)$ and variance σ_X^2 is:

$$\sigma_X = \sqrt{\sigma_X^2}$$

Example 6.1: Bus Waiting Times

The Blacksburg Bus Company is interested in the wait time of its passengers between 8:00 AM and 8:00 PM. Suppose that the minimum wait time is 1 minute and the maximum wait time is 15 minutes.

Let X be the wait time and $f_X(x) = \begin{cases} \frac{1}{14} & 1 \leq x \leq 15 \\ 0 & \text{otherwise} \end{cases}$

A. What values can X take (what is the support of X)?

X can take any value between $\pm\infty$, but we are most interested in the values on support, i.e., $1 \leq x \leq 15$

B. Is the total area under the curve equal to 1? (If the area under the curve is **not** equal to 1, then this is not a legitimate probability distribution.)

$$\int_{-\infty}^{\infty} f_X(x) dx = \int_{-\infty}^1 0 dx + \int_1^{15} \frac{1}{14} dx + \int_{15}^{\infty} 0 dx = 1 ?$$

$$\int_1^{15} \frac{1}{14} dx = \left[\frac{1}{14} x \right]_1^{15} = \frac{15}{14} - \frac{1}{14} = \frac{14}{14} = 1$$

So yes, it is a legitimate probability distribution.

C. What is the probability that a passenger will wait between 10 and 15 minutes?

$$P(10 < X < 15) = \int_{10}^{15} \left(\frac{1}{14} \right) dx = 0.3571$$

D. What is the expected wait time for passengers between the hours of 8 AM and 8 PM?

$$E(X) = \mu_X = \int_1^{15} x \cdot \frac{1}{14} dx = 8 \text{ minutes}$$

E. What is the standard deviation of the wait time between 8 AM and 8 PM?

$$\sigma_X = \sqrt{\int_1^{15} (x - 8)^2 \left(\frac{1}{14} \right) dx} = 4.04 \text{ minutes}$$

Calculator Tutorial 6.1: Evaluating Definite Integrals

You can evaluate **definite integrals** to find probabilities using a TI-83 or TI-84 graphing calculator. (Screenshots are not included for TI-83 instructions.)

- (1) Click the MATH button on your calculator to bring up the MATH menu. Use the down arrow key to navigate to item 9: fnInt(.
- (2) Click ENTER. If you are using a TI-84, the syntax for a definite integral will appear on your screen.

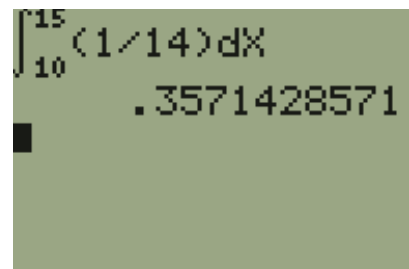
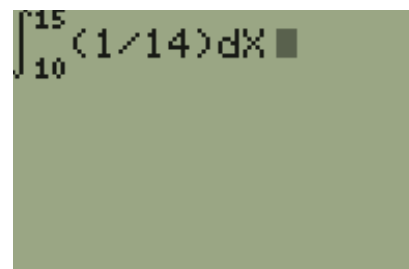
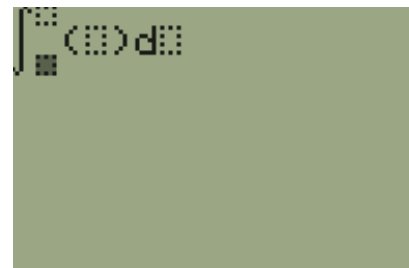
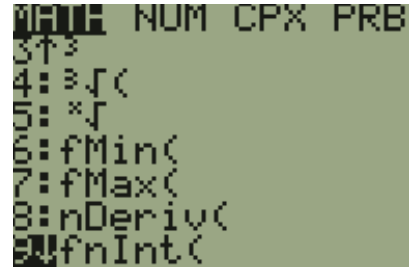
If you are using a TI-83, your screen will display fnInt(on the screen.

- (3) On a TI-84, you will enter the lower and upper limits of integration, the pdf $f(x)$, and the variable X in the proper boxes. You can type the variable using the X, T, θ , n button on your calculator.

If you are using a TI-83, you will type $f(x)$, X , the lower limit, and the upper limit (in this order) immediately after fnInt(, separated by the comma button on your calculator.

- (4) Click ENTER to calculate the value of the definite integral.

This example calculates $P(10 \leq X \leq 15)$ for the random variable described in Example 6.1.



MY NOTES:

6.3 THE NORMAL DISTRIBUTION

6.3.1 INTRODUCTION TO THE NORMAL DISTRIBUTION

One of the most important continuous distributions is the **normal distribution**. The normal distribution is a specific continuous distribution defined by a pdf with parameters mean μ and standard deviation σ :

Formula 6.4: The Normal Probability Density Function

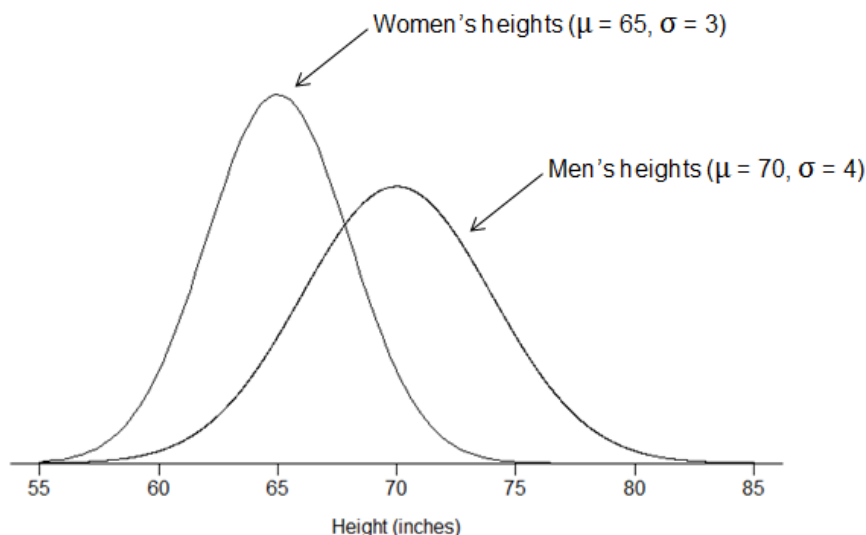
Let X be a normal random variable with mean μ and standard deviation σ . Then its probability density function (pdf) can be represented by the formula:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

We can also say the random variable X follows normal distribution with parameters μ and σ , or simply, $X \sim \mathcal{N}(\mu, \sigma)$.

Normal distributions have several distinctive properties:

- (1) A particular normal distribution is specified (completely characterized) by giving its mean μ and standard deviation σ .
- (2) All normal distributions have the **same overall shape**: symmetric, unimodal and bell shaped. For example, consider the graph of the two normal distributions below for the heights of women and men. The mean μ determines the location of the center of the distribution. Note that the standard deviation σ is relevant to the width of the bell shape.



- (3) All normal random variables X can take on values between $-\infty$ and $+\infty$.
- (4) The **area under the normal curve** for any interval of values for a normal random variable X represents either:
- The proportion of the population with the characteristic described by the interval of values, or
 - The probability that a randomly selected individual from the population will have the characteristic described by the interval of values.

Suppose $X \sim \mathcal{N}(\mu, \sigma)$, and we want to find the probability that X is in range (a, b) , i.e., $P(a < X < b)$. Using the second property of a pdf, which we learned previously, this probability is the area under the normal pdf curve between a and b , which can be obtained by integration. That is:

$$P(a < X < b) = \int_a^b f_X(x) dx = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

The idea is simple, but it's hard to solve such an integral by hand. (It's even hard to type this normal pdf into calculator!) Thus, we need to use some techniques to solve this problem.

Since many techniques are based on the **standard** normal distribution, let's get familiar with it.

6.3.2 STANDARD NORMAL DISTRIBUTION

A normal distribution with a mean of 0 and a standard deviation of 1 is called the standard normal distribution.

Formula 6.5: The Standard Normal Probability Density Function

Let Z be a normal random variable with mean 0 and standard deviation 1. Then its probability density function (pdf) is called the standard normal pdf, which is represented by the formula:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-0)^2}{2(1)^2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad -\infty < z < \infty$$

Note, the capital letter Z is usually reserved for standard normal random variable in statistics.

6.3.3 RELATIONSHIP BETWEEN STANDARD AND NON-STANDARD NORMAL DISTRIBUTIONS

Suppose there is a standard normal random variable Z (with $\mu = 0$ and $\sigma = 1$), and a non-standard normal random variable X (with $\mu \neq 0$ or $\sigma \neq 1$). If you draw the curves of their pdf's, you will find that the variable X 's pdf curve can be obtained by shifting Z 's pdf curve by μ and scaling it by σ . Thus, the z-score of a non-standard normal random variable will play a key role in building a bridge between standard and non-standard normal random variables.

Any non-standard normal random variable X can be converted into a standard normal random variable Z by taking its z-score. The relationship between them is:

$$P(X < x) = P\left(Z < \underbrace{\frac{x - \mu}{\sigma}}_{\text{z-score of } x}\right)$$

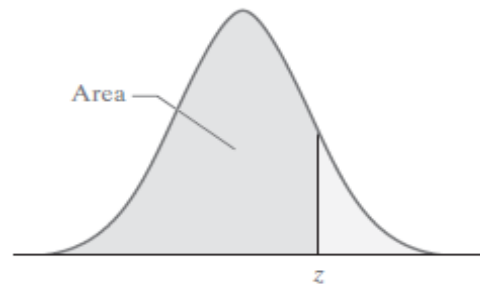
By this relationship, we can calculate the probabilities of any non-standard normal distribution by techniques based on the standard normal distribution.

6.3.4 FINDING PROBABILITIES FOR THE NORMAL DISTRIBUTION USING A NORMAL TABLE

Normal tables give probabilities for the standard normal distribution.

The table we will use for our course gives the area to the left of a specified z-score in a standard normal distribution. That is

$$\text{Area} = F_Z(z) = P(Z < z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$



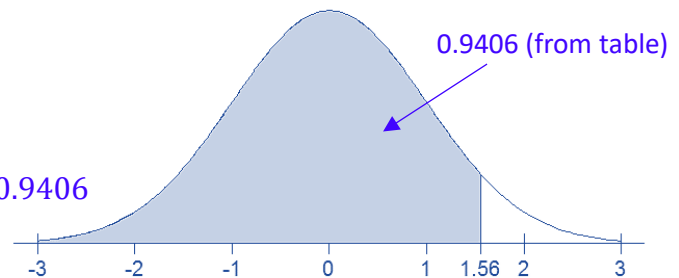
Therefore, for a non-standard normal variable X , to calculate the probability $P(X < x)$, we only need to find the z-score of x . We then use this z-score to find the corresponding probability in a normal table. The other types of probabilities like $P(X > x)$ or $P(a < X < b)$ can be obtained by using the probability rules, such as the complement rule.

Example 6.2: Finding the Probability from normal distribution by normal table

Suppose $X \sim \mathcal{N}(\mu = 0.88, \sigma = 2)$. Find the following probabilities using the **normal table**. Draw a sketch to go along with your answer. Include integral notation with the appropriate bounds.

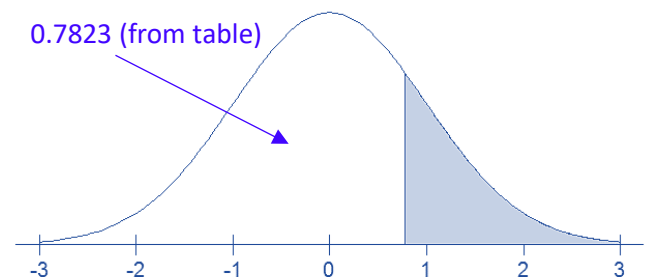
A. Find $P(X < 4)$.

$$\begin{aligned} P(X < 4) &= P\left(Z < \frac{4 - 0.88}{2}\right) \\ &= P(Z < 1.56) = \int_{-\infty}^{1.56} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 0.9406 \end{aligned}$$



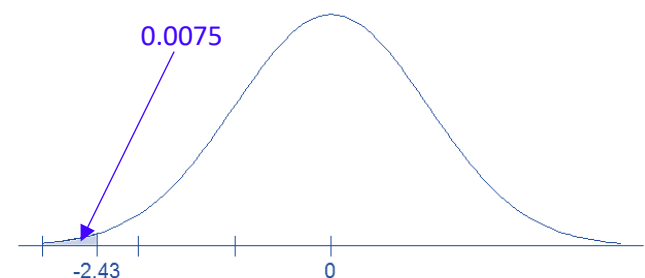
B. Find $P(X > 2.44)$.

$$\begin{aligned} P(X > 2.44) &= P\left(Z > \frac{2.44 - 0.88}{2}\right) \\ &= P(Z > 0.78) = \int_{0.78}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\ &= 1 - 0.7823 = 0.2177 \end{aligned}$$



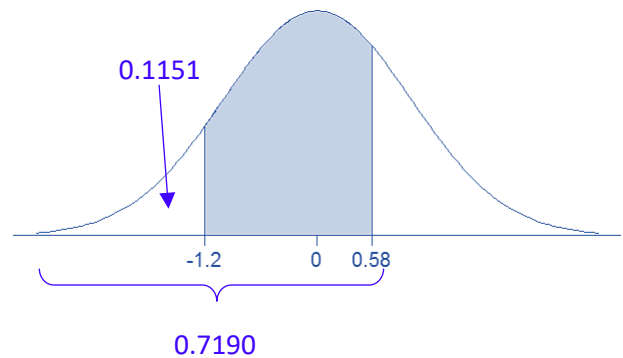
C. Find $P(X \leq -3.98)$.

$$\begin{aligned} P(X \leq -3.98) &= P\left(Z \leq \frac{-3.98 - 0.88}{2}\right) \\ &= P(Z \leq -2.43) = \int_{-\infty}^{-2.43} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\ &= 0.0075 \end{aligned}$$



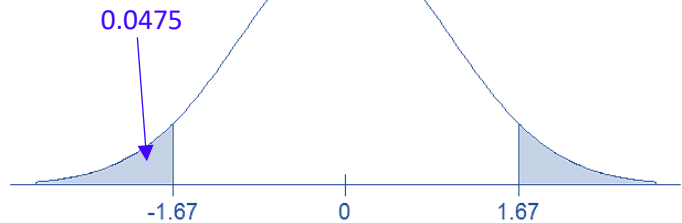
D. Find $P(-1.52 \leq X \leq 2.04)$.

$$\begin{aligned}
 &P(-1.52 \leq X \leq 2.04) \\
 &= P\left(\frac{-1.52 - 0.88}{2} \leq Z \leq \frac{2.04 - 0.88}{2}\right) \\
 &= P(-1.20 \leq Z \leq 0.58) \\
 &= \int_{-1.2}^{0.58} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\
 &= 0.7190 - 0.1151 \\
 &= 0.6039
 \end{aligned}$$



E. Find $P(|X| > 4.22)$. (Hint: Use symmetry of normal pdf curve)

$$\begin{aligned}
 &P(|X| > 4.22) = P\left(|Z| > \frac{4.22 - 0.88}{2}\right) = P(|Z| > 1.67) = P(Z < -1.67) + P(Z > 1.67) \\
 &= \int_{-\infty}^{-1.67} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz + \int_{1.67}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\
 &= 2(0.0475) \quad (\text{by symmetry}) \\
 &= 0.0950
 \end{aligned}$$



6.3.5 FINDING PROBABILITIES FROM A NORMAL DISTRIBUTION USING A CALCULATOR (PART 1)

The **normalcdf** function in the calculator can also help us find the probabilities from a standard normal distribution. In this part, remember to set $\mu = 0$ and $\sigma = 1$ to indicate a standard normal distribution. Also, set “lower” to a negatively large number to integrate $f(x)$ from lower limit.

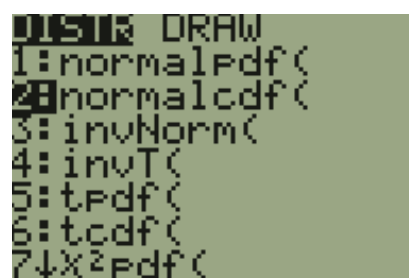
We restrict these settings in this part since we want to copy the idea from normal table.

Calculator Tutorial 6.2: Finding Normal Probabilities (part 1)

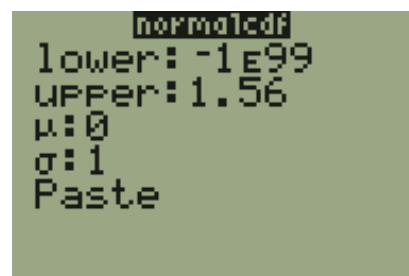
To find **normal probabilities** in your TI-84 calculator, rather than having to type the formula for the pdf of the normal distribution into an integral each time, you can use a built-in function.

(1) Press 2ND + VARS to navigate to the distribution menu.

(2) Use your arrow keys to navigate to item 2: normalcdf(.



(3) Enter your **bounds** of integration and the **mean** and **standard deviation** of the standard normal distribution. If your bounds are $-\infty$ or $+\infty$, you can use the values $-1E99$ or $1E99$ to represent a very large number. Type E by pressing 2ND + the COMMA key.



Remember: For a standard normal distribution, the mean and standard deviation will always be 0 and 1, respectively.

(4) Navigate down to Paste and click ENTER to put the function on the main screen, then click ENTER again to compute the probability.



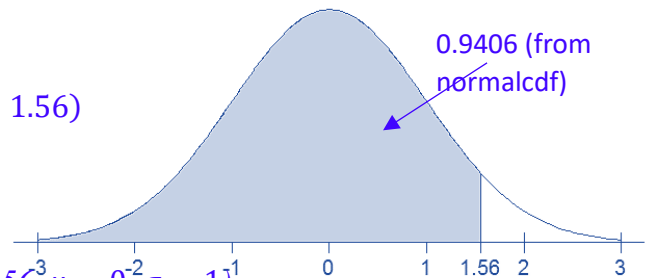
This example calculates $P(Z < 1.56)$.

Example 6.3: Practice with normalcdf in calculator (part 1)

Suppose $X \sim \mathcal{N}(\mu = 0.88, \sigma = 2)$. Find the following probabilities using the **normalcdf** function in your calculator with settings the same as normal table, i.e., lower = -1E99, $\mu = 0$, $\sigma = 1$. Draw a sketch to go along with your answer.

A. Find $P(X < 4)$.

$$P(X < 4) = P\left(Z < \frac{4 - 0.88}{2}\right) = P(Z < 1.56)$$



By normalcdf (lower = -1E99, upper = 1.56, $\mu = 0$, $\sigma = 1$)

We get $P(Z < 1.56) = 0.9406$

B. Find $P(X > 2.44)$.

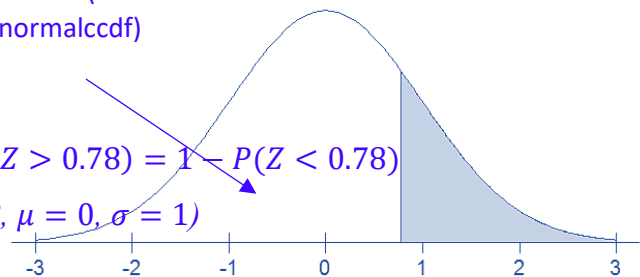
$$P(X > 2.44) = P\left(Z > \frac{2.44 - 0.88}{2}\right) = P(Z > 0.78) = 1 - P(Z < 0.78)$$

By normalcdf (lower = -1E99, upper = 0.78, $\mu = 0$, $\sigma = 1$)

We get $P(Z < 0.78) = 0.7823$.

Thus, $1 - P(Z < 0.78) = 1 - 0.7823 = 0.2177$

0.7823 (from
normalcdf)



You can verify the rest of the results by the calculator on your own.

6.3.6 FINDING PROBABILITIES FROM NORMAL DISTRIBUTION USING A CALCULATOR (PART 2)

Although you are now familiar with the relationship between the standard and non-standard normal distribution, you don't need to convert them every time when finding probabilities. The `normalcdf` function allows us to enter any specific μ and σ to find the probability directly from a non-standard normal distribution. Moreover, this calculator function is not a strictly defined cdf as we learned before since it gives the option to enter the lower limit as something other than $-\infty$. You can treat this `normalcdf` function similar to the integration function and it gives us the convenience of calculating the probability of a normal random variable X in any range, rather than less than a certain value.

Calculator Tutorial 6.3: Finding Normal Probabilities (part 2)

To find **normal probabilities** in your TI-84 calculator, rather than having to type the formula for the pdf of the normal distribution into an integral each time, you can use a built-in function.

(1) Press `2ND + VARS` to navigate to the distribution menu.

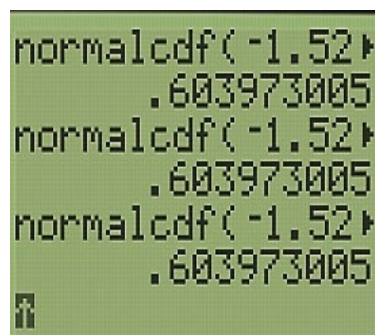


(2) Use your arrow keys to navigate to item 2: `normalcdf`(.

(3) Enter your **bounds** of integration and the **mean** and **standard deviation** of the normal distribution. If your bounds are $-\infty$ or $+\infty$, you can use the values $-1E99$ or $1E99$ to represent a very large number. Type E by pressing `2ND + the COMMA key`.



(4) Navigate down to Paste and click `ENTER` to put the function on the main screen, then click `ENTER` again to compute the probability.



This example calculates $P(-1.52 < X < 2.04) = 0.6040$, where $X \sim \mathcal{N}(\mu = 0.88, \sigma = 2)$.

Example 6.4: Practice with normalcdf in calculator (part 2)

Suppose $X \sim \mathcal{N}(\mu = 0.88, \sigma = 2)$. Find the following probabilities using the **normalcdf** function in the calculator with appropriate values for parameters.

- A. Find $P(X \leq -3.98)$.

By normalcdf (lower = -1E99, upper = -3.98, $\mu = 0.88$, $\sigma = 2$)

We get $P(X \leq -3.98) = 0.0075$

- B. Find $P(-1.52 \leq X \leq 2.04)$.

By normalcdf (lower = -1.52, upper = 2.04, $\mu = 0.88$, $\sigma = 2$)

We get $P(-1.52 \leq X \leq 2.04) = 0.6040$

Example 6.5: Exam Grades

A statistics teacher finds that the exam grades for the second midterm in her business statistics class have a normal distribution with a mean of 82 and a standard deviation of 8.

- A. Find the proportion of students who scored below 85.

*We could look at $z = 0.38$ on the table to find the corresponding probability, **or** we could use the TI-84. You must give at least one appropriate probability statement.*

$$P(X < 85) = \int_{-\infty}^{85} \frac{1}{8\sqrt{2\pi}} e^{-\frac{(x-82)^2}{2(8^2)}} dx \quad \text{or} \quad P(Z < 0.375) = \int_{-\infty}^{0.375} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

Using the calculator: $P(X < 85) = 0.6462$ Using the table: $P(Z < 0.38) = 0.6480$

- B. Find the proportion of students who scored below 70.

$$P(X < 70) = 0.0668$$

- C. Find the proportion of students who scored above 70.

$$P(X > 70) = 1 - 0.0668 = 0.9332$$

- D. Find the proportion of students who scored between 70 and 85.

$$P(70 < X < 85) = \int_{70}^{85} \frac{1}{8\sqrt{2\pi}} e^{-\frac{(x-82)^2}{2(8^2)}} dx = 0.5794$$

6.3.7 INVERSE NORMAL PROBABILITY PROBLEMS

Consider the previous problem from another perspective.

A statistics teacher finds that the exam grades for the second midterm in her business statistics class have a normal distribution with a mean of 82 and a standard deviation of 8.

Find the score that separates the top 10% of scores from the lowest 90% of scores. We are trying to find the value of x such that $P(X < x) = 0.9$ (equivalently we find the value of x such that $P(X > x) = 0.1$). The equation we need to solve is:

$$0.9 = \int_{-\infty}^x \frac{1}{8\sqrt{2\pi}} e^{-\frac{(t-82)^2}{2(8^2)}} dt$$

In other words, to what value would we need to integrate to have an area to the left equal to 0.9?

Fortunately, our calculators have an inverse function that will solve this equation. Most calculators will require you to enter the area under the curve to the **left**, hence the emphasis on using the area to the left. If a problem is written using the right side of the distribution, you will simply need to use the complement rule to convert the problem to consider the left side.

Calculator Tutorial 6.4: Computing Values from Probabilities

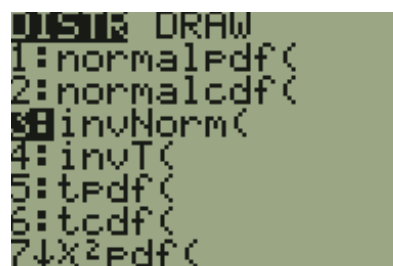
Your TI-84 can be used to find the **value** associated with a **probability**.

(1) Click 2ND + VARS to navigate to the distribution menu.

(2) Use the arrow keys to select item 3: invNorm(.

(3) Enter the probability (area to the left of the value), the mean, and the standard deviation.

(4) Select Paste and click the ENTER key twice. The value associated with the probability you entered will be displayed on the screen.



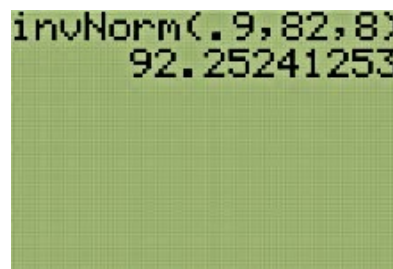
```

DISTR DRAW
1:normalPdf(
2:normalcdf(
3:invNorm(
4:invT(
5:tpdf(
6:tcdf(
7:χ²pdf(
  
```



```

invNorm
area: .9
μ: 82
σ: 8
Paste
  
```



```

invNorm(.9, 82, 8)
92.25241253
  
```

Example 6.6: Exam Grades

A statistics teacher finds that the exam grades for the second midterm in her business statistics class have a normal distribution with a mean of 82 and a standard deviation of 8.

- A. Use the inverse Normal function on your calculator to find the score that separates the top 10% of scores from the lowest 90% of scores.

We want to find x such that $P(X < x) = 0.9$.

$$0.9 = \int_{-\infty}^x \frac{1}{8\sqrt{2\pi}} e^{-\frac{(t-82)^2}{2(8^2)}} dt$$

By calculator, $\text{invNorm}(\text{area} = 0.9, \mu = 82, \sigma = 8)$, we get $x = 92.25$

- B. Suppose the teacher will have a reward of homemade chocolate chip cookies for her students who score in the top 20%. What is the cutoff score needed to determine if a student receives the cookies?

We want to find x such that $P(X > x) = 0.2$.

Recall that for most calculators, the area in invNorm function means the area under the curve to the left. So, if the area under the curve to the right was given like this problem, we need to convert the area to the left first, i.e.,

We want to find x such that $P(X < x) = 0.8$.

Equivalently, we want to find x such that:

$$0.8 = \int_{-\infty}^x \frac{1}{8\sqrt{2\pi}} e^{-\frac{(t-82)^2}{2(8^2)}} dt$$

By calculator, $\text{invNorm}(\text{area} = 0.8, \mu = 82, \sigma = 8)$, we get $x = 88.73$

Thus, any student who scores above 88.73 will receive the cookies.

- C. Suppose the teacher gives an F to students scoring below the 10th percentile. She offers these students the opportunity to take another test to raise their score to a D. What is the cutoff score needed to determine which students get the opportunity to raise their score?

Recall what it means to be at the 10th percentile: 10% of the scores will be equal to or lower than this value.

Thus, we want to find x such that $P(X < x) = 0.1$.

Equivalently, we want to find x such that:

$$0.1 = \int_{-\infty}^x \frac{1}{8\sqrt{2\pi}} e^{-\frac{(t-82)^2}{2(8^2)}} dt$$

By calculator, $\text{invNorm}(\text{area} = 0.1, \mu = 82, \sigma = 8)$, we get $x = 71.75$

Thus, any student scoring below 71.75 will receive an F and an opportunity to raise their score to a D.

MY NOTES:

PRACTICE EXERCISES FOR SECTION 6

1. The distribution of recovery times from mild dehydration x hours after receiving treatment can be modeled by:

$$f_X(x) = \begin{cases} 12x^2 - 12x^3 & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Let the continuous random variable X represent the number of hours until recovery after receiving treatment for dehydration.

- (a) Verify that $f_X(x)$ is a legitimate probability density function.

$$\int_0^1 (12x^2 - 12x^3) dx = 1$$

- (b) Find the probability that the recovery time is between 0.35 and 0.7 hours.

$$P(0.35 < X < 0.7) = \int_{0.35}^{0.7} (12x^2 - 12x^3) dx = 0.5252$$

- (c) Find the probability that a patient's recovery time is exactly 0.1 hours.

$$P(X = 0.1) = 0 \text{ by the properties of a continuous probability distribution.}$$

- (d) Calculate the probability that the recovery time is less than 30 minutes.

$$30 \text{ minutes} = 0.5 \text{ hours}$$

$$P(0 < X < 0.5) = \int_0^{0.5} (12x^2 - 12x^3) dx = 0.3125$$

- (e) What is the expected time it will take for a patient to recover after being treated?

$$\mu_X \text{ or } E(X) = \int_0^1 x (12x^2 - 12x^3) dx = 0.6 \text{ hours}$$

- (f) What is the standard deviation of times it takes for treated patients to recover?

$$\sigma_X = \sqrt{\int_0^1 (x - 0.6)^2 (12x^2 - 12x^3) dx} = \sqrt{0.04 \text{ hours}^2} = 0.2 \text{ hours}$$

2. Normal distributions are always symmetric and unimodal. Suppose you want to analyze the exam scores from a STAT 3090 course and you believe the scores are normally distributed. To simplify things, you convert all of the exam scores to z-scores. What percent of the exam scores would you expect to have z-scores between -2 and 2 ? Find this percent by both Empirical Rule and normal pdf, then compare the results.

(a) *By the Empirical Rule:*

The standard normal distribution has mean $\mu = 0$ and standard deviation $\sigma = 1$.

Between -2 and 2 represents values within 2 standard deviations from the mean. By the Empirical Rule, approximately 95% of the exam scores will have z-scores between -2 and 2 .

(b) *By the normal pdf:*

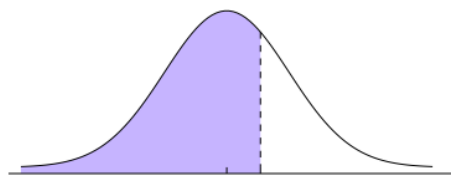
$$P(-2 < Z < 2) = \int_{-2}^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 0.9545 \text{ by normalcdf (lower} = -2, \text{upper} = 2, \mu = 0, \sigma = 1)$$

*Note, if a random variable follows normal distribution, normal pdf and integration can calculate the **exact** probabilities from it. Though the Empirical Rule is applicable here since the normal distribution is bell-shaped, it only gives an **approximation**.*

3. Calculate the following probabilities for a standard normal random variable Z . Include an appropriately labeled sketch.

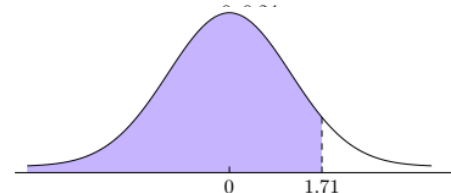
(a) $P(Z < 0.24)$

$$P(Z < 0.24) = \int_{-\infty}^{0.24} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 0.5948$$



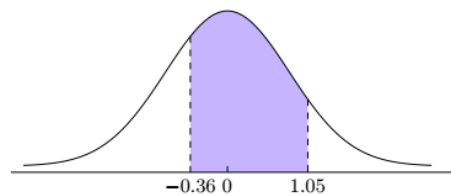
(b) $P(Z \geq 1.71)$

$$P(Z > 1.71) = \int_{1.71}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 0.0436$$



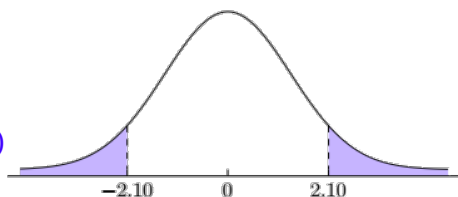
(c) $P(-0.36 < Z < 1.05)$

$$\begin{aligned} P(-0.36 < Z < 1.05) &= \int_{-0.36}^{1.05} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\ &= 0.8531 - 0.3594 = 0.4937 \end{aligned}$$



(d) $P(|Z| > 2.10)$

$$\begin{aligned} P(|Z| > 2.10) &= P(Z > 2.10) + P(Z < -2.10) \\ &= 2(0.0179) = 0.0358 \end{aligned}$$



4. A shoe retailer wants to expand their sales on a global level and decides to begin marketing to other countries. The retailer wants to ensure that they are having their manufacturer make the correct number of shoes in each size, so he requests that the resident statistician perform an analysis. She finds that shoe sizes are normally distributed, with the shoe size in Country A having a mean of 6.5 and standard deviation of 0.4, and the shoe size in Country B having a mean of 7.5 and standard deviation of 1.2.

- (a) What is the percentile rank of size 7 in Country A? Equivalently, at what percentile would a size 7 be in Country A? Interpret your result in terms of a proportion of the population.

$$z = \frac{7 - 6.5}{0.4} = 1.25, \quad P(\text{Shoe Size in A} < 7) \text{ or } P(Z < 1.25) = 0.8944$$

89.44% of the population in Country A has a shoe size of 7 or smaller.

- (b) What is the percentile rank of size 7 in Country B? Equivalently, at what percentile would a size 7 be in Country B? Interpret your result in terms of a proportion of the population.

$$z = \frac{7 - 7.5}{1.2} = -0.40, \quad P(\text{Shoe Size in B} < 7) \text{ or } P(Z < -0.40) = 0.3446$$

34.46% of the population in Country B has a shoe size of 7 or smaller.

- (c) Suppose that the manufacturer currently only has the technology to make size 7 shoes and below. On which country should the retailer focus their marketing efforts?

In Country A, there is a larger proportion of individuals who wear a size 7 or smaller shoe, so the retailer should focus on marketing to Country A.

5. Babies' birth weights are normally distributed with a mean of 120 ounces and a standard deviation of 20 ounces. What would the weight of a newborn be who falls in the 85th percentile for birth weights? What would the cutoff of weight be for the top 25% heaviest babies?

We want find x such that $P(X < x) = 0.85$. We can use the inverse Normal function to find the value of x . $\text{invNorm}(\text{area}=0.85, \mu = 120, \sigma = 20)$ The 85th percentile therefore is 140.7 oz.

Next, we want to find x such that $P(X > x) = 0.25$, convert the right-tail area to left-tail area first, $\text{area} = 1 - 0.25 = 0.75$, then $\text{invNorm}(\text{area}=0.75, \mu = 120, \sigma = 20)$ gives us result of 133.5 oz.

SECTION 7: SAMPLING DISTRIBUTIONS

HAWKES SECTIONS 8.1 – 8.3

7.1 A DILEMMA

Recall what we learned in Section 1. A statistic is a numerical descriptive measure of a sample. All numerical summaries calculated from a sample are statistics, e.g., sample mean $\bar{x}$, sample proportion $\hat{p}$, sample variance s^2 . These statistics calculated from our samples can be used to estimate parameters from a population, e.g., population mean μ , population proportion p , population variance σ^2 .

The statistics we calculated in Section 1 were from a **particular** sample (like a fixed dataset). They were all values related to that specific sample. Recall also from Section 1 that our goal in sampling was to get a representation of the whole population. Thus, we took a **random** sample from the population of interest.

Example 7.1: Coffee

Suppose you want to know the average amount of coffee a college student on your campus drinks (in ounces) in the morning.

- A. What **parameter** do you want to estimate?

The true average amount of coffee that all college students on your campus drink in the morning.

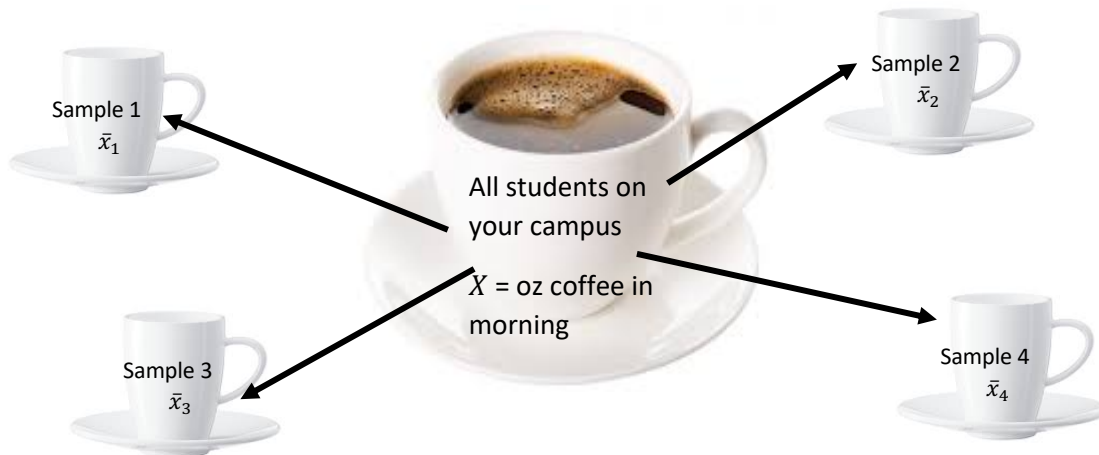
- B. What **statistic** could you find from a sample of students to estimate the parameter?

The sample mean $\bar{x}$.

Let's take a step back. We have a population, with some variable of interest, from which we will select a simple random sample. Before the sample is selected and observed, we know what variable we want to measure, the parameter we want to estimate and the statistic we will use to estimate it. However, we don't know which members of the population will be selected. Therefore, of course, we don't know what value the statistic will be.

Suppose that for the previous example we decide to take a simple random sample of size 50 of college students on our campus and poll them. We are going to assume that these students will all respond to the poll. Before taking the sample, we do not know who the individuals in the sample will be. We do not know how much coffee they drink in the morning and we do not know the value of the sample mean until we actually contact them.

Let X be the amount of coffee each student drinks in the morning. Based on our knowledge of probability it sounds like X is a random variable. X can be any value in the range $[0, \infty)$. There are many random samples of size 50 that we could collect from this population. Since we are using simple random sampling, each of these samples of size 50 are equally likely to be our sample. Each of these samples will have its own sample mean $\bar{x}$.



How then can we use one sample statistic to estimate a population parameter? Each sample we could potentially select of the same size and sampling method from the same population can produce a different value for the statistic. Doesn't that make value of the sample statistic also a random variable? In the coffee example, each sample mean $\bar{x}$ will have units of ounces and it can take values in the range $[0, \infty)$. What do we do?

As it turns out, sample statistics are indeed random variables in their own right. Each type of statistic, such as $\bar{x}$ or $\hat{p}$, or s^2 , will have its own probability distribution to describe how the probabilities will be distributed over its possible values. We call the probability distribution of a statistic a [sampling distribution](#) since the distribution comes about through the process of sampling.

With the sampling distribution of a statistic in hand, which is simply a special probability distribution, we can also find some numerical summaries from this distribution, such as the mean and standard deviation. You will see some long names in this section such as "the mean of sampling distribution of sample mean", or "the standard deviation of sampling distribution of sample proportion" etc. Don't be scared by them - they are just some numerical summaries from some probability distributions.

It is through our knowledge of probability distributions that we can learn how to use the statistics we calculate as [estimates](#) for population parameters. In this section we will explore the sampling distribution of two important statistics, the sample mean and the sample proportion. These distributions will play key roles in the methods of statistical inference that will be introduced in later sections.

The concepts in this section are some of the most difficult ones you will encounter. To assist you in visualizing sampling distributions, we will often use simulations to approximate them. To simulate a sampling distribution:

1. Take many samples of size n from the population such as 100 or more samples
2. For each sample of size n , calculate the statistic of interest
3. Plot all of the statistics in a histogram. This histogram will simulate the shape of the sampling distribution.
4. Using all of the statistics, calculate their mean, which simulates the mean of the sampling distribution.
5. Using all of the statistics, calculate their standard deviation, which simulates the standard deviation of the sampling distribution.

We need to make some compromises in the notation we learned in the probability section. When referring to the name of a random variable, we tend to use a capital letter. When we talk about the sampling distribution of a statistic, we will need to use context to help us determine if we are talking about the statistic as a random variable or a specific calculated value. This is the notation used in your text.

7.2 SAMPLING DISTRIBUTION OF THE SAMPLE MEAN $\bar{x}$

We've had a first look at the sampling distribution of sample mean $\bar{x}$ in Example 7.1. Let's consider the following questions.

- If we generate an infinite number of samples from a population of values, let's call the random variable involved X , and calculate $\bar{x}$ for each sample and then plot all of the $\bar{x}$'s in a histogram, will it converge to a curve?
- If the random variable X follows a specific distribution, what will the curve look like for the distribution of the values of $\bar{x}$?
- How do the parameters of the random variable X and the sample size n affect the shape, center, and spread of the sampling distribution of the values for $\bar{x}$?
- Is the center of this sampling distribution a good estimate for the population mean μ_X ?

You will find the answers in the following information about the sampling distribution of the sample mean $\bar{x}$.

Properties of the Sampling Distribution of the Sample Mean

Suppose a simple random sample of size n is taken from a large population in which the variable of interest X has mean μ_X and standard deviation σ_X . Then the sampling distribution of the sample mean has the following properties:

- **Mean (measure of center):** $\mu_{\bar{x}} = E(\bar{x}) = \mu_X$

The mean of the sampling distribution of sample mean is equal to the population mean of the random variable X . In other words, the sample mean is centered around the true population mean.

- **Standard deviation (measure of spread):** $\sigma_{\bar{x}} = \sigma_X / \sqrt{n}$

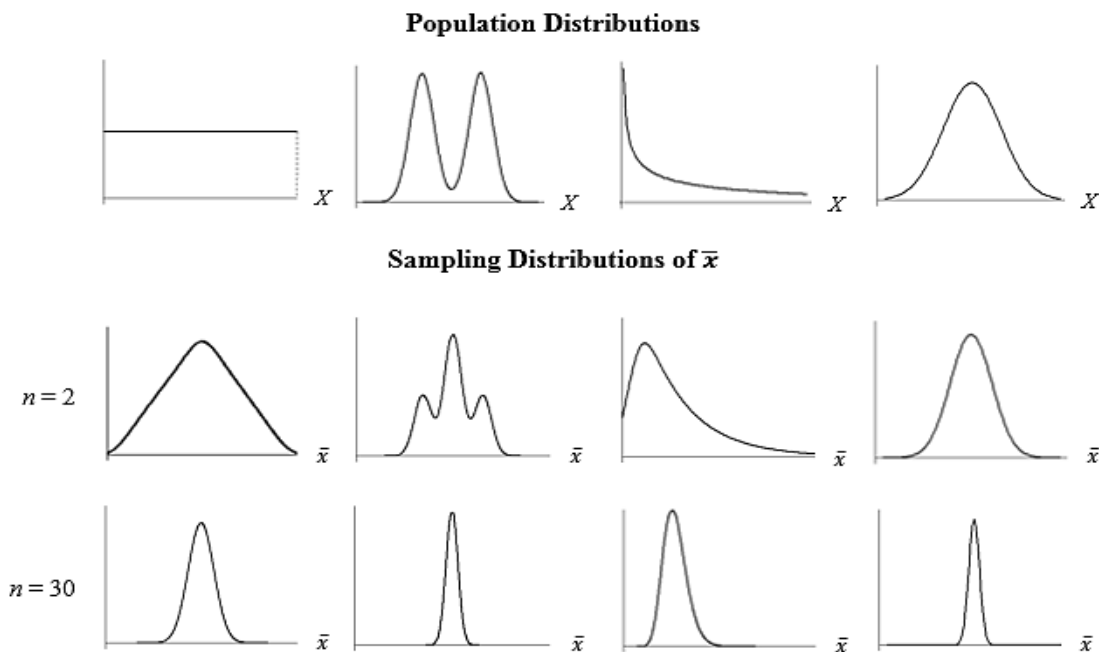
The standard deviation of the sampling distribution of the sample mean is called the standard error of $\bar{x}$.

- The standard error describes on average how much each value of $\bar{x}$ varies from the mean of the sample means.
- From the formula, we can see that the values of $\bar{x}$ tend to vary less than individual observations.
- The standard error of $\bar{x}$ is relevant to the width of the sampling distribution. As n increases, the denominator increases, causing standard error of $\bar{x}$ to decrease, and the width of the curve of its distribution to be narrower. Thus, with larger samples, the values of the sample mean tend to vary less around the true population mean.
- **Shape:**
 - If the distribution of the population variable X is approximately normal, the sampling distribution of the sample mean will be approximately normal regardless of the sample size n . Therefore $\bar{x} \sim N\left(\mu_X, \sigma_X / \sqrt{n}\right)$.
 - Otherwise, if the population distribution of X is not normally distributed, or even if it is unknown, if the sample size n is large enough, the sampling distribution of the sample mean will be approximately normal by the Central Limit Theorem.

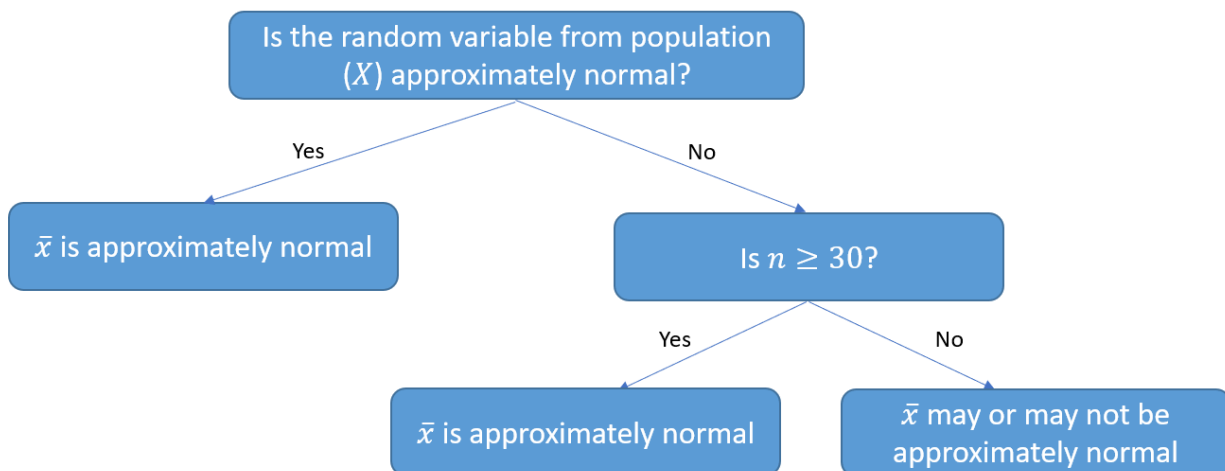
7.2.1 THE CENTRAL LIMIT THEOREM

If n is large (for this course, $n \geq 30$), the sampling distribution of the sample mean will be approximately normal regardless of the shape of the distribution of the population variable X . The larger the sample size, the better the normal approximation will be.

Central Limit Theorem Illustration



To determine the shape of the sampling distribution of the sample mean, you can ask yourself a series of questions...



Give it a try! You can explore the sampling distribution for $\bar{x}$ with the following simulation

<http://www.rossmanchance.com/applets/OneSample.html>

7.2.2 Z-SCORE FORMULA FOR SAMPLE MEANS

If the sampling distribution of $\bar{x}$ is approximately normal, then we can use the z-score formula and the standard normal table to find the probabilities for $\bar{x}$.

Formula 7.1: Z-Score for Sample Mean

Consider the sampling distribution for sample mean $\bar{x}$ with mean $\mu_{\bar{x}}$ and standard error $\sigma_{\bar{x}}$. The **z-score** for an observed **sample mean** $\bar{x}$ is:

$$z = \frac{\text{observation} - \text{mean}}{\text{standard deviation}} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_X}{\sigma_X / \sqrt{n}}$$

where μ_X and σ_X are mean and standard deviation of population variable X , i.e., they are population mean and population standard deviation respectively. Their subscripts of X can be omitted, i.e., μ and σ .

Be careful to note the difference in z-score formulas when finding probabilities of values for *individual observations* versus finding probabilities of values for means of *samples*. The denominators are different and can alter the corresponding probability significantly!

Individual	Sample
$z = \frac{x - \mu_X}{\sigma_X}$	$z = \frac{\bar{x} - \mu_X}{\sigma_X / \sqrt{n}}$

Example 7.2: (Practice with Normal Table) Fertilizer Retailer

A company fills bags with fertilizer for retail sale. The weights of the bags of fertilizer have a normal distribution with a mean weight of 15 pounds and standard deviation of 1.70 pounds.

- A. Describe the sampling distribution for $\bar{x}$ for a sample of size n in this situation using the shape, center, and spread.

The weights of the bags of fertilizer are normally distributed, so the distribution of $\bar{x}$ will be approximately normal with a mean of 15 pounds and standard deviation of $1.70/\sqrt{n}$ pounds.

- B. What is the probability that a **single** randomly-selected bag of fertilizer will weigh between 14 and 16 pounds?

$$z_{\text{lower}} = \frac{14 - 15}{1.70} = -0.59$$

$$z_{\text{upper}} = \frac{16 - 15}{1.70} = 0.59$$

$$P(14 < X < 16) = P(-0.59 < Z < 0.59) = P(Z < 0.59) - P(Z < -0.59) = 0.4436$$

- C. If 35 bags of fertilizer are randomly selected, find the probability that the **average weight** of the 35 bags will be between 14 and 16 pounds.

$$z_{\text{lower}} = \frac{14 - 15}{1.70/\sqrt{35}} = -3.48$$

$$z_{\text{upper}} = \frac{16 - 15}{1.70/\sqrt{35}} = 3.48$$

$$P(14 < \bar{x} < 16) = P(-3.48 < Z < 3.48) = P(Z < 3.48) - P(Z < -3.48) = 0.9995$$

Alternatively, if you don't want to use the z-score or normal table, you can use the **normalcdf** function in calculator to directly find probabilities for $\bar{x}$. Recall how we find probabilities from a non-standard normal distribution by using the normalcdf function in Section 6. Be careful with the **standard deviation** you are going to enter. Since we want to find probabilities from the $\bar{x}$ now, we need to enter $\bar{x}$'s standard deviation, that is, $\sigma_x/\sqrt{n}$.

Example 7.3: (Practice with Calculator) Oil Change Times

The shape of the distribution of the time required to get an oil change at a 10-minute oil change facility is somewhat skewed right. Records indicate that the mean time for an oil change is 11.4 minutes with standard deviation 3.2 minutes.

- A. To compute probabilities regarding the sample mean using the normal distribution, what would be required?

We would need to have a sample size of at least 30 observations ($n \geq 30$) to apply the Central Limit Theorem since the shape of the distribution of oil change times is not normal.

- B. What is the probability that a random sample of 40 oil changes would result in a sample mean of less than 10 minutes? Find the probability by both table and calculator.

Since $n = 40 > 30$, $\bar{x}$ is approximately normal.

$$\text{By table: } P(\bar{x} < 10) = P\left(Z < \frac{10-11.4}{3.2/\sqrt{40}}\right) = P(Z < -2.77) = 0.0028$$

$$\text{By calculator: } P(\bar{x} < 10) = \text{normalcdf}(\text{lower} = -1E99, \text{upper} = 10, \mu = 11.4, \sigma = 3.2/\sqrt{40}) = 0.0028$$

- C. What is the probability that a random sample of 35 oil changes would result in a sample mean of between 10 and 12 minutes? Find the probability by both table and calculator.

Since $n = 35 > 30$, $\bar{x}$ is approximately normal.

$$\text{By table: } P(10 < \bar{x} < 12) = P\left(\frac{10-11.4}{3.2/\sqrt{35}} < Z < \frac{12-11.4}{3.2/\sqrt{35}}\right) = P(-2.59 < Z < 1.11) = P(Z < 1.11) - P(Z < -2.59) = 0.8615$$

$$\text{By calculator: } P(10 < \bar{x} < 12) = \text{normalcdf}(\text{lower} = 10, \text{upper} = 12, \mu = 11.4, \sigma = 3.2/\sqrt{35}) = 0.8615$$

Example 7.4: Relationship between Standard Error and Sample Size

Suppose that the standard deviation of the sampling distribution for $\bar{x}$ is $\sigma_{\bar{x}}$. In terms of $\sigma_{\bar{x}}$, what would be the standard deviation for $\bar{x}$ if...

- A. The sample quadrupled in size?

We know that $\sigma_{\bar{x}} = \frac{\sigma_X}{\sqrt{n}}$. Let $\sigma_{\bar{x}_{new}} = \frac{\sigma_X}{\sqrt{4n}} = \frac{1}{\sqrt{4}} \frac{\sigma_X}{\sqrt{n}} = \frac{1}{2} \sigma_{\bar{x}}$. The standard error is half as large.

- B. The sample was half as large?

We know that $\sigma_{\bar{x}} = \frac{\sigma_X}{\sqrt{n}}$. Let $\sigma_{\bar{x}_{new}} = \frac{\sigma_X}{\sqrt{\frac{1}{2}n}} = \frac{1}{\sqrt{\frac{1}{2}}} \frac{\sigma_X}{\sqrt{n}} = \sqrt{2} \sigma_{\bar{x}}$. The standard error increases by a factor of $\sqrt{2}$.

Example 7.5: Comparing Standard Errors

Which of the following will provide the **largest** standard deviation of the sampling distribution (in other words, the largest standard error) for $\bar{x}$?

- A. Random sample of size 50 from a population with $\mu = 35$ and $\sigma = 10$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{50}} \approx 1.414$$

- B. Random sample of size 30 from a population with $\mu = 35$ and $\sigma = 10$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{30}} \approx 1.826$$

- C. Random sample of size 50 from a population with $\mu = 15$ and $\sigma = 20$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{50}} \approx 2.828$$

- D. Random sample of size 30 from a population with $\mu = 15$ and $\sigma = 20$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{30}} \approx 3.651$$

MY NOTES:

7.3 SAMPLING DISTRIBUTION FOR THE SAMPLE PROPORTION $\hat{p}$

We now turn our attention to numerical measures summarizing *qualitative* data. Suppose that we wish to study a population in which each individual of the population either does or does not have a certain characteristic. Recall that the **population proportion** p is the fraction of a population that possesses some characteristic of interest:

$$p = \frac{x}{N},$$

where x is the number of observations in the population with the characteristic and N is the population size.

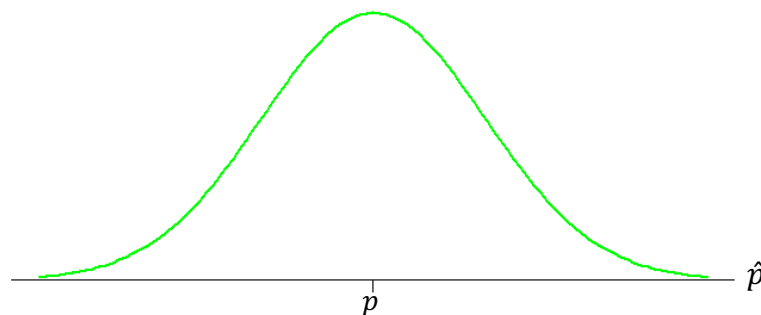
Since the population is usually too large to measure, we estimate the parameter p using the statistic **sample proportion** $\hat{p}$:

$$\hat{p} = \frac{x}{n} = \frac{\text{number of successes}}{\text{sample size}}$$

where x is the number of observations in the sample with the characteristic and n is the sample size.

For example, suppose that there are x successes in a binomial experiment with n trials, but we do not know the value of the probability of success p . We can use the value of $\hat{p} = \frac{x}{n}$ as an estimate for the probability of success p .

Just like the sample mean $\bar{x}$ from a random sample, the sample proportion $\hat{p}$ from a random sample is also a **random variable** whose value will vary. The **sampling distribution** of sample proportion $\hat{p}$ describes possible values that $\hat{p}$ from a sample of size n could take and how often it takes those values.



Give it a try! You can explore the sampling distribution for $\hat{p}$ with the following simulation...

<http://www.rossmanchance.com/applets/OneProp/OneProp.htm?candy=1>

Properties of the Sampling Distribution of the Sample Proportion

Suppose a simple random sample of size n is taken from a large population where the proportion of the population having a characteristic of interest is p . Then the sampling distribution of the sample proportion $\hat{p}$ has the following properties:

- **Mean (center):** $\mu_{\hat{p}} = p$

The mean (center) of the sampling distribution of the sample proportion $\hat{p}$ is equal to the population proportion. In other words, the values of $\hat{p}$ will vary around the true value of p .

- **Standard deviation (spread):** $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

The standard deviation of the sampling distribution of sample proportion $\hat{p}$ is also called the standard error of $\hat{p}$. It implies that as sample size increases, the spread of the sampling distribution decreases. In other words, as the sample size increases, the values for $\hat{p}$ vary less around the value of p .

- **Shape:** For large samples, the sampling distribution of $\hat{p}$ is approximately normal. A sample size is considered “large” if:

$$np \geq 5 \quad \text{and} \quad n(1 - p) \geq 5$$

Note, the sample size n cannot sole determine the normality of $\hat{p}$, which is different from $\bar{x}$. No matter how large the n is, we always need to check both $np \geq 5$ and $n(1 - p) \geq 5$ for the normality of $\hat{p}$.

7.3.1 Z-SCORE FORMULA FOR SAMPLE PROPORTIONS

If the sampling distribution of $\hat{p}$ is approximately normal, then we can use the z-score formula and the standard normal table to find the probability of obtaining certain $\hat{p}$ values.

Formula 7.2: Z-Score for Sample Proportion

Consider the sampling distribution for sample proportion $\hat{p}$ with mean $\mu_{\hat{p}}$ and standard error $\sigma_{\hat{p}}$. The **z-score** for an observed **sample proportion** $\hat{p}$ is:

$$z = \frac{\text{observation} - \text{mean}}{\text{standard deviation}} = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Example 7.6: (Practice with Normal Table) Household Dogs

According to the American Humane Society, 47% of American households own a dog. Let $\hat{p}$ represent the proportion of households that own a dog in a random sample of 100 American households.

- A. Describe the sampling distribution of $\hat{p}$.

Remember: How do we describe a distribution? (Shape, Center, Spread)

Shape: $np = 100(0.47) = 47 \geq 5$ and $n(1 - p) = 100(0.53) \geq 5$

So $\hat{p}$ is approximately normal.

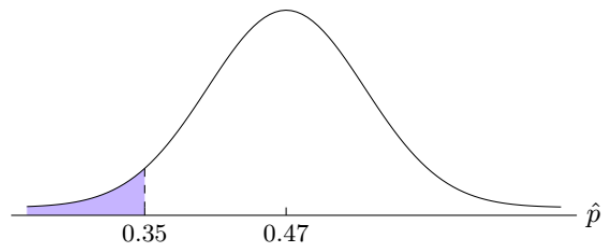
Center (or Mean): $\mu_{\hat{p}} = p = 0.47$

Spread (or Standard Deviation or Standard Error): $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.47(0.53)}{100}} = 0.0499$

- B. What is the probability that less than 35% of households in a random sample of 100 American households own a dog? Write down the value of $\hat{p}$ and draw a sketch to go along with your answer.

$$\hat{p} = 0.35$$

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.35 - 0.47}{0.0499} = -2.40$$



$$P(\hat{p} < 0.35) = P(Z < -2.40) = 0.0082$$

- C. Suppose you actually obtain a sample in which 35 out of 100 randomly selected American households own a dog. Would this cause you to suspect that the Humane Society's claim that 47% of American households own a dog is too high? Explain.

Yes. The probability of observing 35 or fewer homes with a dog in a sample of 100 is too low (0.0082) if we assume 47% of homes own a dog is correct. Our sample provides evidence that the true proportion of dog owners is less than 47%.

Alternatively, you can use **normalcdf** function in calculator to find such probability without calculating the z-score. Make sure to enter the correct mean and standard deviation - they should be from $\hat{p}$.

Example 7.7: (Practice with Calculator) Voter Polling

Suppose that a particular candidate for public office (in a race with only two candidates) is in fact favored by 48% of the voting population in the district.

A polling organization takes a random sample of 625 voters and will use $\hat{p}$, the sample proportion, to estimate p . What is the approximate probability that $\hat{p}$ will be greater than 0.5, causing the polling organization to incorrectly predict the result of the upcoming election?

What information are we given?

True proportion favoring a candidate: $p = 0.48$

Sample size: $n = 625$

Sample proportions in which we are interested: $\hat{p} > 0.5$

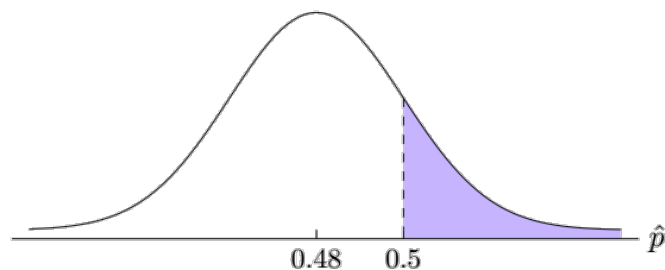
What do we know about the distribution of $\hat{p}$?

$$\mu_{\hat{p}} = p = 0.48$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.48(0.52)}{625}} = 0.0200$$

Since $np = 625(0.48) = 300 \geq 5$ and $n(1-p) = 625(0.52) = 325 \geq 5$, $\hat{p}$ is approximately normally distributed.

*We can use normalcdf function in calculator to find probabilities for $\hat{p}$ instead of z-score or normal table. By calculator: $P(\hat{p} > 0.5) =$
normalcdf(lower = 0.5, upper =
1E99, $\mu = 0.48$, $\sigma = 0.02$) = 0.1587*



Remember to enter the correct mean of $\hat{p}$, i.e., p , and the correct standard error of $\hat{p}$, i.e., $\sqrt{p(1-p)/n}$.

The probability that the polling organization incorrectly predicts the election result is 0.1587.

MY NOTES:

PRACTICE EXERCISES FOR CHAPTER 7

1. Given that a continuous random variable X is normally distributed with a mean of 40 and a standard deviation of 13, calculate the probability that a sample of size 49 has a sample mean of... (You can try using both normal table and calculator.)

- (a) Greater than 37

Since X is normally distributed, $\bar{x}$ is normally distributed regardless of n .

$$P(\bar{x} > 37) = P\left(Z > \frac{37 - 40}{13/\sqrt{49}}\right) = P(Z > -1.62) = 0.9469$$

Or 0.9474 if using the standard normal table or the rounded value of z .

- (b) At least 42.5

Since X is normally distributed, $\bar{x}$ is normally distributed regardless of n .

$$P(\bar{x} \geq 42.5) = P\left(Z \geq \frac{42.5 - 40}{13/\sqrt{49}}\right) = P(Z \geq 1.35) = 0.0891$$

Or 0.0885 if using the standard normal table or the rounded value of z .

- (c) Between 39 and 43

Since X is normally distributed, $\bar{x}$ is normally distributed regardless of n .

$$P(39 < \bar{x} < 43) = P\left(\frac{39 - 40}{13/\sqrt{49}} < Z < \frac{43 - 40}{13/\sqrt{49}}\right) = P(-0.54 < Z < 1.62) = 0.6518$$

Or 0.6528 if using the standard normal table or the rounded values of z .

- (d) No more than 35

Since X is normally distributed, $\bar{x}$ is normally distributed regardless of n .

$$P(\bar{x} \leq 35) = P\left(Z \leq \frac{35 - 40}{13/\sqrt{49}}\right) = P(Z \leq -2.69) = 0.0035$$

Or 0.0036 if using the standard normal table or the rounded value of z .

2. All Clear Windows makes windows for use in homes and commercial buildings. The standards for glass thickness call for the glass to average 0.375 inches with a standard deviation of 0.050 inches. Let $\bar{x}$ represent the mean thickness of 50 randomly selected windows.

(a) Can we use a normal distribution to calculate probabilities for $\bar{x}$? Explain.

Yes. Since $n = 50 \geq 30$, $\bar{x}$ approximately follows normal distribution by the Central Limit Theorem.

(b) Describe the mean (center) and standard error (spread) of the sampling distribution of $\bar{x}$.

Center: $\mu_{\bar{x}} = 0.375$ inches, Spread: $\sigma_{\bar{x}} = \frac{0.050}{\sqrt{50}} = 0.007$ inches

(c) Suppose a random sample of $n = 50$ windows yields a mean thickness of 0.392 inches. What is the likelihood of observing a sample with a mean thickness at least as thick as ours? (In other words, what is the probability that $\bar{x}$ is greater than 0.392 inches?)

If using the exact value for standard error $\sigma_{\bar{x}}$:

$$P(\bar{x} > 0.392) = P\left(Z > \frac{0.392 - 0.375}{0.050/\sqrt{50}}\right) = P(Z > 2.40) = 0.0081$$

Or 0.0082 if using the standard normal table or the rounded value of z .

If using the rounded value of standard error $\sigma_{\bar{x}}$ found in Part (b):

$$P(\bar{x} > 0.392) = P\left(Z > \frac{0.392 - 0.375}{0.007}\right) = P(Z > 2.43) = 0.0076$$

Or 0.0075 if using the standard normal table or the rounded value of z .

3. Suppose the lifespan of a particular brand of calculator battery, X , is approximately normally distributed with a mean of 75 hours and a standard deviation of 10 hours.

- (a) Let $\bar{x}$ be the mean lifespan of a sample of n batteries. Does $\bar{x}$ follow a normal distribution? Explain.

Yes. $\bar{x}$ approximately follows normal distribution since the population X follows a normal distribution.

- (b) What is the probability that a single randomly selected battery from the population will have a lifespan between 70 and 80 hours?

$$P(70 < X < 80) = P\left(\frac{70 - 75}{10} < Z < \frac{80 - 75}{10}\right) = P(-0.5 < Z < 0.5) = 0.3829$$

- (c) What is the probability that 16 randomly selected batteries from the population will have a mean lifespan between 70 and 80 hours?

$$P(70 < \bar{x} < 80) = P\left(\frac{70 - 75}{10/\sqrt{16}} < Z < \frac{80 - 75}{10/\sqrt{16}}\right) = P(-2 < Z < 2) = 0.9545$$

4. A nationwide survey analyzing trends in popular media found that 81% of U.S. college students prefer British baking shows over American baking shows. You are interested to see if this result holds at your university. You take a sample of 140 students on campus and find that 125 of them prefer watching British baking shows.

- (a) Can you use the normal distribution to approximate probabilities for the sample proportion $\hat{p}$ of students at your university who prefer British baking shows? Check the appropriate conditions to explain your answer.

Yes. Since $np = 140(0.81) = 113.4 \geq 5$ and $n(1 - p) = 140(0.19) = 26.6 \geq 5$, we can say that $\hat{p}$ is approximately normal.

- (b) You find that the probability of obtaining a sample where $\hat{p}$ is at least as great as your sample is $P\left(\hat{p} > \frac{125}{140}\right) = P(Z > 2.50) = 0.0062$. Would this cause you to suspect that the national result is an over- or an underestimate for your university?

The probability of obtaining a sample of 140 where 89% or more prefer British baking shows is small, meaning that this is a rare event. This may cause us to suspect that the national result underestimates the true proportion at our university.

5. Leslie has been tasked with putting together a statistical report regarding the use of a particular park in her hometown. Previous data show that 72% of the residents living in the town visited the park in the last month. She wants to know the probability that more than 98 individuals in a random sample of 150 residents have visited the park in the last month.
- (a) Based on the information given, is the sample proportion $\hat{p}$ of residents who have visited the park in the last month approximately normally distributed? Check the appropriate conditions to justify your answer.

$$np = 150(0.72) = 108 \geq 5 \text{ and } n(1 - p) = 150(0.28) = 42 \geq 5$$

Yes, $\hat{p}$ is approximately normal.

- (b) What is the sample proportion $\hat{p}$ for Leslie's sample of 150 residents? Round your answer to two decimal places.

$$\hat{p} = \frac{98}{150} = 0.65$$

- (c) What is the probability that more than 98 individuals in a random sample of 150 residents have visited the park in the last month?

$$P(\hat{p} > 0.65) = P\left(Z > \frac{0.65 - 0.72}{\sqrt{\frac{0.72(1-0.72)}{150}}}\right) = P(Z > -1.91) = 1 - P(Z < -1.91) = 1 -$$

$$0.0281 = 0.9719 \text{ or by calculator } P(\hat{p} > 0.65) = \text{normalcdf}\left(\text{lower} = 0.65, \text{upper} =$$

$$1E99, \mu = 0.72, \sigma = \sqrt{\frac{0.72(1-0.72)}{150}}\right) = 0.9719$$

SECTION 8: CONFIDENCE INTERVALS

HAWKES SECTIONS 9.6 AND 9.7

Sampling distributions provide the basis for a powerful use of statistics: **statistical inference**. Statistical inference involves making statements or answering questions about population parameters using sample data. We can make decisions and predictions about populations, even if we have data for relatively few subjects from that population.

Recall that populations are mostly large (e.g., all Clemson students who drink coffee every morning) and in almost all cases it is not possible to find the exact value of their parameters (e.g., their mean, standard deviation etc.). In this section, we will learn methods for **estimating** the value of a population parameter, and in the next section, we will learn methods for **testing hypotheses** about the parameter values.

8.1 ESTIMATING POPULATION PARAMETERS

A target parameter is the unknown population parameter that we are interested in estimating.

- For quantitative data, this is typically the population mean μ and population standard deviation σ .
- For qualitative data, this is typically the population proportion possessing some characteristic of interest (i.e., the proportion of successes) p .

Again, a target parameter of a population is unknown, and we would never know what value it takes. Estimation (scientific estimations if we want them to be reliable with some confidence) is the only way to learn about the parameter.

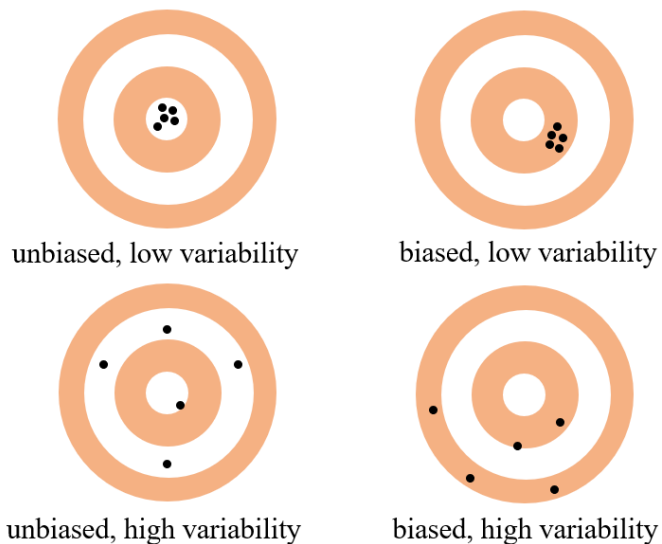
8.1.1 POINT ESTIMATES

A point estimate is a single number that is our “best guess” for the target parameter. The following table provides a few examples of estimators.

Parameter	Point Estimate
μ	$\bar{x}$
p	$\hat{p}$
σ	s

Think back to the beginning of the semester where we talked about taking a random sample. Our goal was to take a sample that was representative of the population. We could then calculate a sample statistic in order to estimate the parameter of interest.

Let's visualize this as a game of darts. In a game of darts, players throw darts at a dart board and try to hit the bullseye on the board. Let the bullseye of the dart board represent the target parameter. Each dart thrown at the board will represent a point estimate. Remember that the point estimate is simply the sample statistic calculated from a random sample. We want the point estimates to come close to the target parameter and we would like them to have low variability. If the point estimates calculated from the samples are centered about the target parameter, we call them unbiased. If the point estimates systematically miss the target parameter and are centered elsewhere, we call them biased. Point estimates with low variability will be clustered close together. Ideally, we would like point estimates to be unbiased and have low variability.



8.1.2 INTERVAL ESTIMATES

An interval estimate is an interval (range) of values around our point estimate that is likely to contain our target parameter. These estimates are more useful than point estimates because they help us gauge the accuracy of a point estimate. From our knowledge of probability, we use sampling distributions to construct these intervals.

A confidence interval (CI) is an interval containing the most plausible values for a parameter based on the results of a random sample from a population. The **general formula** for this type of interval is:

$$\text{point estimate} \pm \underbrace{(\text{critical value}) \times (\text{standard error})}_{\text{Margin of error}}$$

The **margin of error** (E or MOE) measures the accuracy of the point estimate being used to estimate a parameter.

The **critical value** is the number of standard errors that we add and subtract from the point estimate to achieve a desired confidence level. This critical value will come from a probability distribution.

Every confidence interval has an associated **confidence level**, which is the expected proportion of confidence intervals that will contain the value of the target parameter if a large number of different samples was obtained.

The confidence level is denoted $(1 - \alpha)100\%$ and is most commonly chosen to be 90% (with $\alpha = 0.10$), 95% (with $\alpha = 0.05$) or 99% (with $\alpha = 0.01$).

8.2 CONFIDENCE INTERVALS FOR POPULATION PROPORTIONS

Formula 8.1: Confidence Interval for Population Proportion

For a sample of size n and a sample proportion $\hat{p}$, the corresponding $(1 - \alpha)100\%$ **confidence interval for p** , the population proportion, is given by:

$$\hat{p} \pm (z_{\alpha/2}) \left(\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right)$$

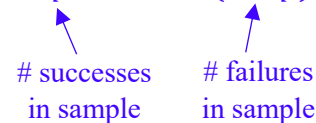
where $z_{\alpha/2}$ is the critical value that puts an area of $\alpha/2$ into the right tail of the standard normal distribution.

Note that, since p is unknown, we substitute $\hat{p}$ for p to estimate standard error: $\hat{\sigma}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

8.2.1 CONDITIONS REQUIRED FOR INFERENCE

Two conditions must be satisfied in order for us to make inference using confidence intervals:

- (1) We must have a random sample selected from the target population.
- (2) The distribution of $\hat{p}$ must be normally distributed: $n\hat{p} \geq 5$ and $n(1 - \hat{p}) \geq 5$



 $n\hat{p}$ → # successes in sample $n(1 - \hat{p})$ → # failures in sample

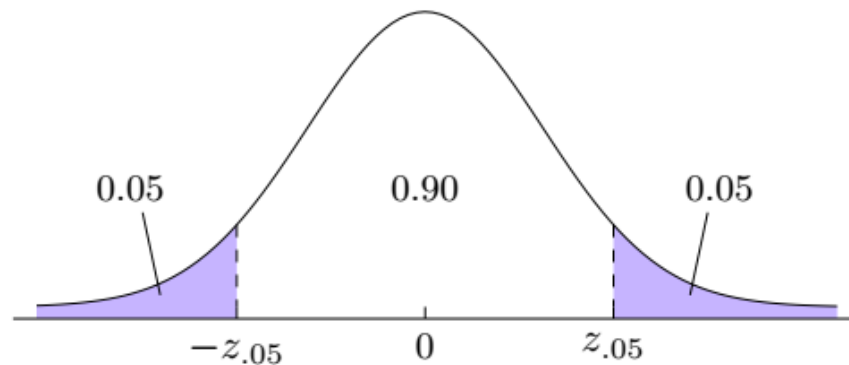
8.2.2 HOW TO FIND CRITICAL VALUES

- (1) Identify the confidence level.
- (2) Find α and divide by 2. $\alpha/2$ is the area in each tail of the standard normal distribution.
- (3) Find the corresponding z-score, $z_{\alpha/2}$. Then $-|z_{\alpha/2}|$ and $|z_{\alpha/2}|$ correspond to the $\alpha/2$ area to the lower and upper tails, respectively.

For example, if you wish to find a 90% confidence interval, the value of α is $1 - 0.90 = 0.1$. Then the corresponding critical value to find is:

$$z_{\alpha/2} = z_{0.1/2} = z_{0.05}.$$

You can find this value using the standard normal table or the invNorm function on your calculator.



$$\text{invNorm}(1 - 0.05, 0, 1) = 1.645$$

Be careful! If you do not subtract $\alpha/2$ from 1, you will get the negative of your critical value. Your critical values for confidence intervals should always be positive.

Example 8.1: Financial Aid

A university dean is interested in determining the proportion of students who receive some sort of financial aid. Rather than examine the records for all students, the dean randomly selects 200 students and finds that 118 of them are receiving financial aid.

- A. Identify the population parameter we wish to estimate.

The true proportion of students on financial aid at this university.

- B. State and check the conditions to create a confidence interval to estimate the parameter.

Condition (1): The sample must be randomly chosen. This is stated in the problem.

Condition (2): The sampling distribution for $\hat{p}$ must be normally distributed. This is true because $200 \left(\frac{118}{200} \right) = 118 \geq 5$ and $200 \left(1 - \frac{118}{200} \right) = 82 \geq 5$.

- C. Use a 95% confidence interval to estimate the true proportion of students on financial aid.

$$\hat{p} = \frac{118}{200} = 0.59$$

$$\alpha = 0.05$$

$$\text{Critical Value: } z_{\alpha/2} = z_{0.025} = 1.96$$

$$\text{CI: } 0.59 \pm 1.96 \sqrt{\frac{0.59(1-0.59)}{200}} \Rightarrow 0.59 \pm 0.0682 \Rightarrow [0.5218, 0.6582]$$

8.2.3 INTERPRETING THE CONFIDENCE LEVEL

The **confidence level** $(1 - \alpha)100\%$ is the percentage of **possible samples** from the population that are likely to result in confidence intervals containing the true parameter value.

For example, for a 95% confidence interval for population proportion...

- Approximately **95%** of all the possible samples of size n that we could take from the population will contain the **true population proportion**.
- Approximately **5%** of the possible samples of size n we could take will not contain the true population proportion.
- There is no way to know whether a particular confidence interval is one that contains the parameter or not!

A confidence level is not the probability that the population parameter falls inside our interval. (Why not?)

We calculate probabilities for random variables; however, the population parameter is **not** random. It is an unknown but **fixed** value.

8.2.4 INTERPRETING CONFIDENCE INTERVALS

We can use the following structure to guide us in writing sentences of interpretation for confidence intervals.

I am / We are 95 % confident that the **population** proportion
(confidence level * 100) (parameter)

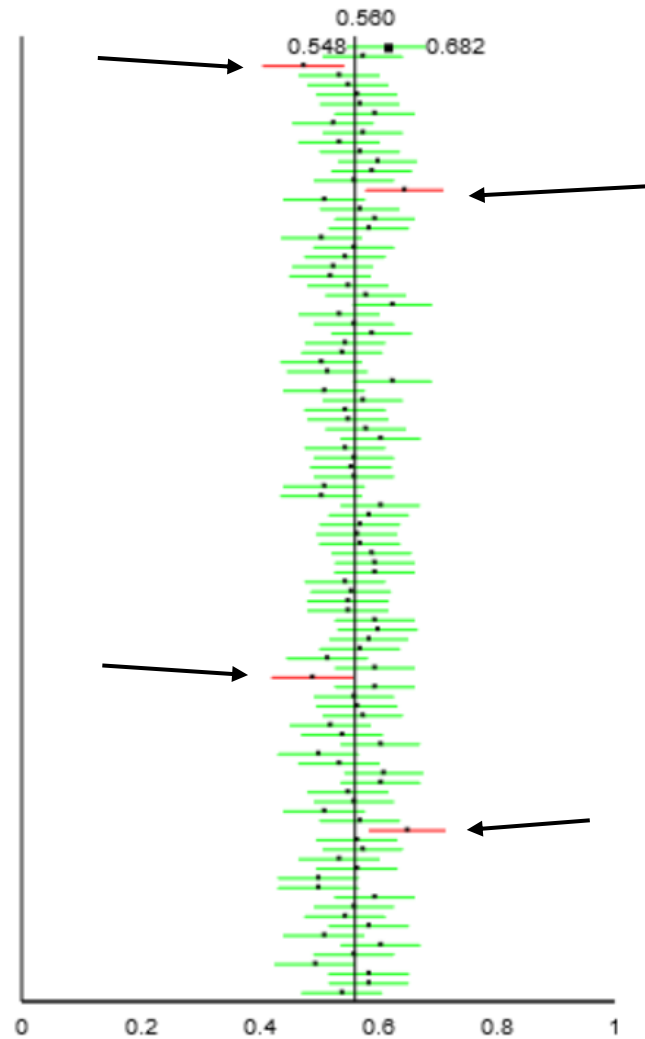
of all students who receive financial aid lies between 0.5218 and 0.6582.
(context of the problem) (lower bound) (upper bound)

8.2.5 VISUALIZING CONFIDENCE INTERVALS

Suppose that we know from the registrar's records in Example 8.1 that the **true** proportion of students receiving some sort of financial aid is 0.56. If we take 100 samples of size $n = 200$ from the population of university students, we will obtain a different confidence interval for each of them.

In this visualization, the center vertical line represents the true population proportion of 0.56, while the horizontal lines span the lengths of the 100 different confidence intervals. The center of each interval is at the point estimate $\hat{p}$, which varies from sample to sample. All but four (96%) of these intervals (look for the arrows) contain p .

In the long run, 95% of all samples of size 200 from this population will give a confidence interval that contains the true proportion p .



Give it a try! You can experiment with this simulation at:

<http://www.rossmanchance.com/applets/ConfSim.html>

Example 8.2: Financial Aid (continued)

A university dean is interested in determining the proportion of students who receive some sort of financial aid. Rather than examine the records for all students, the dean randomly selects 200 students and finds that 118 of them are receiving financial aid.

- A. Find a 90% confidence interval to estimate the true proportion of students on financial aid.

The only thing that changes from Example 8.1 is the critical value.

$$\hat{p} = \frac{118}{200} = 0.59 \quad \alpha = 0.1 \quad \text{Critical Value: } z_{\alpha/2} = z_{0.05} = 1.645$$

$$\text{CI: } 0.59 \pm 1.645 \sqrt{\frac{0.59(1 - 0.59)}{200}} \Rightarrow 0.59 \pm 0.0572 \Rightarrow [0.5328, 0.6472]$$

- B. Interpret this confidence interval.

We are 90% confident that the true / population proportion of students who receive financial aid is between 0.5328 and 0.6472.

- C. Interpret the confidence level.

If we took many samples of size 200 from this population and found a confidence interval for each sample, approximately 90% of the confidence intervals would contain the true proportion of students who receive financial aid.

- D. Find and interpret a 99% confidence interval to estimate the true proportion of students on financial aid.

$$\text{CI: } 0.59 \pm 2.576 \sqrt{\frac{0.59(1 - 0.59)}{200}} \Rightarrow 0.59 \pm 0.0896 \Rightarrow [0.5004, 0.6796]$$

We are 99% confident that the true / population proportion of students who receive financial aid is between 0.5004 and 0.6796.

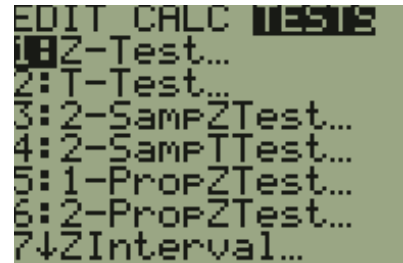
- E. Interpret the confidence level.

If we took many samples of size 200 from this population and found a confidence interval for each sample, approximately 99% of the confidence intervals would contain the true proportion of students who receive financial aid.

Calculator Tutorial 8.1: Finding Confidence Intervals for Proportion

You can find **confidence intervals for proportion** using your TI-83 or TI-84 graphing calculator.

- (1) Click the STAT button on your calculator and use the arrow keys to navigate to the TESTS menu.

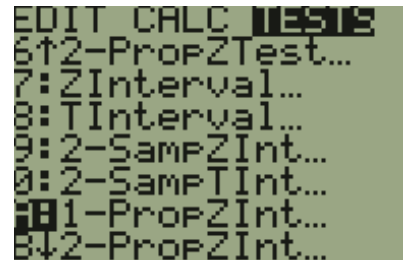


```

EDIT CALC TESTS
1:Z-Test...
2:T-Test...
3:2-SampZTest...
4:2-SampTTest...
5:1-PropZTest...
6:2-PropZTest...
7:ZInterval...

```

- (2) Use the down arrow key to navigate to item A: 1-PropZInt... Click ENTER.

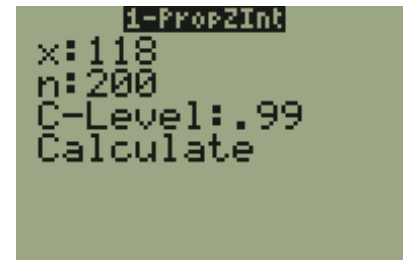


```

EDIT CALC TESTS
6:2-PropZTest...
7:ZInterval...
8:TInterval...
9:2-SampZInt...
0:2-SampTInt...
A:1-PropZInt...
B:2-PropZInt...

```

- (3) Enter the number of successes x , the sample size n , and the confidence level C-Level as a decimal number. Click CALCULATE. This function will only work if you have a whole number for the number of successes. If you only know the proportion of successes, you cannot use this function.

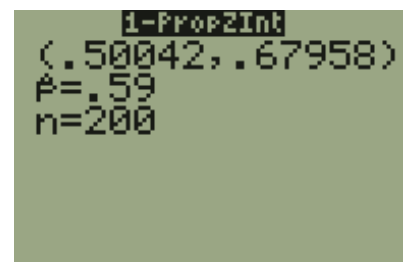


```

1-PropZInt
x:118
n:200
C-Level:.99
Calculate

```

- (4) Your calculator will display a confidence interval, the corresponding sample proportion, and the sample size.



```

1-PropZInt
(.50042,.67958)
p=.59
n=200

```

8.2.6 HOW CONFIDENCE INTERVALS BEHAVE

Increasing the **confidence level** increases the width of the confidence interval.

Increasing **sample size** decreases the width of the confidence interval.

Example 8.3: Financial Aid (continued)

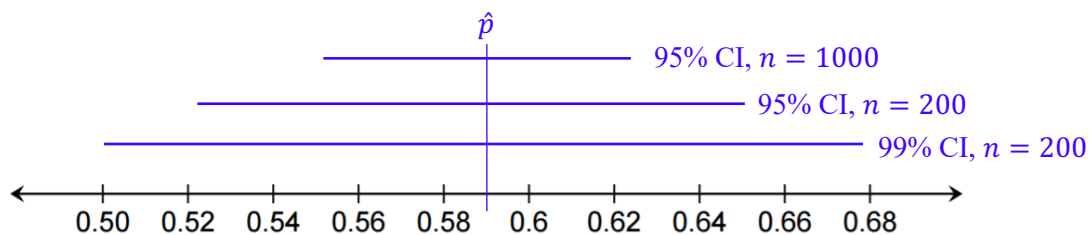
In Example 8.1, you found the 95% confidence interval to be [0.5218, 0.6582].

What happens if we increase the **confidence level** from 95% to 99%?

$$0.59 \pm 2.576 \sqrt{\frac{0.59(1-0.59)}{200}} \Rightarrow [0.5004, 0.6796], \text{ we see the confidence interval is wider}$$

What happens if we have a sample size of 1000?

$$0.59 \pm 1.96 \sqrt{\frac{0.59(1-0.59)}{1000}} \Rightarrow [0.5595, 0.6205], \text{ we see the confidence interval is narrower}$$



8.2.7 SAMPLE SIZE CALCULATIONS

If we want to create a confidence interval that has a certain margin of error and a certain confidence level, then we can estimate what the **minimum** sample size is to create such an interval.

Formula 8.2: Sample Size for Proportion

For a $(1 - \alpha)100\%$ confidence interval for population proportion, the **minimum sample size** required to achieve a desired margin of error E is:

$$n = \frac{(z_{\alpha/2})^2 \hat{p}(1 - \hat{p})}{E^2}$$

where $\hat{p}$ is an estimate of p obtained in a previous study. If no such estimate is available, we can obtain conservatively large values of n by using $\hat{p} = 0.5$.

Note: *Always* round the sample size n **up** to the nearest integer to ensure that the sample size will be sufficient to achieve the specified reliability.

Example 8.4: Home Ownership

An urban economist wishes to estimate the proportion of Americans who own their homes. What sample size should be obtained if he wishes the estimate to be within 0.02 with 90% confidence?

No $\hat{p}$ is given, so we use $\hat{p} = 0.5$, $E = 0.02$, and $z_{\alpha/2} = 1.645$.

$$n = \frac{(1.645)^2(0.5)(1 - 0.5)}{(0.02)^2} = 1691.27 \approx 1692$$

We need a sample size of 1692 Americans.

Example 8.5: Semester System

In a college student poll, it is of interest to estimate the proportion p of students in favor of changing from a quarter-system to a semester system. How many students should be polled so that we can estimate the population proportion to within 0.09 using a 99% confidence interval? *A small study has been done that found a sample proportion of 0.35.*

This time, we are given $\hat{p} = 0.35$; $E = 0.09$, and $z_{\alpha/2} = 2.576$.

$$n = \frac{(2.576)^2(0.35)(1 - 0.35)}{(0.09)^2} = 186.375 \approx 187$$

We need to poll 187 students.

MY NOTES:

8.3 CONFIDENCE INTERVALS FOR POPULATION MEANS

Similar to the way we find interval estimates for population proportions, we can find confidence intervals for **population mean**.

Recall: What are the characteristics of the sampling distribution for sample mean $\bar{x}$?

- The **mean** of the sampling distribution of $\bar{x}$ is $\mu_{\bar{x}} = \mu$.
- The **standard error** of the sampling distribution of $\bar{x}$ is $\sigma_{\bar{x}} = \sigma / \sqrt{n}$.
- The sampling distribution of $\bar{x}$ is approximately normal if either...
 - *The sample size is “large enough” ($n \geq 30$)*
 - *X is approximately normally distributed*

In practice, we will not know the population standard deviation, so we will use s , the **sample standard deviation**, to estimate σ . To estimate the **standard error** for $\bar{x}$, we will use:

$$\widehat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{n}}$$

To account for the extra variability introduced by this estimation for the standard error, rather than using the standard normal (z) distribution, critical values are calculated using a t-distribution.

Note: Your textbook shows confidence intervals for mean where the population variance, and thus σ , is known. However, σ is **very rarely known** if the population mean is unknown! In this course, we will concentrate only on the confidence interval for mean when σ is **unknown**, which is the most useful.

Some textbooks state that if $n \geq 30$, then you may use a z-distribution rather than a t-distribution to find critical values. This is not necessarily the best method. **In this course, we will always use the t-distribution for finding confidence intervals for μ .**

8.3.1 THE t -DISTRIBUTION

The t -distribution was developed by William Gosset, an English statistician who worked at the Guinness Brewery in Ireland and published his findings using the pseudonym “Student.” Gosset was interested in problems involving small samples; for example, some of his work investigated the chemical properties of barley, where sample sizes might be as few as three.

Formula 8.3: The t -Statistic

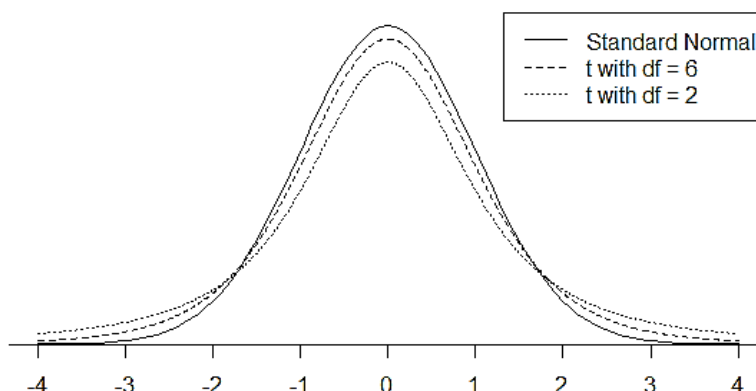
Suppose that a random sample of size n is drawn from a population that is normally distributed with mean μ and standard deviation σ . Then the **t -statistic**, denoted

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

follows the **t -distribution** with $n - 1$ **degrees of freedom**.

A **t -distribution** has several distinctive properties:

- (1) All t -distributions are bell-shaped and symmetric about zero, similar to the standard normal distribution.
- (2) Similar to the normal distribution, the t -distribution is a family of curves.
- (3) A particular t -distribution is specified by giving its degrees of freedom (df).
- (4) For inference about a single population mean, the degrees of freedom are $df = n - 1$.
- (5) The t -distribution has thicker tails and is more spread out than the standard normal distribution.
- (6) As df increases, the t -distribution approaches the standard normal distribution.



8.3.2 CONFIDENCE INTERVALS FOR POPULATION MEAN

Recall: What is the *general* formula for a confidence interval?

$$\text{point estimate} \pm \underbrace{(\text{critical value}) \times (\text{standard error})}_{\text{Margin of error}}$$

Formula 8.4: Confidence Interval for Population Mean

For a sample of size n and a sample mean $\bar{x}$, the corresponding $(1 - \alpha)100\%$ **confidence interval for μ** , the population mean, is given by:

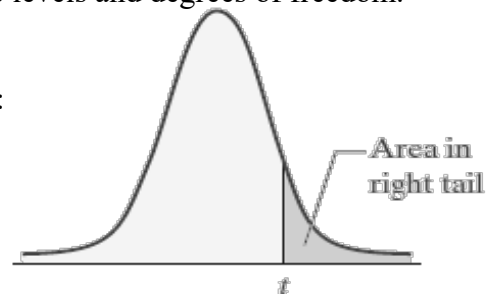
$$\bar{x} \pm (t_{\alpha/2}) \left(s / \sqrt{n} \right)$$

where $t_{\alpha/2}$ is the critical value from the t -distribution table.

The table of values used in this course for the t -distribution gives the area to the right of a specified t critical value for selected $(1 - \alpha)$ confidence levels and degrees of freedom.

Standard notation for specifying a t critical value for a $(1 - \alpha)$ confidence level and $n - 1$ degrees of freedom is:

$$t_{\alpha/2, n-1}$$



where $\alpha/2$ is the area to the right of the t -value and $n - 1$ is the degrees of freedom.

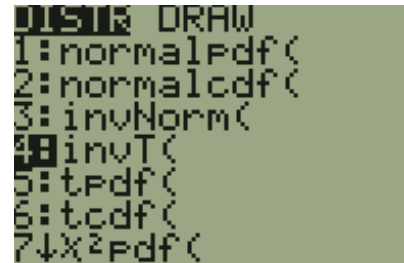
Notice that the values in the last row of the t -distribution table are identical to the standard normal z critical values for those confidence levels. This is because the larger your sample size is (and thus your degrees of freedom), the closer the t -distribution is to the standard normal distribution. Use the z -values listed in the last row when the degrees of freedom in a problem are greater than 1,000.

If the desired degrees of freedom are not listed in the table, choose the closest number in the df column.

Calculator Tutorial 8.2: Finding t Critical Values

Some calculators, including the TI-84, have the ability to find **t critical values** for a specified confidence level.

- (1) Click 2ND + VARS and use the down arrow key to navigate to item 4: invT(. Click ENTER.



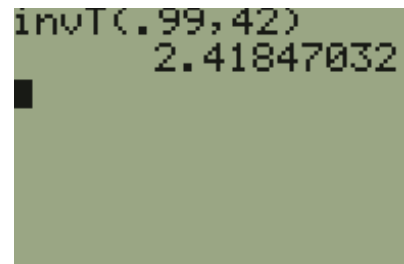
```
DISTR DRAW
1:normalpdf(
2:normalcdf(
3:invNorm(
4:invT(
5:tpdf(
6:tcdf(
7:χ²pdf(
```

- (2) Type the area to the **left** of the desired t -value and the degrees of freedom. Use the arrow keys to navigate to Paste and click ENTER.



```
invT
area: .99
df: 42
Paste
```

- (3) The formula will be displayed on your main screen. Click Enter again to compute the t critical value.



```
invT(.99,42)
2.41847032
```

This example computes $t_{.01,42}$.

Note: The TI-83 calculator does not have the invT(function pre-installed. You can, however, program the function manually. Learn how by following the steps at:

<https://www.youtube.com/watch?v=5Ft5eZVJtPk>

Example 8.6: Finding t Critical Values

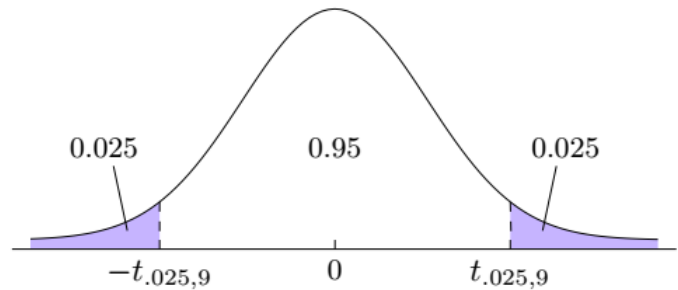
Use the t -table or your calculator to find the specified t critical values.

- A. What is the t critical value for a 95% confidence interval for μ from a sample of 10 observations?

$$\alpha = 0.05 \Rightarrow \alpha/2 = 0.025$$

$$df = n - 1 = 10 - 1 = 9$$

$$t_{.025,9} = 2.262$$

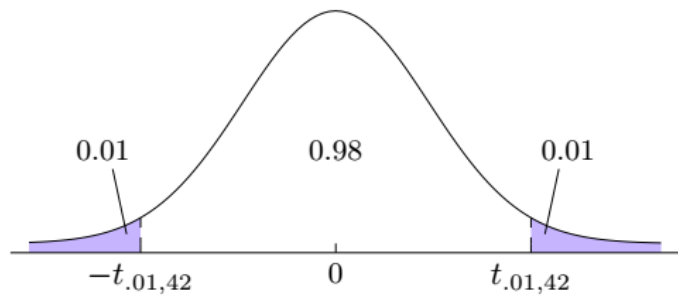


- B. What is the t critical value for a 98% confidence interval for μ from a sample of size 43?

$$\alpha = 0.02 \Rightarrow \alpha/2 = 0.01$$

$$df = n - 1 = 43 - 1 = 42$$

$$t_{.01,42} \approx t_{.01,40} = 2.423$$



8.3.3 CONDITIONS REQUIRED FOR INFERENCE

Two conditions must be satisfied in order for us to make inference using confidence intervals:

- (1) We must have a random sample selected from the target population.
- (2) The distribution of $\bar{x}$ must be normally distributed, which occurs if either...
 - The sample size is “large enough” ($n \geq 30$).
 - X is approximately normally distributed.

For **large** samples of size n , a confidence interval can be computed for **non-normal** populations. This is because the procedure for constructing confidence intervals is robust; that is, it is accurate despite minor departures from normality.

For **small** samples of size n , confidence intervals are appropriate **only** if the population is approximately normal. You can check this condition using a normal probability plot to verify that the data come from a normally distributed population and a box plot to verify that there are no outliers.

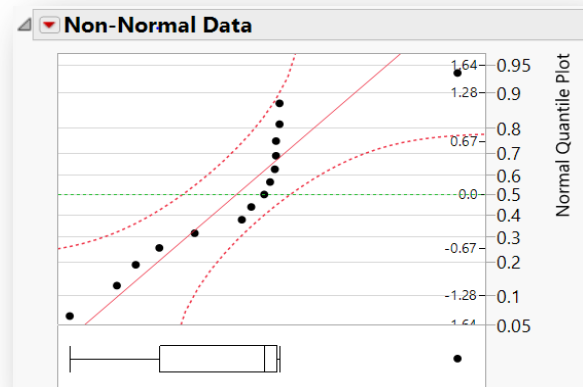
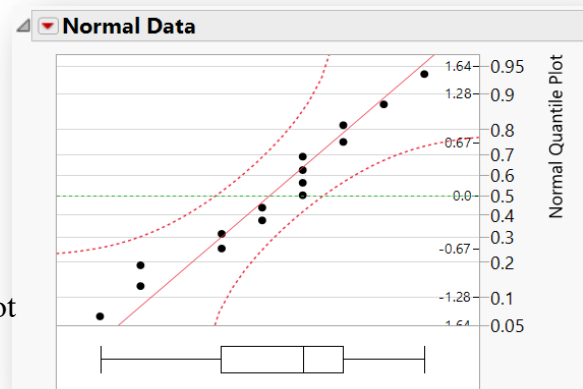
NORMAL PROBABILITY PLOT

Consider the following output produced for two different data sets in JMP.

If the data come from a normally distributed population, the points on the normal probability plot should be roughly linear and fall within the curved bounds.

There should also be no outliers shown by the boxplot.

If the points on the normal probability plot are **not** roughly linear or fall outside the curved bounds, or if outliers or strong skewness are present in the boxplot, a confidence interval using t critical values is not an appropriate method.



Example 8.7: Sleep Time for College Students

Based on a random sample of 52 full-time college students, a 99% confidence interval for the mean number of hours that full-time college students sleep per night is [6.73, 7.67] hours.

- A. What is the population of interest?

Full-time college students

- B. What is the parameter of interest?

$\mu = \text{true mean number of hours that full-time college students sleep per night}$

- C. What is the sample?

The 52 full-time college students in the study

- D. Have the conditions necessary for constructing a confidence interval been met?

(1) Random sample – stated in the problem

(2) The sampling distribution for $\bar{x}$ is approximately normal since $n = 52 \geq 30$

- E. Interpret the confidence interval in context of the problem.

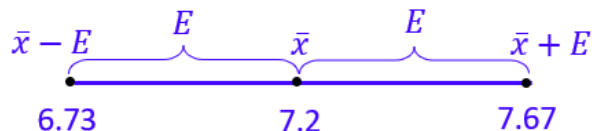
We are 99% confident that the true mean hours college students sleep per night is between 6.73 and 7.67 hours.

- F. Based on the interval, is it possible that the mean number of hours full-time college students sleep per night is 8 hours? Is it likely?

It is possible that the interval we constructed does not capture the true mean. However, it is unlikely since 8 falls outside the 99% confidence interval.

- G. What is the margin of error for this interval? (Hint: It's half the width of the interval.)

$$E = \frac{7.67 - 6.73}{2} = \frac{0.94}{2} = 0.47$$



- H. What is the point estimate for the population mean?

$$\bar{x} = 6.73 + 0.47 = 7.2$$

Example 8.8: Employee Age

A group of 19 randomly selected employees has a mean age of 22.4 years with a standard deviation of 3.8 years. Assume that the population of ages has a normal distribution. You wish to construct a confidence interval to estimate μ , the true mean age of all employees at this business.

A. What is the population of interest?

The employees at this business

B. What is the parameter of interest?

μ = mean age of all the employees at this business

C. What is the sample?

The 19 randomly selected employees

D. State and check the necessary conditions to create a confidence interval to estimate the true mean age of employees.

(1) Random sample – stated in the problem

(2) The sampling distribution for $\bar{x}$ is approximately normal since X is normal.

E. Construct and interpret a 99% confidence interval to estimate the true mean age of employees.

$$\bar{x} = 22.4, s = 3.8, n = 19$$

$$\alpha = 0.01 \Rightarrow \alpha/2 = 0.005, df = 19 - 1 = 18, t_{.005, 18} = 2.878$$

$$CI: \bar{x} \pm t_{.005, 18} \left(\frac{s}{\sqrt{n}} \right) \Rightarrow 22.4 \pm (2.878) \left(\frac{3.8}{\sqrt{19}} \right) \Rightarrow [19.891, 24.909] \text{ years}$$

We are 99% confident that the true mean age of employees at this business is between 19.891 and 24.909 years old.

F. Interpret the confidence level.

If we were to take many samples of size 19 from this population and form a confidence interval for each sample, approximately 99% of the intervals would contain the true mean age of employees.

Calculator Tutorial 8.3: Finding Confidence Intervals for Mean

You can find **confidence intervals for mean** using your TI-83 or TI-84 graphing calculator.

- (5) Click the STAT button on your calculator and use the arrow keys to navigate to the TESTS menu.

```

EDIT CALC TESTS
1:Z-Test...
2:T-Test...
3:2-SampZTest...
4:2-SampTTest...
5:1-PropZTest...
6:2-PropZTest...
7:ZInterval...

```

- (6) Use the down arrow key to navigate to item 8: TInterval... Click ENTER.

```

EDIT CALC TESTS
3:2-SampZTest...
4:2-SampTTest...
5:1-PropZTest...
6:2-PropZTest...
7:ZInterval...
8:TInterval...
9:2-SampZInt...

```

- (7) Use the right arrow key and click ENTER to select Stats. Enter the sample mean $\bar{x}$, the sample standard deviation S_x , the sample size n , and the confidence level C-Level as a decimal number. Click CALCULATE.

```

TInterval
Inpt:Data Stats
x:22.4
Sx:3.8
n:19
C-Level:.99
Calculate

```

- (8) Your calculator will display a confidence interval for mean with the summary statistics you entered.

```

TInterval
(19.891,24.909)
x=22.4
Sx=3.8
n=19

```

8.3.4 SAMPLE SIZE CALCULATIONS

Similar to the manner in which we determine the sample size required to achieve a desired margin of error for population proportion, we can do the same for mean. However, recall that the margin of error for a confidence interval for mean is

$$E = t_{\alpha/2, n-1} \left(\frac{s}{\sqrt{n}} \right).$$

The critical value, $t_{\alpha/2, n-1}$, depends on the sample size n , which is the quantity we wish to estimate! In order to estimate the sample size for a given margin of error, we must estimate the t critical value using a standard normal z -value. We will also need an estimate for s which will often come from a previous study or pilot study.

Formula 8.5: Sample Size for Mean

For a $(1 - \alpha)100\%$ confidence interval for population mean, the **minimum sample size** required to achieve a desired margin of error E is:

$$n = \left(\frac{z_{\alpha/2} \cdot s}{E} \right)^2$$

Note: Always round the sample size n **up** to the nearest integer to ensure that the sample size will be sufficient to achieve the specified reliability.

Example 8.9: How Long Would You Like to Live?

In a survey of 35 adult Americans, it was found that the mean age (in years) that people would like to live is 87.9 years with a standard deviation of 15.5 years. How many adult Americans would need to be surveyed to estimate the true mean age people would like to live within 2 years and with 95% confidence?

$$s = 15.5, E = 2, z_{\alpha/2} = 1.96$$

$$n = \left(\frac{z_{\alpha/2} \cdot s}{E} \right)^2 = \left(\frac{1.96 \cdot 15.5}{2} \right)^2 = 230.7361 \approx 231 \text{ Americans}$$

MY NOTES:

PRACTICE EXERCISES FOR CHAPTER 8

1. One frequently used test to measure reading ability of children is the DRP, or Degree of Reading Power. It is known that DRP scores are normally distributed. A researcher suspects that the mean score μ of all third-graders in Henrico County Schools is different from the national mean of 32. To test her suspicion, she administers the DRP to a random sample of 22 Henrico County third-grade students. Their scores are recorded in the following table.

40	26	39	17	42	18	24	43	46	27	19
47	19	26	37	34	15	45	41	39	31	46

- (a) Find the t critical value for a 90% confidence interval for the true mean μ of DRP scores in Henrico County Schools.

$$\alpha = 1 - 0.90 = 0.1 \Rightarrow \alpha/2 = 0.05 \qquad n = 22 \Rightarrow df = 22 - 1 = 21$$

$$t_{.05,21} = 1.72$$

- (b) State and verify the conditions for the confidence interval.

(1) Sample must be randomly selected from the population – stated in the problem.

(2) $\bar{x}$ must be approximately normal – true since X is stated to be normal.

- (c) Compute and interpret the 90% confidence interval for the mean DRP score in Henrico County Schools. (Hint: First find sample mean and sample standard deviation.)

$$\bar{x} = 32.77, s = 10.86$$

$$\bar{x} \pm t_{.05,21} \left(\frac{s}{\sqrt{n}} \right) \Rightarrow 32.77 \pm 1.72 \left(\frac{10.86}{\sqrt{22}} \right) \Rightarrow [28.79, 36.75]$$

- (d) Use the confidence interval you constructed in Part (c) to comment on whether you agree with the researcher's suspicion. Explain your reasoning clearly.

It is reasonable to disagree with her suspicion. Since 32 is in the interval, we don't have evidence that the true mean DRP score for Henrico County 3rd-graders is different than 32.

- (e) Suppose that the distribution of DRP scores was unknown. What would the researcher have needed to do differently in the study to construct a valid confidence interval?

She would have needed to collect DRP scores from a larger sample of at least 30 students.

2. In a July 2001 research note, the U.S. Department of Transportation reported the results of the National Occupant Protection Use Survey. One focus of the survey was to determine the level of cell phone use by drivers while they are driving a motor passenger vehicle. Data collected by observers at randomly selected intersections across the country revealed that, in a sample of 1165 drivers, 35 were using their cell phones.

- (a) What is the point estimate for p , the true driver cell phone rate?

$$\hat{p} = \frac{35}{1165} \approx 0.03$$

- (b) Check and verify the two conditions for a 98% confidence interval.

(1) Sample must be randomly selected from the population – stated in the problem.

(2) $n\hat{p} = 1165(.03) = 34.95 \geq 5$ and $n(1 - \hat{p}) = 1165(.97) = 1130.05 \geq 5$
 $\Rightarrow$ The distribution of $\hat{p}$ is approximately normal.

- (c) Compute a 98% confidence interval for the proportion of drivers that are using their cell phones.

$$\hat{p} \pm z_{.01} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \Rightarrow 0.03 \pm 2.33 \sqrt{\frac{.03(.97)}{1165}} \Rightarrow [0.0184, 0.0417]$$

- (d) Interpret the interval that you found in Part (c).

We are 98% confident that the true proportion of drivers using their cell phones while driving is between 0.0184 and 0.0417.

- (e) Interpret the confidence level used to construct the confidence interval.

If we found many samples of 1165 and constructed confidence intervals for population proportion for each, we would expect approximately 98% of the intervals to contain the true population proportion p .

3. In a survey of 40 Clemson students, it was found that the mean age (in years) when they would like to get married is 27.4 years old with a standard deviation of 6 years. How many Clemson students would need to be surveyed to estimate the mean age at which Clemson students would like to get married within 1.5 years with 90% confidence?

$$n = \left(\frac{z_{.05} \cdot s}{E} \right)^2 = \left(\frac{1.645 \cdot 6}{1.5} \right)^2 = 43.2964 \approx 44 \text{ students}$$

4. A magazine company is planning to survey customers to determine the proportion who will renew their subscription for the coming year. The magazine wants to estimate the population proportion with 95% confidence and a margin of error equal to 0.03. What is the minimum sample size required?

$$n = \frac{(z_{.025})^2 \hat{p}(1 - \hat{p})}{E^2} = \frac{(1.96)^2 (0.5)^2}{(.03)^2} = 1067.111 \approx 1068 \text{ customers}$$

SECTION 9: HYPOTHESIS TESTING

HAWKES CHAPTER 10

Another important method for making statistical inference about a population is the **hypothesis test**.

In conducting a statistical hypothesis test, the goal is to determine whether the data collected support certain claims or predictions being made about the population parameter. For example, we could use hypothesis tests to answer questions such as...

- Is the amount of pollutants released by an industrial smokestack **higher** than the EPA mandated limit of 5 parts per million (PPM)?
- Has the public opinion about global warming **changed** over time?
- Is the average tax refund **lower** than the claim made by the IRS?

Every hypothesis test has an associated significance level known as α , which is related to the value of alpha in a $(1 - \alpha)100\%$ confidence interval.

For example, Reese's pieces are supposed to be 50% orange, 25% yellow, and 25% brown. An assembly line worker alerts the quality control manager at her plant that one machine appears to be producing too few orange candies. The manager will need to decide if the machine is working properly or if it needs maintenance. We can use a hypothesis test to determine if the machine is producing the correct percentage of orange Reese's pieces, or if the percentage of orange is too low. Once we have determined if the percentage of orange candies being produced is correct or not, the manager can make an informed decision about the machine.

We will consider 3 hypothesis tests in this course.

1. Hypothesis test for a proportion
2. Hypothesis test for a mean
3. Hypothesis test for a variance/standard deviation

9.1 HYPOTHESIS TESTING: BASIC STEPS (HATS)

Textbooks often list several steps to take when conducting a hypothesis test. We have reduced the number of steps down to four. You can easily remember the 4 steps using the word **HATS**

H: Hypotheses: State the null and alternative **hypotheses** using population parameters.

A: Assumptions: Certain **assumptions** or conditions must be present for the hypothesis test to be valid, similar to confidence intervals.

T: Test: Collect evidence (sample data) to test the hypothesis. Analyze the data to assess the plausibility of the hypothesis.

S: Summarize: Use the results from the test step to decide to reject or fail to reject the null hypothesis. Write a **summary** of the conclusion based upon your analysis in terms of the alternative hypothesis using the population's context and parameters.

Let's look at these steps in a little more detail.

9.1.1 HYPOTHESES

Hypotheses (the plural of **hypothesis**) are statements regarding the nature of the population parameter about which you are making inference. Every hypothesis test has two hypotheses about the population parameter:

- The null hypothesis (H_0) is the claim we wish to test. It states that the population parameter is **equal** to a particular value, and it is *assumed to be true* unless sufficient evidence is gathered to reject it. It usually represents the **status quo** and is a statement of *no change* or *no effect*.
- The alternative hypothesis (H_1 or H_a) is the **research hypothesis**. It is the statement we wish to find evidence to support as being true. It states that the population parameter falls in some **alternative range** of values.

Hypotheses may be structured in one of three ways.

Left-Tailed Test	Right-Tailed Test	Two-Tailed Test
H_0 : parameter = value	H_0 : parameter = value	H_0 : parameter = value
H_1 : parameter <u><</u> value	H_1 : parameter <u>></u> value	H_1 : parameter <u>≠</u> value

Left- and right-tailed tests are referred to as **one-tailed** or **one-sided tests**, and sometimes a two-tailed test is referred to as a **two-sided test**.

The alternative hypothesis should express a challenge to the null hypothesis. It is usually what we had in mind before we decided to collect data and can be thought of as the situation for which we wish to find evidence. We should always specify the hypotheses we wish to test *before* we look at the data.

If you do not have a specific direction in mind when conducting a hypothesis test, use a two-sided test.

Example 9.1: Stating Hypotheses

Determine the null and alternative hypotheses in each of the following scenarios.

- A. Nationally, the average price of an existing single-family home is \$245,700. A local realtor believes the average price of homes is higher in Clemson.

Let μ be the average price of all homes in Clemson.

$$H_0: \mu = \$245,700 \quad (\text{no difference})$$

$$H_1: \mu > \$245,700 \quad (\text{right-tailed test})$$

- B. In 2005, 70% of Americans reported that they were in favor of the death penalty for a person convicted of murder. A researcher wonders if the percentage of Americans in favor of the death penalty is different today.

Let p be the population proportion of Americans in favor of the death penalty today.

$$H_0: p = 0.70$$

$$H_1: p \neq 0.70 \quad (\text{two-tailed test})$$

- C. Reese's pieces are supposed to be 50% orange, 25% yellow, and 25% brown. An assembly line worker alerts the quality control manager at her plant that one machine appears to be producing too few orange candies.

Let p be the true proportion of orange Reese's pieces the machine produces.

$$H_0: p = 0.50$$

$$H_1: p < 0.50 \quad (\text{left-tailed test})$$

Note: When defining the parameter to be tested, you need to somehow refer to the population. For example, you can define the parameter as interest as the **population** proportion, the **true** proportion, or the proportion of **all** individuals with the characteristic of interest.

9.1.2 ASSUMPTIONS/CONDITIONS

Similar to when we worked with confidence intervals, certain conditions must be present in order for a hypothesis test to be valid. Each type of hypothesis test will have its own set of conditions, which we will see as we work through examples. As with confidence intervals, assumptions should always be stated and checked.

9.1.3 TEST

We test our hypothesis by collecting sample data, which serves as evidence for our alternative hypothesis. To determine whether this evidence is sufficient or not, we compute a test statistic, which is a statistic computed from our sample. It may be a z-score, a *t*-score, or a chi-square value (more on this later). The test statistic measures how close our data is to the value in the hypothesis statements. In other words, is our data consistent with the null hypothesis?

We determine what constitutes “consistent with the null hypothesis” by considering the sampling distribution of the **test statistic** under the assumption that H_0 is true.

9.1.4 SUMMARIZE

If the sample data are **consistent** with the null hypothesis H_0 , we say that we do not reject or fail to reject the null hypothesis.

If the sample data are **inconsistent** with the null hypothesis and **supportive of the alternative** hypothesis H_1 , we reject the null hypothesis in favor of the alternative.

You use one of the following two approaches to make your decision about the null hypothesis.

Approach 1: Rejection Region Approach

Approach 2: P-Value Approach

Once you have made a decision to reject or fail to reject the null hypothesis, write a **concluding statement** about the parameter in context of the problem. You should always write your conclusion in terms of the **alternative hypothesis** stating that you have sufficient evidence to suggest the alternative hypothesis or insufficient evidence to suggest the alternative.

Below is a template you can use when constructing your concluding statement.

At the _____ significance level, there is _____ evidence to suggest
(alpha) (sufficient / insufficient)
that _____.
(alternative hypothesis in context of the problem)

Note: We never prove the null hypothesis to be true! Why? Let's find out in the next section.

9.2 TYPES OF ERRORS

There are two possible decisions that we can make when we perform a hypothesis test:

- If the sample data are **consistent** with the null hypothesis, we...
 - Do not reject H_0 .
 - State that there is insufficient evidence to support H_1 .
- If the sample data are **inconsistent** with the null hypothesis, we...
 - Reject H_0
 - State that there is sufficient evidence to support H_1 .

However, our decision is based upon a *sample*, which contains incomplete information about our population, so we may unknowingly commit an **error** and make an incorrect decision regarding the null hypothesis. There are **two** possible errors that can result from hypothesis testing.

- A Type I error occurs if we **reject** the null hypothesis when it is actually true. The probability of committing this type of error is α (the significance level of the experiment).
- A Type II error occurs if we **do not reject** the null hypothesis when the alternative hypothesis is true. The probability of committing this type of error is β .

If you increase the probability of committing the first type of error, you will decrease the probability of committing the second type, and vice versa.

We can summarize the possible decisions we can make about H_0 and the possible errors we could commit in the following table.

		<i>Reality</i>	
		H_0 True	H_1 True
<i>Decision</i>	Do Not Reject H_0	<i>Correct</i>	<i>Type II Error</i>
	Reject H_0	<i>Type I Error</i>	<i>Correct</i>

You can think of the hypothesis testing procedure as analogous to a **jury trial**, in which the hypotheses of interest are

$$H_0: \text{defendant is innocent} \quad \text{and} \quad H_1: \text{defendant is guilty.}$$

The null hypothesis of innocence is *assumed true* unless there is sufficient evidence to support the alternative hypothesis of guilt beyond a reasonable doubt.

There are two possible verdicts that a jury could arrive at regarding the defendant:

- Reject H_0 (guilty)
- Fail to reject H_0 (not guilty)

Just as a defendant is never declared innocent in a jury trial, we **never** accept H_0 as true in a hypothesis test.

It is possible that the jury could make a mistake. If we relate to the two possible errors the jury could commit to the errors that can be committed in hypothesis testing, Type I and Type II errors can be described as the following.

- **Type I Error:** An innocent person is convicted.
- **Type II Error:** A guilty person is acquitted.

Example 9.2: Safe Drinking Water

The Environmental Protection Agency (EPA) declares drinking water to be unsafe if the concentration of lead is greater than 15 parts per billion (ppb). Let μ denote the true mean concentration of lead from a community water source. A water system that monitors lead levels might use the lead level measurements from a sample of water specimens to test the hypotheses

$$H_0: \mu = 15 \text{ ppb (water is safe)} \quad \text{and} \quad H_1: \mu > 15 \text{ ppb (water is unsafe).}$$

If a water source is found to violate the EPA standard, it can no longer be used as a community water source.

- A. Describe a Type I error (rejecting a true H_0) in context of this situation.

Concluding that the water is unsafe, when in reality, the lead levels are actually acceptable.

- B. What are the consequences of making a Type I error in context of this situation?

Unnecessarily eliminating a community water source.

- C. Describe a Type II error (not rejecting a false H_0) in context of this situation.

Not concluding the water is unsafe, when in reality, the lead levels are too high.

- D. What are the consequences of making a Type II error in the context of this situation?

The public unknowingly drinking unsafe water, putting them at risk of health problems due to lead consumption (e.g. high blood pressure, hearing loss, kidney damage, etc.)

To reduce the chances of committing a Type II error when the consequences are significant, one can sacrifice some significance. Setting $\alpha = 0.10$ instead of a more restrictive 0.05 or 0.01 makes rejecting H_0 more likely, as some test statistic values that would not be rejected with a smaller α would be rejected at this higher significance level.

Example 9.3: Effective Medicine

A medical researcher is evaluating the safety of a drug that is suspected to cause serious side effects among patients. Suppose that, if these serious side effects occur in more than 1% of patients, the drug will be recalled. Otherwise, the serious side effects are considered to be anomalous and not caused by the drug. The researcher collects data on a large sample of patients taking the drug to test the hypothesis

$$H_0: p = 0.01 \quad \text{against} \quad H_1: p > 0.01.$$

- A. Describe a Type I error in context of this situation.

Concluding that we have evidence that the drug causes serious side effects in patients, when in reality, it does not cause serious side effects.

- B. What are the consequences of making a Type I error in context of this situation?

The drug is recalled when it does not need to be. Patients potentially miss out on a beneficial drug.

- C. Describe a Type II error in context of this situation.

Concluding that we do not have evidence that the drug causes serious side effects, when in reality, it does cause serious side effects.

- D. What are the consequences of making a Type II error in the context of this situation?

The drug is not recalled and causes serious side effects among patients.

MY NOTES:

9.3 HYPOTHESIS TESTS FOR PROPORTION

We will now introduce the **hypothesis test for p** , the population proportion. Just like confidence intervals for p , the test statistic in this test is based on the **standard normal distribution**. We will follow the same procedure **HATS**. Recall there are **2** approaches for hypothesis testing.

TABLE 9.1: HYPOTHESIS TESTING FOR PROPORTION - REJECTION REGION APPROACH

HYPOTHESES

- Define p in context of the problem.
- The hypotheses can be structured in one of three ways:

LEFT-TAILED TEST	RIGHT-TAILED TEST	TWO-TAILED TEST
$H_0: p = p_0$	$H_0: p = p_0$	$H_0: p = p_0$
$H_1: p < p_0$	$H_1: p > p_0$	$H_1: p \neq p_0$

where p_0 is the hypothesized value of the population proportion.

ASSUMPTIONS / CONDITIONS

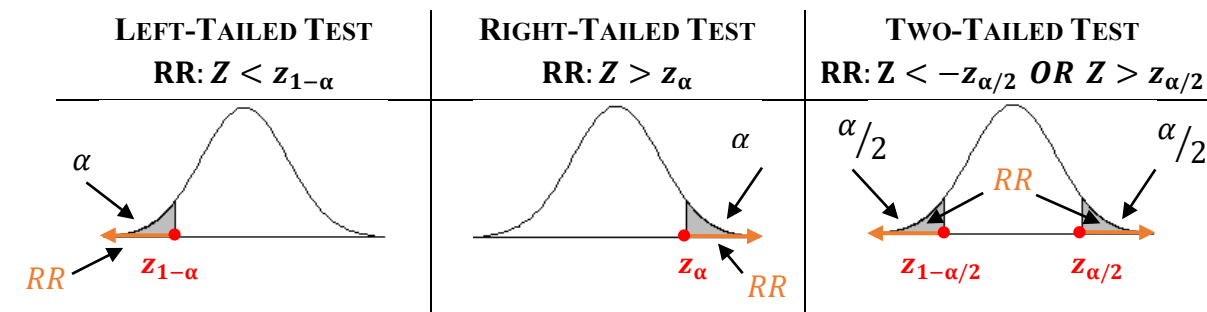
- The sample data come from a random sample or randomized experiment.
- The sampling distribution for $\hat{p}$ is approximately normal. This will be true if $np_0 \geq 5$ and $n(1 - p_0) \geq 5$.

TEST STATISTIC

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

which follows the standard normal distribution.

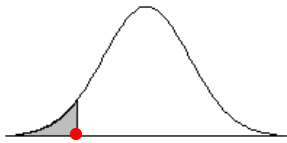
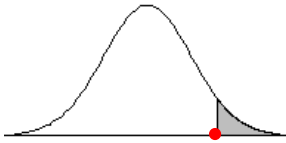
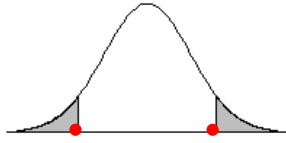
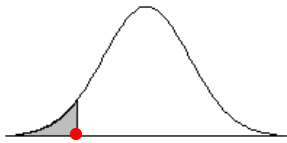
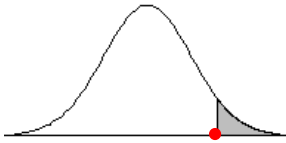
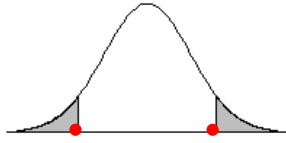
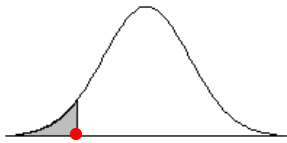
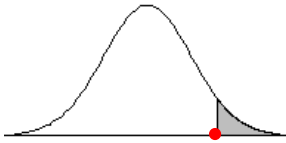
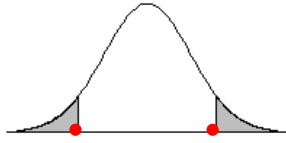
FIND REJECTION REGION BASED ON THE SIGNIFICANCE LEVEL α OF THE TEST*



* Some textbooks refer to the rejection region (RR) as the **decision rule**.

SUMMARIZE

- Reject H_0 if the test statistic lies within the rejection region.
In context of the study, state that there is sufficient evidence to support H_1 .
- Fail to reject H_0 if the test statistic does not lie in the rejection region.
In the context of the study, state that there is not sufficient evidence to support H_1 .

TABLE 9.2: HYPOTHESIS TESTING FOR PROPORTION - P-VALUE APPROACH											
<u>HYPOTHESES</u> ➤ Define p in context of the problem. ➤ The hypotheses can be structured in one of three ways: <table> <tr> <th>LEFT-TAILED TEST</th><th>RIGHT-TAILED TEST</th><th>TWO-TAILED TEST</th></tr> <tr> <td>$H_0: p = p_0$</td><td>$H_0: p = p_0$</td><td>$H_0: p = p_0$</td></tr> <tr> <td>$H_1: p < p_0$</td><td>$H_1: p > p_0$</td><td>$H_1: p \neq p_0$</td></tr> </table> where p_0 is the hypothesized value of the population proportion.			LEFT-TAILED TEST	RIGHT-TAILED TEST	TWO-TAILED TEST	$H_0: p = p_0$	$H_0: p = p_0$	$H_0: p = p_0$	$H_1: p < p_0$	$H_1: p > p_0$	$H_1: p \neq p_0$
LEFT-TAILED TEST	RIGHT-TAILED TEST	TWO-TAILED TEST									
$H_0: p = p_0$	$H_0: p = p_0$	$H_0: p = p_0$									
$H_1: p < p_0$	$H_1: p > p_0$	$H_1: p \neq p_0$									
<u>ASSUMPTIONS / CONDITIONS</u> ➤ The sample data come from a random sample or randomized experiment. ➤ The sampling distribution for $\hat{p}$ is approximately normal. This will be true if $np_0 \geq 5$ and $n(1 - p_0) \geq 5$.											
<u>TEST STATISTIC</u> $z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$ which follows the standard normal distribution.											
CALCULATE P-VALUE <table> <tr> <th>LEFT-TAILED TEST</th><th>RIGHT-TAILED TEST</th><th>TWO-TAILED TEST</th></tr> <tr> <td>  $P\text{-value} = P(Z < z_0)$ </td><td>  $P\text{-value} = P(Z > z_0)$ </td><td>  $P\text{-value} = 2 \cdot P(Z > z_0)$ </td></tr> </table>			LEFT-TAILED TEST	RIGHT-TAILED TEST	TWO-TAILED TEST	 $P\text{-value} = P(Z < z_0)$	 $P\text{-value} = P(Z > z_0)$	 $P\text{-value} = 2 \cdot P(Z > z_0)$			
LEFT-TAILED TEST	RIGHT-TAILED TEST	TWO-TAILED TEST									
 $P\text{-value} = P(Z < z_0)$	 $P\text{-value} = P(Z > z_0)$	 $P\text{-value} = 2 \cdot P(Z > z_0)$									
<u>SUMMARIZE</u> ➤ Reject H_0 if $P\text{-value} \leq \alpha$. In context of the study, state that there is sufficient evidence to support H_1. ➤ Fail to reject H_0 if $P\text{-value} > \alpha$. In the context of the study, state that there is not sufficient evidence to support H_1.											

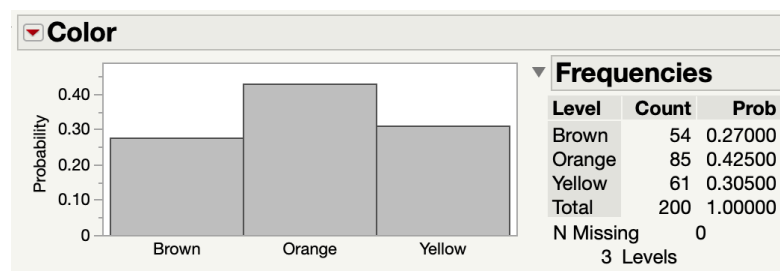
Up to and including the calculation of the test statistic, the steps of a hypothesis test are the same regardless of which approach is used. The difference in the two approaches appears where we make the decision to reject or fail to reject the null hypothesis. In the Rejection Region approach you will compare the [test statistic](#) to a [rejection region](#) that is found using the significance level, α , of the test. In the P-value approach, you [calculate a P-value](#) for your test statistic and compare it [directly](#) to the significance level, α , of the test.

The following example will illustrate these steps of the hypothesis test using both approaches. We will begin with the **H**, **A**, and **T** steps since those steps are the same for both the Rejection Region approach and the P-value approach.

Example 9.4: Reese's Pieces

Reese's pieces are supposed to be 50% orange, 25% yellow, and 25% brown. An assembly line worker alerts the quality control manager at her plant that one machine appears to be producing too few orange candies. The manager will need to decide if the machine is working properly or if it needs maintenance. You are asked to perform a hypothesis test at the significance level of $\alpha = 0.05$. You take a random sample of 200 Reese's pieces from the machine and have JMP calculate the proportion of orange candies.

Here are the results of your sample:



A. Define the appropriate **Hypotheses**.

Let p = true proportion of orange Reese's pieces the machine produces

$$H_0: p = 0.50$$

$$H_1: p < 0.50$$

B. State and verify that the proper **Assumptions** hold.

(1) Random sample – stated in the problem.

(2) $\hat{p}$ is normally distributed since

$$np_0 = 200(0.50) = 100 \geq 5 \text{ and } n(1 - p_0) = 200(1 - 0.50) = 100 \geq 5$$

C. Calculate the **Test statistic**.

$$\hat{p} = \frac{85}{200} = 0.425$$

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} = \frac{0.425 - 0.50}{\sqrt{\frac{0.50(1 - 0.50)}{200}}} \approx -2.12$$

Now that we have our test statistic, we move on to the **S** step: **Summarize**. We will use the test statistic to determine if we will reject or fail to reject the null hypothesis and write a final concluding statement.

9.3.1 APPROACH 1: REJECTION REGION

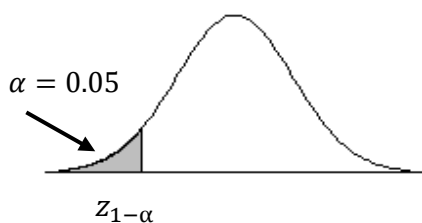
In the Rejection Region Approach (decision rule) we find the test statistic that separates a decision to reject the null hypothesis from a decision to fail to reject the null hypothesis. The test statistics that fall into the rejection region represent those that put the sample statistic too far away from the hypothesized value to believe that the null hypothesis is true.

In a test for a proportion, we will use the [Standard Normal Distribution](#) to find the rejection region. We need to find a location or locations along the horizontal axis that correspond(s) to putting α into the appropriate tail(s). This/these location(s) will be our critical value(s).

- For a left-tail test, any test statistic [less than or equal](#) to the critical value will be in the rejection region causing us to reject the null hypothesis.
- For a right-tail test, any test statistic [greater than or equal](#) to the critical value will be in the rejection region causing us to reject the null hypothesis.
- For a two-tail test, we put $\alpha/2$ into the left and right tails and find the associated critical values for each tail. Any test statistic [less than or equal](#) to the left tail critical value OR [greater than or equal](#) to the right tail critical value will be in the rejection region causing us to reject the null hypothesis.

Example 9.5: Reese's Pieces

Use the Rejection Region Approach to complete the S step for the Reese's Pieces problem.



$$P(Z < z_{1-.05}) = 0.05$$

$$z_{1-.05} = -1.645.$$

This value can be found using a Standard Normal Table, or using the invNorm function on your calculator: invNorm(0.05, 0, 1).

Thus, our critical value is -1.645 and our rejection region is:

$$RR: Z < -1.645$$

Since the test statistic $z_0 = 0$, $-2.12 < -1.645$ (in the rejection region), we reject H_0 . At the 5% significance level, there is sufficient evidence to conclude that less than 50% of the candies produced by the machine are orange.

9.3.2 APPROACH 2: P-VALUE

Before we use the P-value approach in an example, let's define what the P-value is and how to use it.

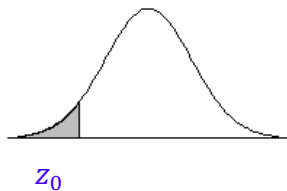
When performing a hypothesis test using the **P-value approach**, we compute a value from the test statistic called the **P-value**. The P-value is the probability of getting your test statistic or something more extreme (i.e. further towards the tail or tails of the appropriate probability distribution) **if** the null hypothesis is true.

If the P-value is small, we have reason to suspect that the null hypothesis might not be true for the population of interest. Small P-values correspond to test statistics that are in the rejection region. When this is the case, we say that we have a statistically significant result with evidence in favor of the alternative hypothesis. We consider the P-value to be small if it is less than the significance level α .

Example 9.6: Reese's Pieces

Use the P-value Approach to complete the S step for the Reese's Pieces problem.

$$z_0 \approx -2.12$$



$$P\text{-value} = P(Z < z_0) = P(Z < -2.12) = 0.0169$$

Since the P-value of $0.0169 \leq \alpha = 0.05$, we reject H_0 .

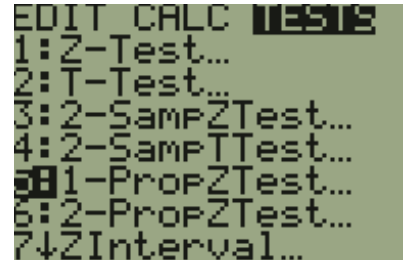
At the 5% significance level, there is sufficient evidence to conclude that less than 50% of the candies produced by the machine are orange.

Calculator Tutorial 9.1: Finding P-Values for Proportion Using 1-PropZTest

You can compute **P-values** for hypothesis tests for proportion using normalcdf and the properties of the standard normal distribution. (See Section 7).

You can also compute **P-values** using a z-test function in your calculator.

1. Click the STAT button, use the arrow keys to navigate to the TESTS menu, then navigate to item 5: 1-PropZTest... Click ENTER.



```

EDIT CALC TESTS
1:Z-Test...
2:T-Test...
3:2-SampZTest...
4:2-SampTTest...
5:1-PropZTest...
6:2-PropZTest...
7:ZInterval...
  
```

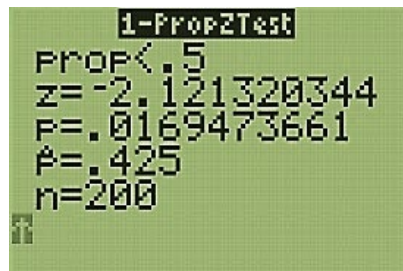
2. Enter the hypothesized population proportion p_0 , the number of individuals x with the desired characteristic in the sample, and the sample size n . Select the appropriate test direction, then select Calculate and click ENTER.



```

1-PropZTest
P0: .5
x: 85
n: 200
PROP≠P0 <P0 >P0
Calculate
  
```

3. The results of the hypothesis test will be displayed on the screen, including values for the test statistic z and the P -value which is simply labeled p . The value of $\hat{p}$ and the sample size n are given.



```

1-PropZTest
PROP<.5
Z=-2.121320344
P=.0169473661
p=.425
n=200
  
```

This example shows the results of the hypothesis test conducted in Example 9.6.

Example 9.7: Health Care (Two-Tailed Test) P-value Approach

A political action group claims that 44% of adult Americans believe that it is the responsibility of the federal government to provide healthcare coverage for all citizens. A recent Gallup survey of randomly selected American adults found that 493 out of 1050 of those polled agreed that the federal government should provide healthcare coverage for all citizens.

Does the Gallup survey provide statistical evidence to suggest that the true proportion of adult Americans who believe healthcare coverage is a responsibility of the federal government is different than 0.44 at the 5% level of significance? **Use the P-value approach.**

- A. Define the parameter of interest and give the appropriate **hypotheses** for this test.

Let p = the true proportion of adult Americans who believe that it is the responsibility of the federal government to provide health care coverage for all citizens.

$$H_0: p = 0.44$$

$$H_1: p \neq 0.44$$

- B. State and check the appropriate **assumptions** for a hypothesis test to be valid.

(1) The sample is randomly selected, as stated in the problem.

(2) The sampling distribution of $\hat{p}$ is approximately normally distributed since

$$np_0 = 1050(0.44) = 462 \geq 5 \text{ and } n(1 - p_0) = 588 \geq 5$$

- C. Calculate the **test** statistic.

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} = \frac{0.4695 - 0.44}{\sqrt{\frac{0.44(0.56)}{1050}}} = 1.93$$

- D. Calculate the P-value and **summarize** your conclusions.

$$P\text{-value} = 2P(Z > |1.93|) = 2 \cdot P(Z > 1.93) = 2(0.0268) = 0.0536$$

Since $P\text{-value} = 0.0536 > \alpha = 0.05$, we fail to reject the null hypothesis.

At the 5% significance level, there is insufficient evidence to suggest that the true proportion of adult Americans who believe that it is the responsibility of the federal government to provide healthcare coverage is different than 0.44.

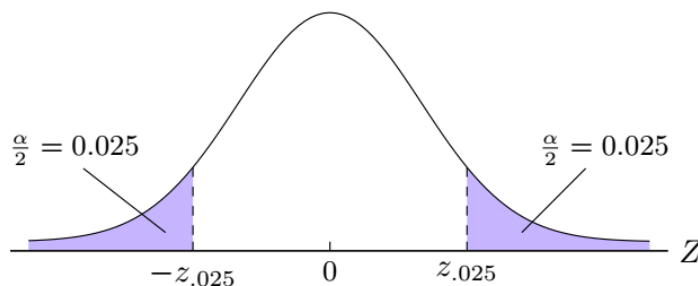
Example 9.8: Health Care (Two-Tailed Test) Rejection Region Approach

A political action group claims that 44% of adult Americans believe that it is the responsibility of the federal government to provide healthcare coverage for all citizens. A recent Gallup survey of randomly selected American adults found that 493 out of 1050 of those polled agreed that the federal government should provide healthcare coverage for all citizens.

Does the Gallup survey provide statistical evidence to suggest that the true proportion of adult Americans who believe healthcare coverage is a responsibility of the federal government is different than 0.44 at the 5% level of significance?

Use the Rejection Region approach. Since the **H**, **A** and **T** steps are the same as in the previous example, simply find the Rejection Region and write the conclusion.

A. Find the Rejection Region.



$$RR: Z > 1.96 \text{ or } Z < -1.96$$

B. **Summarize** your conclusions.

Since the test statistic $z_0 = 1.93$ is not in the rejection region $(-1.96 < z_0 < 1.96)$, we fail to reject the null hypothesis.

At the 5% significance level, there is insufficient evidence to suggest that the true proportion of adult Americans who believe that it is the responsibility of the federal government to provide healthcare coverage is different than 0.44.

MY NOTES:

9.4 HYPOTHESIS TESTS FOR MEAN

Like the confidence interval for population mean, a **hypothesis test for μ** is based on the **t -distribution**. (Some textbooks state that the standard normal distribution may be used for tests with large sample sizes. We will solely use the t -distribution.) We will follow the same procedure **HATS**. Recall there are **2** approaches for hypothesis testing.

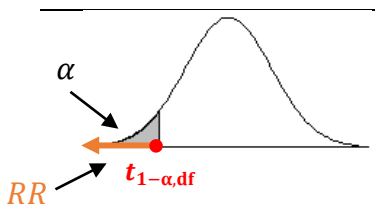
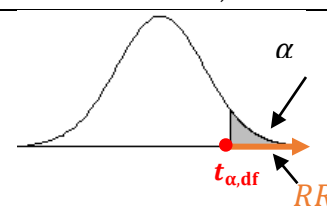
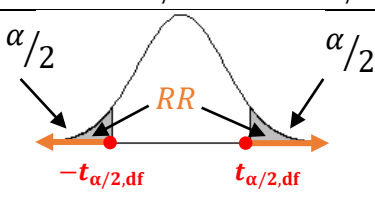
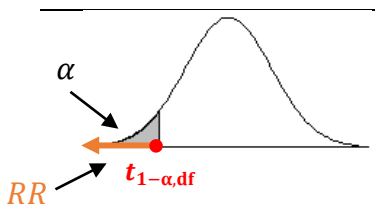
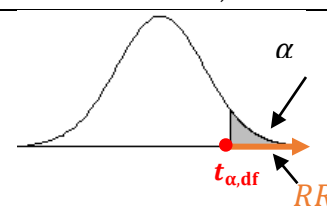
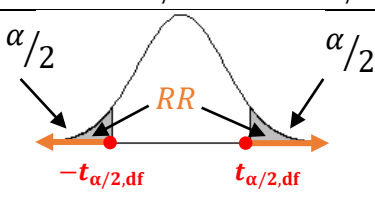
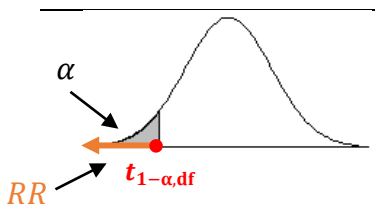
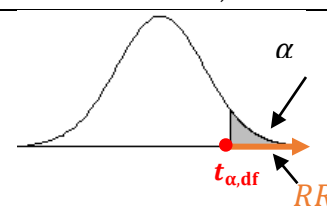
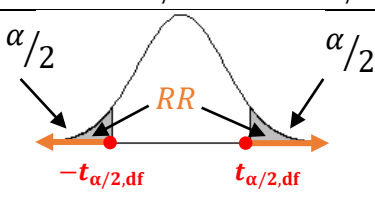
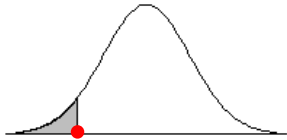
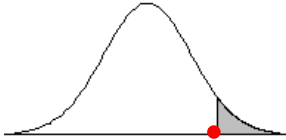
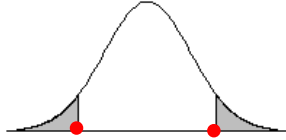
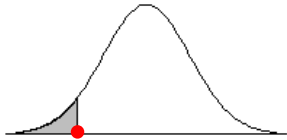
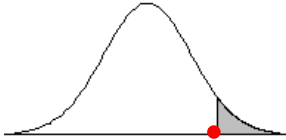
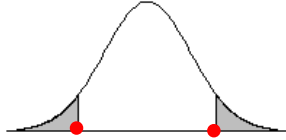
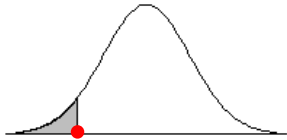
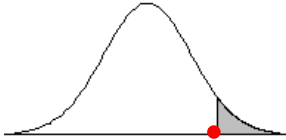
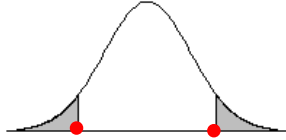
TABLE 9.3: HYPOTHESIS TESTING FOR MEAN – REJECTION REGION APPROACH											
<u>HYPOTHESES</u> ➤ Define μ in context of the problem. ➤ The hypotheses can be structured in one of three ways: <table> <tr> <th>LEFT-TAILED TEST</th><th>RIGHT-TAILED TEST</th><th>TWO-TAILED TEST</th></tr> <tr> <td>$H_0: \mu = \mu_0$</td><td>$H_0: \mu = \mu_0$</td><td>$H_0: \mu = \mu_0$</td></tr> <tr> <td>$H_1: \mu < \mu_0$</td><td>$H_1: \mu > \mu_0$</td><td>$H_1: \mu \neq \mu_0$</td></tr> </table> where μ_0 is the hypothesized value of the population mean.			LEFT-TAILED TEST	RIGHT-TAILED TEST	TWO-TAILED TEST	$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$	$H_1: \mu < \mu_0$	$H_1: \mu > \mu_0$	$H_1: \mu \neq \mu_0$
LEFT-TAILED TEST	RIGHT-TAILED TEST	TWO-TAILED TEST									
$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$									
$H_1: \mu < \mu_0$	$H_1: \mu > \mu_0$	$H_1: \mu \neq \mu_0$									
<u>ASSUMPTIONS / CONDITIONS</u> ➤ The sample data come from a random sample or randomized experiment. ➤ The sampling distribution for $\bar{X}$ is approximately normal. This will be true if X is normal or $n \geq 30$.											
<u>TEST STATISTIC</u> $t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ which follows the t-distribution with $n - 1$ degrees of freedom.											
FIND REJECTION REGION BASED ON THE SIGNIFICANCE LEVEL α OF THE TEST* <table> <tr> <th>LEFT-TAILED TEST RR: $T < t_{1-\alpha, df}$</th><th>RIGHT-TAILED TEST RR: $T > t_{\alpha, df}$</th><th>TWO-TAILED TEST RR: $T < -t_{\alpha/2, df}$ OR $T > t_{\alpha/2, df}$</th></tr> <tr> <td></td><td></td><td></td></tr> </table> * Some textbooks refer to the rejection region (RR) as the decision rule.			LEFT-TAILED TEST RR: $T < t_{1-\alpha, df}$	RIGHT-TAILED TEST RR: $T > t_{\alpha, df}$	TWO-TAILED TEST RR: $T < -t_{\alpha/2, df}$ OR $T > t_{\alpha/2, df}$						
LEFT-TAILED TEST RR: $T < t_{1-\alpha, df}$	RIGHT-TAILED TEST RR: $T > t_{\alpha, df}$	TWO-TAILED TEST RR: $T < -t_{\alpha/2, df}$ OR $T > t_{\alpha/2, df}$									
											
<u>SUMMARIZE</u> ➤ Reject H_0 if the test statistic lies within the rejection region. In context of the study, state that there is sufficient evidence to support H_1. ➤ Fail to reject H_0 if the test statistic does not lie in the rejection region. In the context of the study, state that there is not sufficient evidence to support H_1.											

TABLE 9.4: HYPOTHESIS TESTING FOR PROPORTION – P-VALUE APPROACH								
<u>HYPOTHESES</u> ➤ Define μ in context of the problem. ➤ The hypotheses can be structured in one of three ways: <table> <tr> <th>LEFT-TAILED TEST</th><th>RIGHT-TAILED TEST</th><th>TWO-TAILED TEST</th></tr> <tr> <td> $H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$ </td><td> $H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$ </td><td> $H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$ </td></tr> </table> where μ_0 is the hypothesized value of the population mean.			LEFT-TAILED TEST	RIGHT-TAILED TEST	TWO-TAILED TEST	$H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$	$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$	$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$
LEFT-TAILED TEST	RIGHT-TAILED TEST	TWO-TAILED TEST						
$H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$	$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$	$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$						
<u>ASSUMPTIONS / CONDITIONS</u> ➤ The sample data come from a random sample or randomized experiment. ➤ The sampling distribution for $\bar{X}$ is approximately normal. This will be true if X is normal <i>or</i> $n \geq 30$.								
<u>TEST STATISTIC</u> $t_0 = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$ which follows the t-distribution with $n - 1$ degrees of freedom.								
CALCULATE THE P-VALUE <table> <tr> <th>LEFT-TAILED TEST</th><th>RIGHT-TAILED TEST</th><th>TWO-TAILED TEST</th></tr> <tr> <td>  t_0 p-value = $P(T < t_0)$ </td><td>  t_0 p-value = $P(T > t_0)$ </td><td>  $- t_0$ t_0 p-value = $2 \cdot P(T > t_0)$ </td></tr> </table>			LEFT-TAILED TEST	RIGHT-TAILED TEST	TWO-TAILED TEST	 t_0 p-value = $P(T < t_0)$	 t_0 p-value = $P(T > t_0)$	 $- t_0$ t_0 p-value = $2 \cdot P(T > t_0)$
LEFT-TAILED TEST	RIGHT-TAILED TEST	TWO-TAILED TEST						
 t_0 p-value = $P(T < t_0)$	 t_0 p-value = $P(T > t_0)$	 $- t_0$ t_0 p-value = $2 \cdot P(T > t_0)$						
<u>SUMMARIZE</u> ➤ Reject H_0 if P-value $\leq \alpha$. In context of the study, state that there is sufficient evidence to support H_1. ➤ Fail to reject H_0 if P-value $> \alpha$. In the context of the study, state that there is not sufficient evidence to support H_1.								

The following example will illustrate these steps of the hypothesis test for a mean using both approaches. We will begin with the **H**, **A**, and **T** steps since those steps are the same for both the Rejection Region approach and the P-value approach.

Example 9.9: Potato Chips

The Better Business Bureau has received several complaints that a potato chip company is underfilling its five-ounce bags of chips. The Bureau randomly selects 25 bags of chips and determines the weight of each bag. The sample average weight of the bags is 4.90 ounces with a standard deviation of 0.15 ounces. It is known that the weight of potato chip bags is normally distributed. Is there evidence at the 1% level that the bags are being underfilled?

- A. What is the significance level of this hypothesis test?

The significance level for the test is $\alpha = 0.01$.

- B. Define the parameter of interest and state the **hypotheses**.

μ = true mean weight of the bags of potato chips at this company

$H_0: \mu = 5$ ounces

$H_1: \mu < 5$ ounces

- C. State and check whether the necessary **assumptions** hold.

(1) Random sample – stated that bags are randomly selected.

(2) $\bar{X}$ must be approximately normal – true since X is normally distributed.

- D. Compute the appropriate **test** statistic.

Information given: $\bar{x} = 4.9$, $s = 0.15$, $n = 25$, $\mu_0 = 5$

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{4.9 - 5}{0.15/\sqrt{25}} = -3.33$$

9.4.1 APPROACH 1: REJECTION REGION

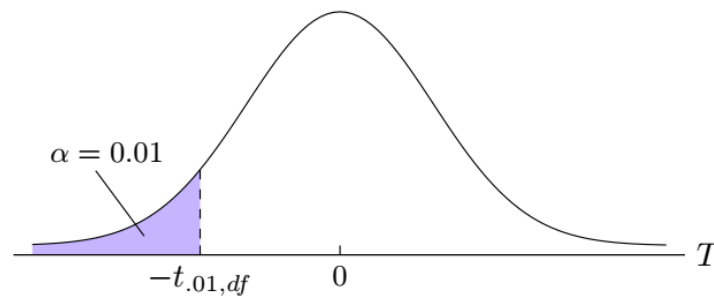
In the Rejection Region Approach (decision rule) we again find the test statistic that separates a decision to reject the null hypothesis from a decision to fail to reject the null hypothesis. The test statistics that fall into the rejection region represent those that put the sample statistic too far away from the hypothesized value to believe that the null hypothesis is true.

In a test for a mean, we will use the [T-distribution](#) to find the rejection region. We need to find a location or locations along the horizontal axis that correspond(s) to putting α into the appropriate tail(s). This/these location(s) will be our critical value(s).

Example 9.10: Potato Chips RR Approach

Use the rejection region approach to complete the S step for the Potato Chips problem

Since this is a left-tailed test, $\alpha = 0.01$ corresponds to the area in the left tail of the distribution. We may find the t -value that separates the lower 0.01 area from the rest of the distribution using the table or the calculator invT function.



$$P(T < t_{.01,24}) = 0.01 \Rightarrow t_{.01,24} = \text{invT}(.01,24) = -2.492$$

Rejection Region (RR): $T < -2.492$

Since $t_0 = -3.33 < -2.492$, we reject H_0 .

At the 1% significance level, there is sufficient evidence to suggest that the true mean weight of the bags of chips is less than 5 ounces. We have reason to suspect that the company is underfilling the bags.

9.4.2 APPROACH 2: P-VALUE

Recall: when performing a hypothesis test using the **P-value approach**, we compute a value from the test statistic called the **P-value**. The P-value is the probability of getting your test statistic or something more extreme (i.e. further towards the tail or tails of the appropriate probability distribution) **if** the null hypothesis is true.

If the P-value is small, we have reason to suspect that the null hypothesis might not be true for the population of interest. Small P-values correspond to test statistics that are in the rejection region. When this is the case, we say that we have a statistically significant result with evidence in favor of the alternative hypothesis. We consider the P-value to be small if it is less than the significance level α .

You can approximate the p-value using the t -distribution table; however, the table is not detailed enough to provide an exact tail probability and can only provide a range of values within which the p-value falls. Statistical software such as JMP or a TI-83 or TI-84 calculator can find the exact p-value for a hypothesis test for μ .

Example 9.11: Potato Chips P-value Approach

Use the P-value approach to complete the **S** step for the Potato Chips problem

Using the calculator: $p\text{-value} = P(T < t_0) = P(T < -3.33) = 0.0014$

Using a table: Since the t -distribution is symmetric, $P(T < -3.33) = P(T > 3.33)$, which we need to use since the t -table only has positive t -values. For $df = n - 1 = 24$, we can see from the table that, since $3.33 > 2.797$, the p -value will be less than 0.005. This is for reference only; in STAT 3090, focus should be placed on computing exact p -values with the TI-84.

At the 1% significance level, there is sufficient evidence to suggest that the true mean weight of the bags of chips at this company is less than 5 ounces.

We have reason to suspect that the company is underfilling the bags.

Calculator Tutorial 9.2: Finding P-Values for Mean Using tcdf

One way to compute **p-values** using the t -distribution and is quite similar to the way we find normal probabilities using normalcdf.

1. Click 2ND + VARS, then use the down arrow key to navigate to item 6: tcdf(. Click ENTER.
2. Enter the lower and upper limits and the degrees of freedom. Then use navigate to Paste and click ENTER.
3. The tcdf(function will be displayed on screen. Click ENTER once more to compute the probability.

This example shows how to find $P(T < -3.33)$.

```

DISTR DRAW
1:normalpdf(
2:normalcdf(
3:invNorm(
4:invT(
5:tpdf(
6:tcdf(
7:χ²pdf(
  
```

```

tcdf(
lower:-1E99
upper:-3.33
df:24
Paste
  
```

```

tcdf(-1E99, -3.33, 24)
.001399522
  
```

Calculator Tutorial 9.3: Finding P-Values for Mean Using T-Test

You can also compute **p-values** using a t -test function in your calculator.

1. Click the STAT button, use the arrow keys to navigate to the TESTS menu, then navigate to item 2: T-Test... Click ENTER.
2. For Inpt, select the Stats option. Enter the hypothesized mean μ_0 , the sample mean $\bar{x}$, the sample standard deviation S_x (s), and the sample size n . Then select the appropriate-tailed hypothesis test. Select Calculate and click ENTER.
3. The results of the hypothesis test will be displayed on the screen, including values for the test statistic t and p -value.

This example shows the results of the hypothesis test conducted in Example 9.11

```

EDIT CALC TESTS
1:Z-Test...
2:T-Test...
3:2-SampZTest...
4:2-SampTTest...
5:1-PropZTest...
6:2-PropZTest...
7:ZInterval...
  
```

```

T-Test
Inpt:Data Stats
μ₀:5
x̄:4.9
Sx:15
n:25
μ:≠μ₀ <μ₀ >μ₀
Calculate Draw
  
```

```

T-Test
μ<5
t=-3.333333333
P=.0013881571
x̄=4.9
Sx=15
n=25
  
```

MY NOTES:

9.5 HYPOTHESIS TESTS FOR VARIANCE

Most statistical inferences are conducted for population mean μ or population proportion p . However, it is sometimes of interest to perform **hypothesis tests for population variance σ^2** (or equivalently, standard deviation σ).

For example, one application of hypothesis tests for σ^2 is in looking at *quality control* to make sure that the production of something does not contain a high amount of variability. This is often of importance in food production or the production of prescription drugs, where it is important that the products be consistent and have a low variance.

The **test statistic** used in hypothesis tests for population variance has a distribution known as the chi-square (χ^2) distribution.

The chi-square distribution is right skewed. The support for the distribution is $X \geq 0$.

Like the t -distribution, this distribution changes shape depending on the number of degrees of freedom, $df = n - 1$.

The upper tail probability of this distribution for various critical values is given in your text and in the Appendix of the Lecture Guide. This χ^2 distribution table is read similarly to the t -distribution table.

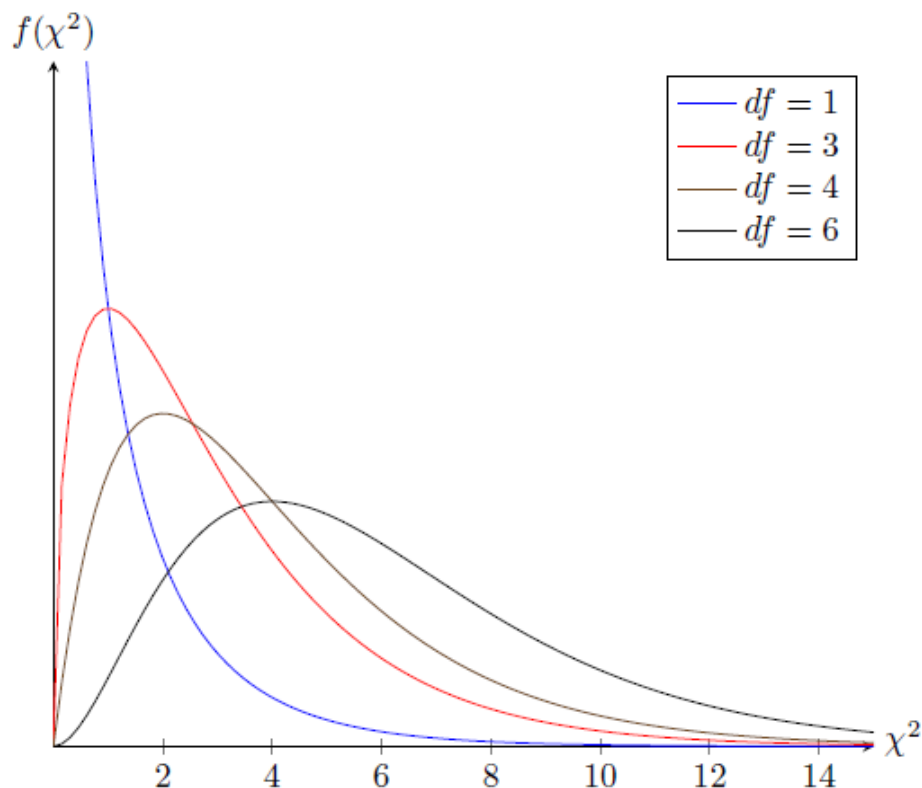


TABLE 9.5: HYPOTHESIS TEST FOR VARIANCE – RR APPROACH**HYPOTHESES**

- Define σ^2 in context of the problem.
- The hypotheses can be structured in one of three ways:

LEFT-TAILED TEST	RIGHT-TAILED TEST	TWO-TAILED TEST
$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 < \sigma_0^2$	$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 > \sigma_0^2$	$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 \neq \sigma_0^2$

where σ_0^2 is the hypothesized value of the population variance.

ASSUMPTIONS / CONDITIONS

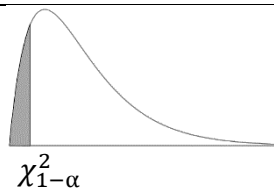
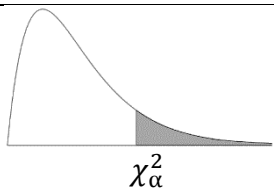
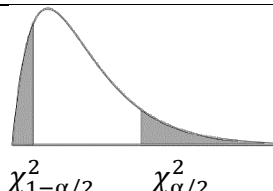
- A random sample must be selected from the target population.
- The population from which the sample is selected must have an approximately normal distribution.

TEST STATISTIC

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

which follows the χ^2 distribution.

FIND REJECTION REGION BASED ON THE SIGNIFICANCE LEVEL α OF THE TEST*

LEFT-TAILED TEST RR: $\chi^2 < \chi_{1-\alpha}^2$	RIGHT-TAILED TEST RR: $\chi^2 > \chi_{\alpha}^2$	TWO-TAILED TEST RR: $\chi^2 < \chi_{1-\alpha/2}^2$ or $\chi^2 > \chi_{\alpha/2}^2$
		

* Some textbooks refer to the rejection region (RR) as the **decision rule**.

SUMMARIZE

- Reject H_0 if the test statistic lies within the rejection region.
In context of the study, state that there is sufficient evidence to support H_1 .
- Fail to reject H_0 if the test statistic does not lie in the rejection region.
In the context of the study, state that there is not sufficient evidence to support H_1 .

TABLE 9.6: HYPOTHESIS TEST FOR VARIANCE – P-VALUE APPROACH**HYPOTHESES**

- Define σ^2 in context of the problem.
- The hypotheses can be structured in one of two ways:

LEFT-TAILED TEST	RIGHT-TAILED TEST	TWO-TAILED TEST
$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 < \sigma_0^2$	$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 > \sigma_0^2$	 N/A Use Rejection Region Approach

where σ_0^2 is the hypothesized value of the population variance.

ASSUMPTIONS / CONDITIONS

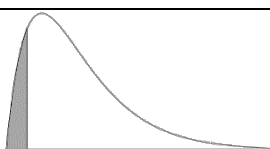
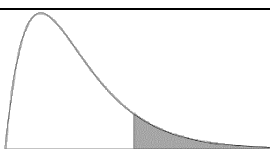
- A random sample must be selected from the target population.
- The population from which the sample is selected must have an approximately normal distribution.

TEST STATISTIC

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

which follows the χ^2 distribution.

CALCULATE THE P-VALUE

LEFT-TAILED TEST	RIGHT-TAILED TEST	TWO-TAILED TEST
 χ_0^2 $P\text{-value} = P(\chi^2 < \chi_0^2)$	 χ_0^2 $P\text{-value} = P(\chi^2 > \chi_0^2)$	 Use Rejection Region Approach

5. CONCLUSION

- Reject H_0 if $p\text{-value} \leq \alpha$.
In context of the study, state that there is sufficient evidence to support H_1 .
- Fail to reject H_0 if $p\text{-value} > \alpha$.
In the context of the study, state that there is not sufficient evidence to support H_1 .

Note: Due to the fact that the chi-square distribution is asymmetric and can only take on positive values, the P-value approach is not appropriate when conducting two-tailed hypothesis tests for population variance. **You must use the rejection region approach to conduct two-tailed tests for σ^2 .**

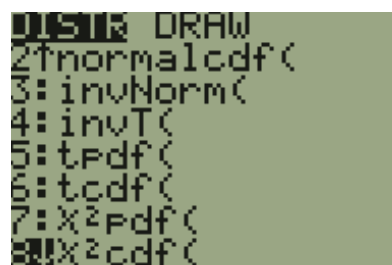
We typically will frame our hypothesis test in terms of variance. However, if a problem asks about the **standard deviation**, it is acceptable to write your hypotheses and summarizing statement in terms of standard deviation. You may do so as long as your values are consistent with the parameter you are describing; for example, if given a hypothesized variance of $\sigma_0^2 = 256$, you must recognize that the corresponding standard deviation is $\sigma_0 = \sqrt{256} = 16$.

Calculator Tutorial 9.4: Finding P-Values for Variance Using χ^2 cdf

To perform a hypothesis test for σ^2 using P-value approach, you will need to compute **P-values**. You can do so using a function called χ^2 cdf, which functions similarly to the tcdf function (see Calculator Tutorial 9.2).

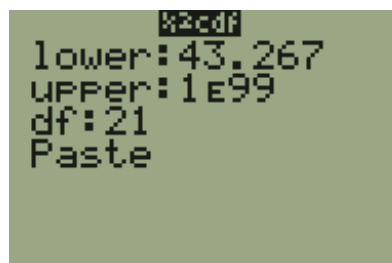
1. Click 2ND + VARS to open the distribution menu, then use the arrow keys to navigate to item 8: χ^2 cdf(.
2. Enter the lower bound, upper bound, and degrees of freedom for the appropriate probability statement used to calculate the p-value. Click ENTER.
3. The function to compute the p-value will appear on the main screen. Click ENTER once more to calculate the p-value.

This example computes $P(\chi^2 > 43.267)$.



```

DISTR DRAW
2:normalcdf(
3:invNorm(
4:invT(
5:tpdf(
6:tcdf(
7:X^2pdf(
8:X^2cdf(
  
```



```

X^2cdf
lower: 43.267
upper: 1E99
df: 21
Paste
  
```



```

X^2cdf(43.267, 1E99
.0029058133
  
```

Example 9.12: Monthly Income RR Approach

In one town, monthly incomes for men with college degrees are found to have a standard deviation of \$650. A random sample of 22 men without college degrees resulted in incomes with a standard deviation of \$933. Assume the monthly income for men is normally distributed. At the 1% significance level, do the data provide sufficient evidence to conclude that the standard deviation, σ , of incomes of men without college degrees in that town is greater than \$650? Use the rejection region approach.

A. Define the appropriate **hypotheses**.

σ^2 = population variance of incomes of men in this town without college degrees

$$H_0: \sigma^2 = 422,500 \quad (\text{since } \sigma_0^2 = 650^2 = 422,500)$$

$$H_1: \sigma^2 > 422,500$$

B. State and verify that the proper **assumptions** hold.

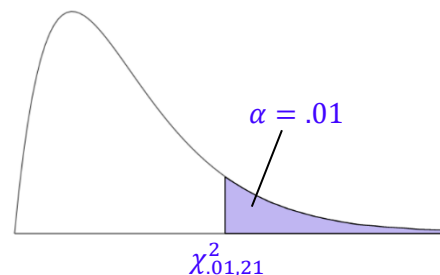
(1) Stated in the problem that the sample of 22 men without degrees was randomly selected.

(2) Stated in the problem that population monthly income for men is normally distributed.

C. Calculate the **test** statistic and find the rejection region

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(22-1)(933^2)}{422,500} = 43.267$$

$$\chi_{.01,21}^2 = 38.932 \Rightarrow \text{RR: } \chi^2 > 38.932$$



D. Write a **summary** of your conclusion in context of the problem.

Since $\chi_0^2 = 43.267 > 38.932$ (i.e. 43.267 is in the rejection region), we reject H_0 .

At the 1% significant level, we have sufficient evidence that...

- the true variance (standard deviation) of the monthly incomes of men without college degrees in this town is greater than 422,500 (\$650). or
- the true variance (standard deviation) of the monthly incomes of men without college degrees in this town is greater than that of men with college degrees. or
- the monthly incomes of men without college degrees are more variable than those of men with college degrees.

Example 9.13: Monthly Income P-value Approach

Using the test statistic found in Part C in example 9.12, complete the hypothesis test for the scenario on the previous page using the p-value approach.

A. Compute the **p-value**.

$$p\text{-value} = P(\chi^2 > 43.267) = 0.0029$$

B. Write a **summary** of your conclusion in the context of the problem.

Since $0.0029 < \alpha = 0.01$, we reject the null hypothesis H_0 .

At the 1% significant level, we have sufficient evidence that...

- *the true variance (standard deviation) of the monthly incomes of men without college degrees in this town is greater than 422,500 (\$650). or*
- *the true variance (standard deviation) of the monthly incomes of men without college degrees in this town is greater than that of men with college degrees. or*
- *the monthly incomes of men without college degrees are more variable than those of men with college degrees.*

Example 9.14: Wait Times

A store manager wishes to investigate whether he should switch from using multiple checkout lines to a single checkout line in his store. He knows that the standard deviation of the waiting times using multiple lines is 5.2 minutes. After switching to a single line, he finds that for a random sample of 29 customers, the waiting times have a standard deviation of 4.3 minutes. Assume that the line waiting times are normally distributed. Use a 0.05 significance level to test whether the standard deviation of the waiting times for a single line differs from 5.2 minutes. Use the rejection region approach.

Hypotheses

σ^2 = population variance of waiting times for a single line

$$H_0: \sigma^2 = 27.04 \quad (\text{since } \sigma^2 = 5.2^2 = 27.04)$$

$$H_1: \sigma^2 \neq 27.04$$

Assumptions

- (1) Random sample – stated in the problem.
- (2) Population is normally distributed – stated in the problem.

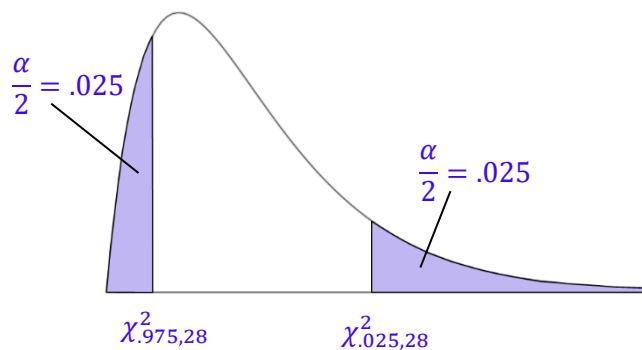
Test

$$\chi_0^2 = \frac{(29 - 1)(4.3^2)}{27.04} = 19.14$$

$$\alpha = 0.05 \Rightarrow \frac{\alpha}{2} = 0.025$$

$$\chi_{.975,28}^2 = 15.308, \chi_{.025,28}^2 = 44.461$$

$$\text{RR: } \chi^2 < 15.308 \text{ OR } \chi^2 > 44.461$$

Summary

Since $15.308 < 19.14 < 44.461$ (i.e. the test statistic is not in the RR), we fail to reject H_0 .

At the 5% significance level, there is insufficient evidence to suggest that the true variability in (standard deviation of) waiting times for a single line is different from 27.04 minutes² (5.2 minutes); that is, different than that of waiting times for multiple lines.

MY NOTES:

PRACTICE EXERCISES FOR SECTION 9

1. In 1990, the mean height of women 20 years of age or older was 63.6 inches based on data obtained by the CDC. Suppose that a random sample of 45 women 20 years and older today results in a mean height of 63.75 inches with a standard deviation of 0.7. Does your sample provide sufficient evidence at the $\alpha = 0.005$ level that women are taller today than in 1990?

(a) Define the parameter of interest and state the hypotheses.

μ = true mean height of women 20 years and older today

$H_0: \mu = 63.6$ in.

$H_1: \mu > 63.6$ in.

(b) List and verify the assumptions needed to perform a hypothesis.

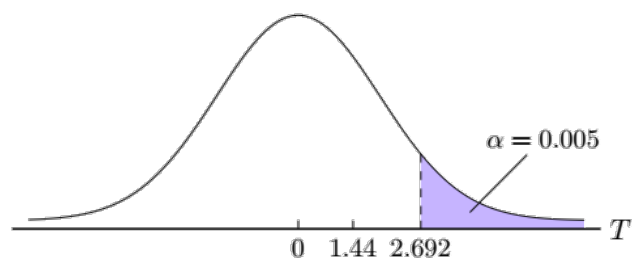
(1) It is stated that a random sample of women 20 years and older was selected.

(2) The distribution of $\bar{X}$ is approximately normal since $n = 45 > 30$.

(c) Compute the appropriate test statistic and find the rejection region. Draw a sketch that shows the shaded rejection region and where your test statistic falls relative to the critical values.

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{63.75 - 63.6}{0.7/\sqrt{45}} = 1.44$$

Critical Value: $t_{.005,44} = 2.692$



(If using the t -table, $t_{.005,44} \approx t_{.005,40} = 2.704$.)

$RR: T > 2.692$

(d) Summarize the results of your test in the context of the problem

Since $t_0 = 1.44$ is not in the rejection region, we fail to reject H_0 .

At the $\alpha = .005$ level, we have insufficient evidence that the true mean height of women is greater than 63.6 inches (that is, that women are taller today than in 1990).

2. A researcher wishes to determine whether full-time college students get an amount of sleep different than the recommended eight hours per night. He takes a random sample of 52 full-time college students and finds the average number of hours slept per night to be 7.2 hours with a standard deviation of 1.27 hours. Does the researcher have evidence at the 5% level that college students do not sleep eight hours per night?

(a) Define the parameter of interest and state the hypotheses.

μ = true mean number of hours slept per night by full-time college students

$H_0: \mu = 8$ hours

$H_1: \mu \neq 8$ hours

(b) State and check whether the necessary assumptions hold.

(1) It is stated that a random sample of 52 full-time college students was selected.

(2) $\bar{X}$ is approximately normal since the sample size is large enough ($n = 52 \geq 30$).

(c) Compute the test statistic and a corresponding p-value.

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{7.2 - 8}{1.27/\sqrt{52}} = -4.542$$

$$p\text{-value} = 2 \cdot P(T < -4.542) = 0.000034$$

(d) Summarize the results of your test.

Since $p\text{-value} = 0.0015 < \alpha = 0.05$, we reject H_0 .

At the 5% significance level, we have sufficient evidence to suggest that the true mean number of hours slept per night by full-time college students is different than eight hours.

3. A survey conducted five years ago by the health center at a university showed that 25% of the students smoked at the time. Since then, the university has hosted several student-led smoking prevention campaigns to educate students about the dangers of smoking and encourage them to pursue healthier habits. This year, a new survey was conducted on a random sample of 200 students from the university, and it was found that 36 of them smoke. Do these data provide convincing evidence at the 4% level to suggest that the percentage of students who smoke has decreased over the last five years?

(a) Define the parameter of interest and state the hypotheses.

p = true proportion of students at this university who smoke

$$H_0: p = 0.25$$

$$H_1: p < 0.25$$

(b) Check whether the appropriate assumptions hold.

(1) It is stated that a random sample of 200 students was selected at the university.

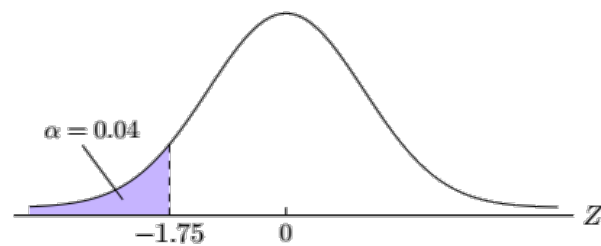
(2) The distribution of $\hat{p}$ is approximately normal since $np_0 = 200(.25) = 50 \geq 5$ and $n(1 - p_0) = 200(0.75) = 150 \geq 5$.

(c) Calculate the test statistic and find the rejection region. Include a sketch.

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} = \frac{0.18 - 0.25}{\sqrt{\frac{0.25(0.75)}{200}}} = -2.286$$

$$\text{Critical Value: } z_{.04} = -1.75$$

$$RR: Z < -1.75$$



(d) Summarize the results of the test in context.

Since the test statistic z_0 is in the rejection region ($-2.286 < -1.75$), reject H_0 .

At the 4% significance level, we have sufficient evidence that the true proportion of students who smoke at this university is less than 0.25 (in other words, the true proportion has decreased from five years ago).

4. The manufacturers of Ramen noodles, a favorite college student staple, are concerned that the sodium levels in their product vary too much. A package of Ramen is supposed to contain 1100 mg of sodium with a standard deviation of 110 mg. Assume that Ramen sodium levels are normally distributed. A random sample of 20 packages was analyzed and the sample standard deviation was found to be 129.57 mg. Is there evidence at the $\alpha = 0.10$ level that the standard deviation of the sodium content exceeds the desired level?

(a) Define the parameter of interest and state the hypotheses.

σ^2 = true variance of sodium levels in Ramen noodles

$$H_0: \sigma^2 = 12,100 \quad \text{since } (110 \text{ mg})^2 = 12,100 \text{ mg}^2$$

$$H_1: \sigma^2 > 12,100$$

(b) Verify that the appropriate assumptions for inference are met.

(1) It is stated that a random sample of 20 Ramen packages is selected.

(2) It is stated that the population (Ramen sodium levels) are normally distributed.

(c) Calculate the test statistic and find the corresponding p-value.

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{19(129.57)^2}{12,100} = 26.362$$

$$p\text{-value} = P(\chi^2 > 26.362) = 0.1204$$

(d) Summarize the results of the hypothesis test in context.

Since the p-value 0.1204 is greater than the significance level $\alpha = 0.10$, we fail to reject H_0 .

At the 10% significance level, we have insufficient evidence to conclude that the true variance (standard deviation) of the sodium levels of Ramen noodles is higher than the desired level of 12,100 mg² (110 mg).

5. An accounting professor wishes to see how the final exam scores for an introductory accounting course this semester vary compared to the previous semester. Last semester, the standard deviation in final exam scores for all students enrolled in the course was 13.5. This semester, the professor randomly selects 41 students and finds the standard deviation of their final exam scores to be 18.09. He knows that the final exam scores are normally distributed. Does he have evidence at the 5% level that the standard deviation of accounting final exam scores has changed this semester?

(a) Define the parameter of interest and state the hypotheses.

σ^2 = true variance of final exam scores for all accounting students this semester

$$H_0: \sigma^2 = 182.25$$

$$H_1: \sigma^2 \neq 182.25$$

(b) Check whether the appropriate assumptions hold.

(1) It is stated that the professor randomly selected 38 students.

(2) The distribution of final exam scores (the population) is normal.

(c) Calculate the test statistic and find the corresponding rejection region.

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(41-1)(18.09)^2}{182.25} = 71.824$$

$$\text{Critical Values: } \chi_{.025,40}^2 = 59.342 \text{ and } \chi_{.975,40}^2 = 24.433$$

$$RR: \chi^2 < 24.433 \text{ or } \chi^2 > 59.342$$

(d) Summarize the results of the hypothesis test in context.

Since the test statistic χ_0^2 is in the rejection region ($71.824 > 59.342$), we reject H_0 .

At the $\alpha = 0.05$ significance level, we have sufficient evidence that the true variance (standard deviation) of accounting final exam scores this semester is different than 182.25 (13.5); that is, it is different than last semester.

6. In an angiogram, heart problems are examined via a small tube (catheter) threaded into the heart from a vein in a patient's leg. It is important that the company manufacturing the catheter maintain a mean diameter of 2.00 mm for the instrument to be safe for patients (the standard deviation is small). Each day, quality control personnel take measurements to test

$$H_0: \mu = 2.00 \text{ mm} \quad \text{against} \quad H_1: \mu \neq 2.00 \text{ mm}$$

at the 1% level. If they discover a problem (in other words, they reject the null hypothesis), they will stop the manufacturing process until it is corrected.

- (a) Define a Type I error in context.

The evidence suggests that the true average diameter is not 2.00 mm when it is actually 2.00 mm.

- (b) Describe a possible consequence of performing a Type I error in this situation.

The company will take action to correct the process when there is no need to do so.

- (c) What is the probability of making a Type I error in this case?

The probability of making a Type I error is $\alpha = 0.01$.

- (d) Define a Type II error in context.

The evidence will suggest that the true average diameter is 2.00 mm when it is actually not 2.00 mm.

- (e) Describe a possible consequence of performing a Type II error in this situation.

The company will not take action to correct the process when they should. This will result in a product that does not meet the specifications and is not safe to use.

7. In hypothesis testing, what is a way that you can decrease the probability of making a Type II error?

We could increase the significance level α .

SECTION 10: LINEAR REGRESSION

(HAWKES 13.1, 13.2, 13.3, 13.5)

Can we describe net worth as a function of age? Is there a relationship between the number of homes sold and interest rates? Can we use crude oil prices to predict gasoline prices? In this section, we will learn a method of statistical inference that will allow us to answer questions such as these.

The goal of regression analysis is to explore the relationship between **explanatory** or independent **variables** and corresponding **response** or dependent **variables**. In this course we will only consider bi-variate data where we have one explanatory variable x , and one response variable y .

(Explanatory and response variables were introduced in Section 1.)

Example 10.1: Relationship between Sales Volume and Advertising

A company wishes to study the relationship between sales volume and the amount of money spent on advertisements.

- A. What is the **explanatory** variable x ?

Amount spent on advertisements

- B. What is the **response** variable y ?

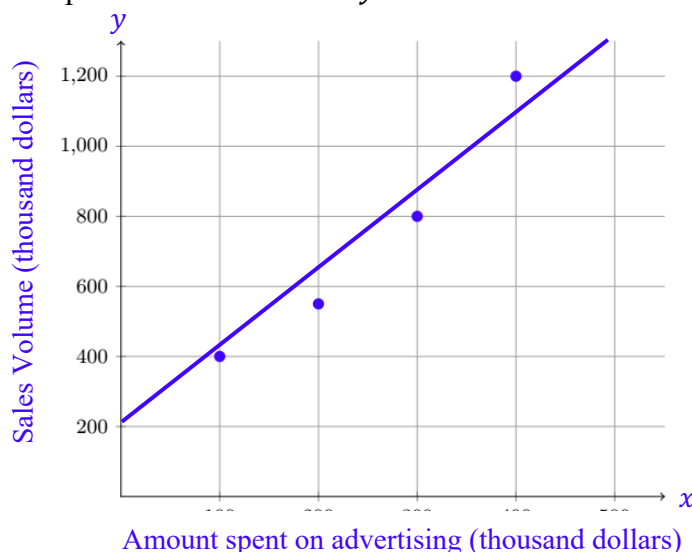
Sales volume

Advertising (thousand dollars)	Sales Volume (thousand dollars)
100	400
200	550
300	800
400	1200

- C. Plot the data from the table on the coordinate plane. Label the x and y axes with the corresponding variables.

We read this graph as “sales volume versus advertising,” or “ y versus x .”

Demonstrate drawing a line through the data points.



10.1 SIMPLE LINEAR REGRESSION MODEL

The **simple linear regression model** is an equation that describes a linear relationship between two variables for a population.

Formula 10.1: Simple Linear Regression Model

For an explanatory variable x and a response variable y , the **simple linear regression model** is made up of two parts, a predictable part which is the equation of a line and an unpredictable part which is the random error.

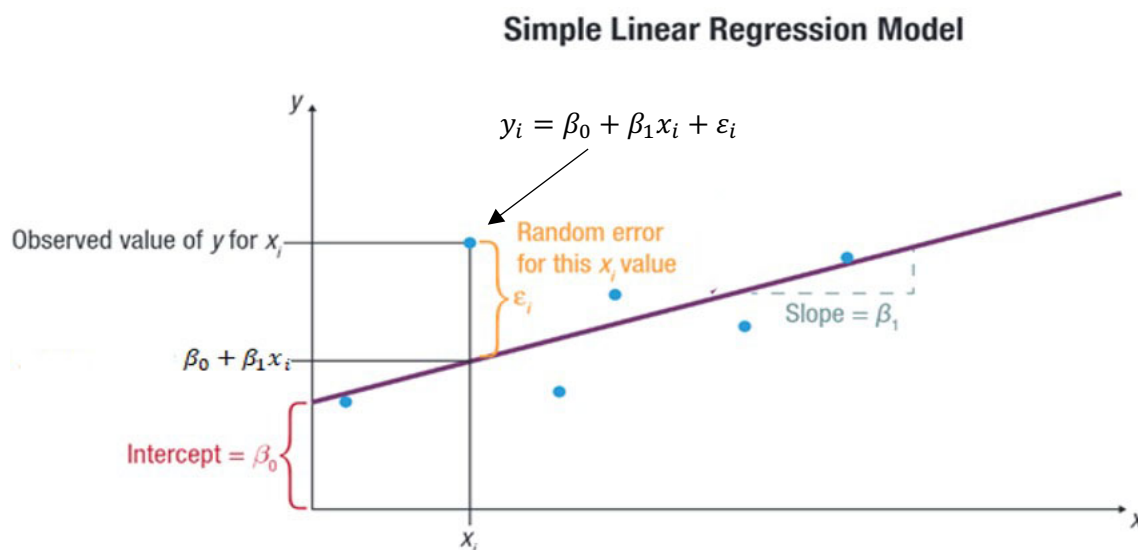
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma)$$

where...

- y_i is the i^{th} observation of the response variable
- x_i is the i^{th} observation of the explanatory variable
- β_0 is the y -intercept, the true but unknown parameter from the population
- β_1 is the slope of the line, the true but unknown parameter from the population
- ε_i is the random error which is assumed to follow a normal distribution with mean 0 and standard deviation σ .

The error terms or deviations from the line are the vertical distances between the line $\beta_0 + \beta_1 x_i$ and the observed value of y_i as demonstrated in the following figure. They result from any other factors, known or unknown, between the explanatory and response variables in the context of the problem, for example measurement error.¹

We can visualize the simple linear regression model using the figure below.²



¹ page 533 of *Statistical Methods and Data Analysis* by R. Lyman Ott and Michael Longnecker.

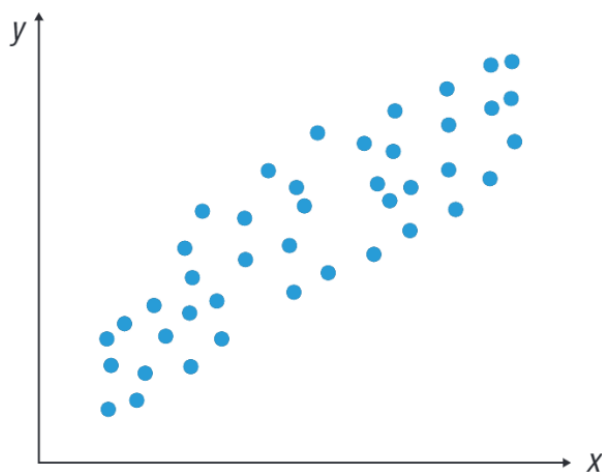
² Figure 13.1.2 on page 754 of *Discovering Business Statistics* by Quinton J. Nottingham and James S. Hawkes.

Since we will rarely expect to see data points fall in exactly a straight line because of the existence of the random error term, ε_i , in Formula 10.1, linear relationships are used to describe general trends in our data.

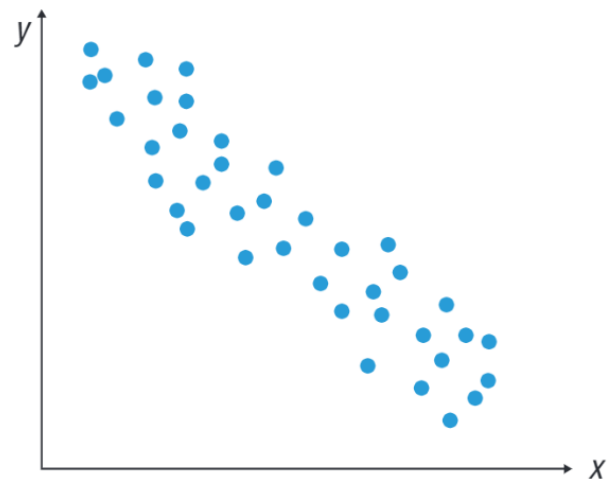
As we read a graph of y versus x from left to right, we ask ourselves: What happens to the value of y as the value of x increases?

Example 10.2: Describing Trends in Data

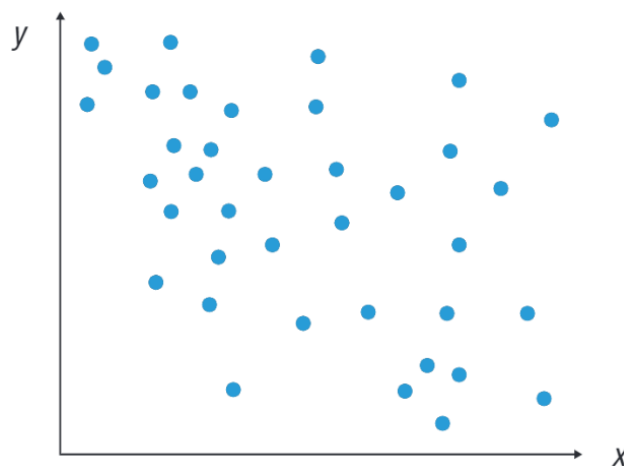
Suppose you want to analyze the relationship between an explanatory variable x and a response variable y . Describe the general trends shown in the following scatterplots³ of data points.



As x increases, y increases. (positive slope)



As x increases, y decreases. (negative slope)



As x increases, there is no observable trend in y .

³ Figure 13.1.3 on page 754 of *Discovering Business Statistics* by Quinton J. Nottingham and James S. Hawkes.

Because it is not practical to observe every possible value of either the explanatory or response variable to analyze whether a linear relationship exists and identify the y-intercept and slope parameters, we can use a sample of data to approximate the simple linear regression model. We do so using the **estimated simple linear regression (SLR) equation**, which is sometimes called the *least squares regression line* (LSRL).

Formula 10.2: Least Squares Regression Line (LSRL)

For a sample of data consisting of observations of an explanatory variable x and a response variable y , the **least squares regression line (LSRL)** is

$$\hat{y}_i = b_0 + b_1 x_i, \quad a \leq x_i \leq b,$$

where $\hat{y}_i$ is the predicted value of y for a given value x_i , b_0 is the estimate of the y -intercept of the regression line, b_1 is the estimate of the slope coefficient of the regression line, and a and b represent the minimum and maximum observed x values, respectively.

The values $\mathbf{b}_0$ and $\mathbf{b}_1$ are point estimates of the parameters β_0 and β_1 , respectively.

The **estimated y-intercept, b_0** , is the average predicted value of y when x is zero. This is only reasonable when $x = 0$ falls within the range of x values from the sample data or at least is very close to the range of x values making it a meaningful value as an explanatory variable.

We say that interpreting the y -intercept is valid only when $x = 0$ “is within the scope of the problem”.

We can interpret the **slope** as the average rate of change of the response (dependent) variable for every unit increase of the explanatory (independent) variable. The units for slope are the units for the dependent variable divided by the units of the independent variable.

The **estimated slope, b_1** , can be interpreted in the following manner:

For every additional _____ in _____, the value
 (one unit increase) (the explanatory variable x)
 of _____ _____ on average by _____.
 (the response variable y) (increases / decreases) (slope with units)

If $b_1 > 0$, then the value of y **increases** with an increase in x .

If $b_1 < 0$, then the value of y **decreases** with an increase in x .

Example 10.3: Understanding the Least Squares Regression Line

Suppose that the following estimated regression line was determined to predict an employee's salary (in dollars \$), y , based on the number of years of experience, x , that the employee has.

$$\hat{y} = 25689.10 + 2148.35x, \quad 0 \leq x \leq 35$$

- A. What is the dependent variable?

Salary

- B. What is the independent variable?

Years of experience

- C. What is the value that estimates β_0 in this equation? Use appropriate symbol and units.

$$b_0 = \$25,689.10$$

- D. What is the value that estimates β_1 in this equation? Use appropriate symbol and units.

$$b_1 = \$2,148.35/\text{year}$$

- E. What is the predicted average salary for an employee with 15 years of experience?

$$\hat{y} = 25689.10 + 2148.35(15) = \$57,914.35$$

- F. Interpret the slope in this situation.

For every additional year of experience, salary increases on average by \$2,148.35.

- G. Interpret the y-intercept in this situation.

For a person with no years of experience ($x = 0$), the estimated average salary is \$25,689.10.

10.2 RESIDUALS IN PREDICTION

As mentioned previously, when we find a least squares regression line, we do not expect all of our data points to fall on the line exactly because of the existence of the random error terms, ε_i , in Formula 10.1. Though the error terms are random and include an unknown parameter σ , some of their properties can still be captured by the difference between observed and predicted values. We call the difference between an observed value of y_i and the predicted value $\hat{y}_i$ found by the LSRL the **estimated error**, or residual.

Formula 10.3: Estimated Error (or Residual)

For an observed value of the response variable y_i and a predicted value $\hat{y}_i$ found by the least squares regression line, the **estimated error** (or **residual**) is given by

$$\text{Estimated Error or Residual} = e_i = \text{Observed } y - \text{Predicted } y = y_i - \hat{y}_i.$$

10.3 SUM OF SQUARED ERRORS

When we have a set of data points, there are many possible lines we could draw through the points to estimate the linear relationship. How do we know which line is the best fit?

One criterion we can use to determine the line of best fit is by utilizing a quantity known as the **sum of squared errors (SSE)**.

Formula 10.4: Sum of Squared Errors (SSE)

The **sum of squared errors (SSE)** is given by

$$\text{SSE} = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2.$$

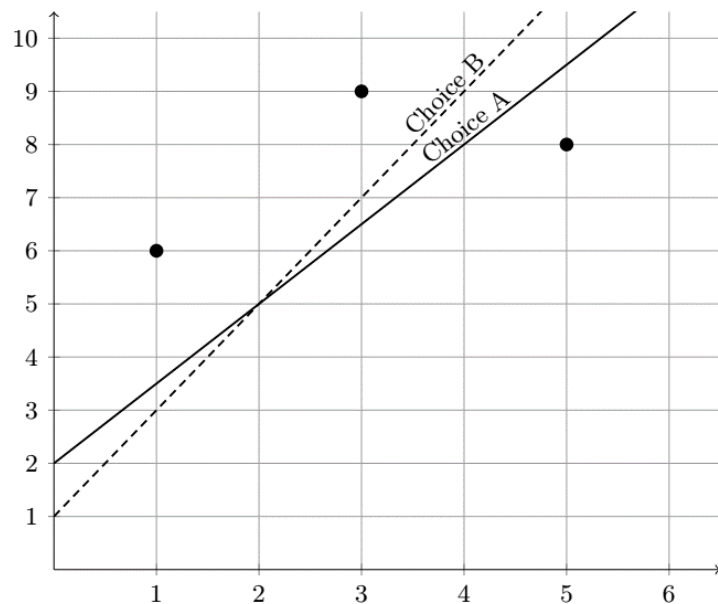
Note, the correct name should be sum of squared **estimate** of errors. We know the error terms ε_i are random with an unknown parameter, so the estimated errors (residuals) e_i can only be used to estimate ε_i . However, in the long history of statistics, the term sum of squared errors has been used by convention.

Example 10.4: Finding the Line of Best Fit Using SSE

Given the data points (1,6), (3,9), and (5,8), determine which of the following lines serves as a better fit for the data:

Choice A: $\hat{y} = 2 + 1.5x$, $x = 1, 3, 5$

Choice B: $\hat{y} = 1 + 2x$, $x = 1, 3, 5$

**Choice A**

x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	6	$2 + 1.5(1) = 3.5$	$6 - 3.5 = 2.5$	6.25
3	9	$2 + 1.5(3) = 6.5$	$9 - 6.5 = 2.5$	6.25
5	8	$2 + 1.5(5) = 9.5$	$8 - 9.5 = -1.5$	2.25
Sum of Squared Errors (SSE)				14.75

Choice B

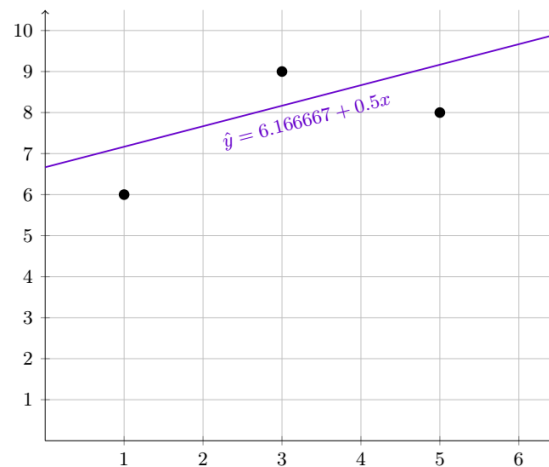
x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	6	$1 + 2(1) = 3$	$6 - 3 = 3$	9
3	9	$1 + 2(3) = 7$	$9 - 7 = 2$	4
5	8	$1 + 2(5) = 11$	$8 - 11 = -3$	9
Sum of Squared Errors (SSE)				22

Therefore, choice A is a better fit because it has a smaller SSE.

From Example 10.4 we observe that different regression lines result in different Sums of Squares Errors (SSEs). Also, it should be clear that there exist infinitely many ways to draw such lines. How do we know which of these many lines is the best line? The term “least squares” part of the least squares regression line refers to finding the line of best fit by minimizing the sum of squared errors. For example, the least squares regression line (LSRL) for the data points in Example 10.4 is

$$\hat{y} = 6.166667 + 0.5x, \quad 1 \leq x \leq 5,$$

which has a SSE value of 2.666667.



Most textbooks provide formulas to use to calculate the best value for estimated y-intercept b_0 and estimated slope b_1 based on minimizing the value of SSE. However, the calculations are complicated and tedious to find by hand. We will utilize technology to efficiently find b_0 and b_1 of the LSRL. Below is the JMP output for the previous example.

Linear Fit

Linear Fit

$y = 6.1666667 + 0.5x$

Summary of Fit

RSquare	0.428571
RSquare Adj	-0.14286
Root Mean Square Error	1.632993
Mean of Response	7.666667
Observations (or Sum Wgts)	3

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	2.0000000	2.00000	0.7500
Error	1	2.6666667	2.66667	Prob > F
C. Total	2	4.6666667		0.5456

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	6.1666667	1.972027	3.13	0.1970
x	0.5	0.57735	0.87	0.5456

Least squares regression line (LSRL)

Sum of squared errors

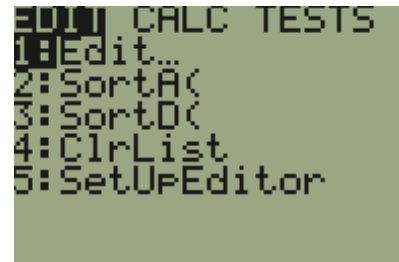
b_0 : estimate of y-intercept

b_1 : estimate of slope

Calculator Tutorial 10.1: Finding the Least Squares Regression Line

The TI-83 and TI-84 calculators have the ability to find the LSRL from a data set.

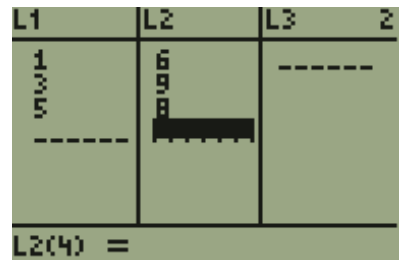
1. Click STAT and select the first option, 1: Edit...



```

EDIT  [CALC] TESTS
1:Edit...
2:SortA(
3:SortD(
4:ClrList
5:SetUpEditor
  
```

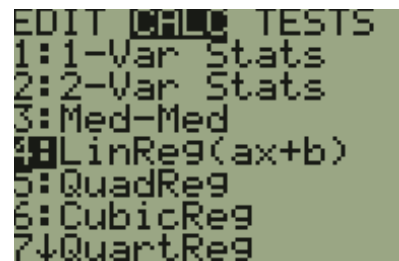
2. Enter the data points for the independent variable x in L1 and those for the dependent variable y in L2.



L1	L2	L3	2
1	6		
2	9		
3	8		
4	7		
5	5		

L2(4) =

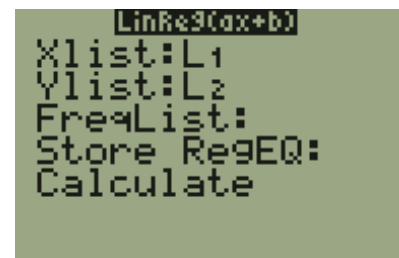
3. Click the STAT button again and navigate to the CALC menu. Select item 4: LinReg(ax + b).



```

EDIT  [CALC] TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7:QuartReg
  
```

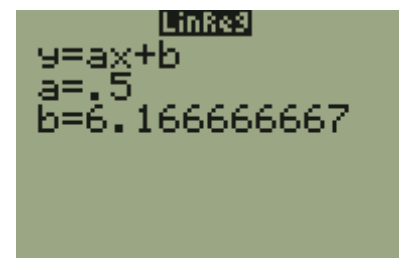
4. For Xlist, enter L1, and for Ylist, enter L2. Leave the other options blank. Use the down arrow keys to select Calculate and click ENTER.



```

LinReg(ax+b)
Xlist:L1
Ylist:L2
FreqList:
Store RegEQ:
Calculate
  
```

The value a represents the slope b_1 and the value b represents the y -intercept b_0 .



```

LinReg
y=ax+b
a=.5
b=6.166666667
  
```

MY NOTES:

10.4 FITTING A SIMPLE LINEAR REGRESSION MODEL

In order to fit a **simple linear regression model** to estimate a linear relationship between one explanatory variable (x) and one response variable (y)...

- (1) Take a **sample** of data consisting of measurements of x and y .
- (2) Create a **scatterplot** of the sample data. Determine whether the points seem to fall in a linear pattern. (Does it seem reasonable to fit a linear model to the data?)
- (3) Fit a **Simple Linear Regression** to the sample data.

Example 10.5: Fitting a Regression Model using JMP

An economist wants to determine the relationship between one's FICO score, x , and the interest rate of a 36-month auto loan, y . The data in the table below represent the interest rate (in percentage points) that a bank might offer on a 36-month auto loan for a sample of FICO scores.

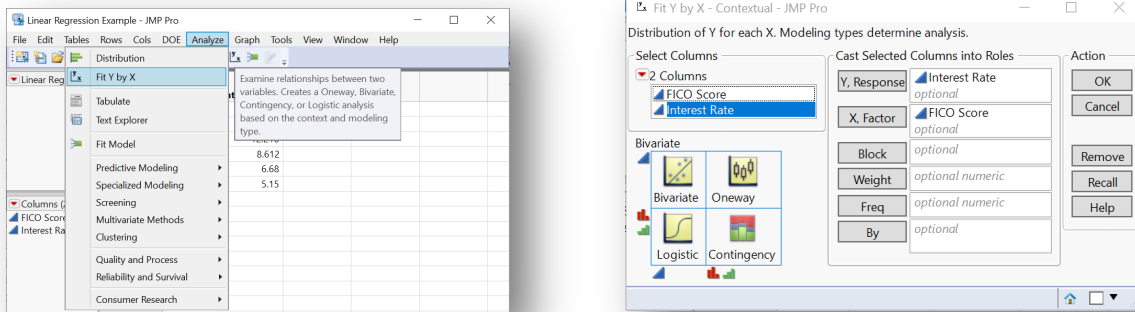
STEP 1: TAKE A SAMPLE OF DATA CONSISTING OF MEASUREMENTS OF x AND y .

Take a Sample from Population		Enter them to JMP	
FICO Score (points)	Interest Rate (percentage points)	FICO Score	Interest Rate
545	18.982	545	18.982
595	17.967	595	17.967
640	12.218	640	12.218
675	8.612	675	8.612
705	6.680	705	6.68
750	5.150	750	5.15

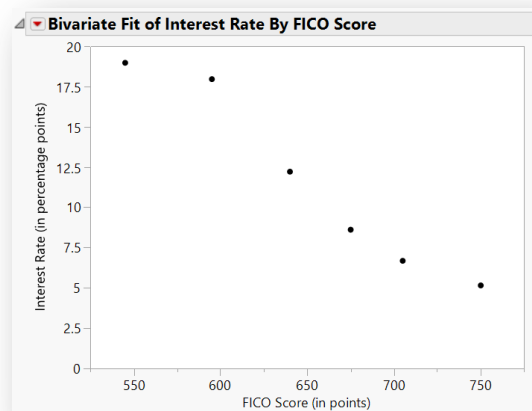
STEP 2: CREATE A SCATTERPLOT OF THE SAMPLE DATA

Do the data points seem to fall in a linear pattern? Does it seem reasonable to fit a linear model to the data?

To create a scatterplot in JMP, select **Analyze >> Fit Y by X**. Click and drag your response and explanatory variables from the **Select Columns** menu to the fields **Y, Response** and **X, Factor**, respectively.



Click **OK**. This will generate a scatterplot of the data, Interest Rate vs. FICO Score.

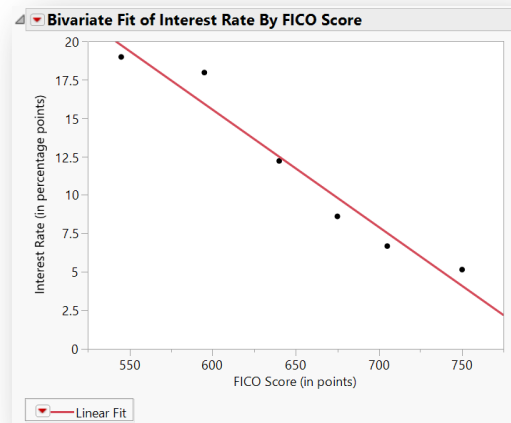
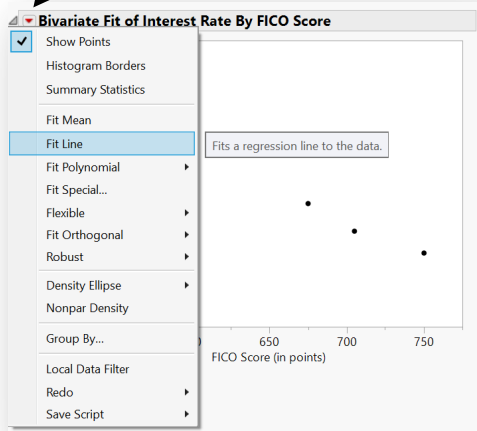


Note: You may wish to adjust your axis values as appropriate. (Here, the y-axis is adjusted to begin at zero.) You can access these settings by right-clicking on the axis you wish to edit and selecting **Axis Settings**.

You can also change the vertical and horizontal axis to add units by double-clicking on the axis title you wish to edit and typing the new label.

STEP 3: FIT A SIMPLE LINEAR REGRESSION TO THE SAMPLE DATA.

Click the **triangle** next to the graph title and select **Fit Line**. The **least squares regression line (LSRL)** will now be shown on your scatterplot.



The output for your **regression analysis** will be printed below the scatterplot. Note that you can find the estimated y -intercept and slope in both the **Linear Fit** equation and the **Parameter Estimates** table.

Linear Fit				
Interest Rate = 61.36874 - 0.0763692*FICO Score				
Summary of Fit				
RSquare		0.952323		
RSquare Adj		0.940404		
Root Mean Square Error		1.424099		
Mean of Response		11.6015		
Observations (or Sum Wgts)		6		
Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	162.03933	162.039	79.8987
Error	4	8.11223	2.028	Prob > F
C. Total	5	170.15157		0.0009*
Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	61.36874	5.597944	10.96	0.0004*
FICO Score	-0.076369	0.008544	-8.94	0.0009*

Example 10.6: Describing the Regression Equation Using JMP Output

Use the JMP output from your regression analysis to answer the following.

- A. Define the least squares regression line equation.

$$\text{Estimated Interest Rate} = 61.36874 - 0.0763692(\text{FICO Score}), 545 \leq \text{FICO Score} \leq 750$$

$$\text{or } \hat{y} = 61.36874 - 0.076392x, \quad 545 \leq x \leq 750$$

- B. Identify the slope, using the appropriate symbol.

$$b_1 = -0.07639692 \text{ percentage point/point}$$

- C. Interpret the slope.

For every additional point on an individual's FICO score, the interest rate decreases on average by 0.07637 percentage points.

- D. Identify the y-intercept, using the appropriate symbol.

$$b_0 = 61.36874 \text{ percentage points}$$

- E. Interpret the y-intercept.

A FICO score of zero is not within the scope of the problem since data was only collected for FICO scores between 545 and 750 points. Therefore, there is no valid interpretation of the y-intercept in terms of this problem.

10.5 EVALUATING THE FIT OF THE SIMPLE LINEAR REGRESSION MODEL

There are several items to consider when determining if the simple linear regression model is a good fit for your data.

- (1) Standard Error
- (2) Coefficient of Determination
- (3) Correlation Coefficient
- (4) Residual Analysis
- (5) Value of Slope β_1

10.5.1 STANDARD ERROR

The **Sum of Squared Errors (SSE)** we learned in Formula 10.4 is the most intuitive measurement to determine how well the model fits our data. The small SSE means the fitted line comes very close to all of the data points, and therefore, the model is predicting values accurately. However, since SSE is relevant to the number of observations n , we would like to find a more standardized measurement which can reflect the “average” distance between observed and predicted values.

The **standard error** (also called **standard error of prediction** or **standard error of the model**), denoted s , is a good candidate to do this job. It is the standard deviation of estimate of errors (residuals) which can be calculated from **variance of errors**, s^2 :

Formula 10.5: Variance of errors

$$s^2 = \frac{\sum (y_i - \hat{y})^2}{n - 2} = \frac{SSE}{n - 2}$$

This looks very similar to the formula for variance we used earlier in the semester. Notice that the denominator is $n - 2$ rather than n or $n - 1$. This is from the degrees of freedom which won't be introduced in STAT 3090.

To find s simply take the square root.

Formula 10.6: Standard Error

$$s = \sqrt{s^2}$$

From the formulas we see, the standard error s represents the average distance that the observed values of y vary from the predicted values of y . In a linear regression model, we want the value of s to be small in comparison to the observed y values.

In the JMP output from your regression analysis, the value of s is labeled as the **Root Mean Square Error** in the **Summary of Fit** table:

Summary of Fit	
RSquare	0.952323
RSquare Adj	0.940404
Root Mean Square Error	1.424099
Mean of Response	11.6015
Observations (or Sum Wgts)	6

We can interpret values of standard error in the following manner.

[illegible]

Example 10.7: Interpreting Standard Error

Interpret the standard error found by your regression analysis.

$s = 1.424099$

For any given FICO score, the observed interest rate should vary from the prediction on average by 1.424 percentage points.

10.5.2 COEFFICIENT OF DETERMINATION

The **coefficient of determination**, R^2 , or r^2 is another commonly used measurement to determine how well our model fits the data. To understand r^2 , we need to understand the essence of the variability of the data.

For a certain sample with observed values of y_i , we can calculate its mean $\bar{y}$. Then, the total variability (called **Sum of Squares Total** or **SST**) of this data set is calculated by:

$$SST = \sum (y_i - \bar{y})^2$$

Recall how we calculate the variance for y_i 's, we just don't take the average here.

Suppose we have a linear regression model to fit this data set, the predicted values $\hat{y}_i$'s can be obtained. Then, the total variability of the data set can be partitioned into two kinds of variabilities – the variability from regression (our model), and the variability from random errors ε_i (see Formula 10.1):

- The variability from regression (called **Sum of Squares Due to Regression** or **SSR**) is calculated by:

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

- The variability from random errors is just the **Sum of Squared Errors (SSE)** we learned in Section 10.3 and calculated by Formula 10.4:

$$SSE = \sum (y_i - \hat{y}_i)^2$$

Therefore,

$$SST = SSR + SSE$$

Now we can define the coefficient of determination r^2 , which is the [*proportion of variation*](#) in y that can be explained by x using this model:

Formula 10.7: Coefficient of Determination r^2

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

If the model fits the data well, the value of r^2 should be close to 1 meaning most of the variation of y can be captured by this model. Alternatively, you can think it as: a good model should predict y 's accurately, i.e., have small SSE. Since the SST of a certain data set is fixed, and $SSR = SST - SSE$, this is equivalent to having a large SSR close to SST, and consequently, the value of r^2 is close to 1.

The **SSR**, **SSE**, and **SST** can be found in the **Sum of Squares** column in the **Analysis of Variance table**, labeled by **Model**, **Error**, and **C. Total** respectively:

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	162.03933	162.039	79.8987
Error	4	8.11223	2.028	Prob > F
C. Total	5	170.15157		0.0009*

We can calculate the value of r^2 by Formula 10.7 using these values, or directly get it from **Summary of Fit** table, labeled by **RSquare**:

Summary of Fit	
RSquare	0.952323
RSquare Adj	0.940404
Root Mean Square Error	1.424099
Mean of Response	11.6015
Observations (or Sum Wgts)	6

You can verify that the r^2 obtained from Analysis of Variance table using Formula 10.7 is the same as that of in Summary of Fit table.

We can interpret the value of r^2 as follows:

Approximately _____% of the variation in _____ can be
 explained by _____.
 (the explanatory variable x) (the response variable y)

Example 10.8: Interpreting the Coefficient of Determination

Interpret the coefficient of determination in the context of your regression analysis.

$$r^2 = 0.9523$$

Approximately 95.23% of the variation in interest rates can be explained by FICO score.

(Note that this a percent variation – it is the same regardless of the units of y in context.)

10.5.3 CORRELATION COEFFICIENT

If you take the square root of the r^2 value (in decimal form), it will give you the value of r , the **correlation coefficient**. The sign on the value of r is dependent on the sign of the estimated slope of the regression equation, b_1 . A positive b_1 (or increasing trend of data) yields a positive r ; a negative b_1 (or decreasing trend of data) yields a negative r .

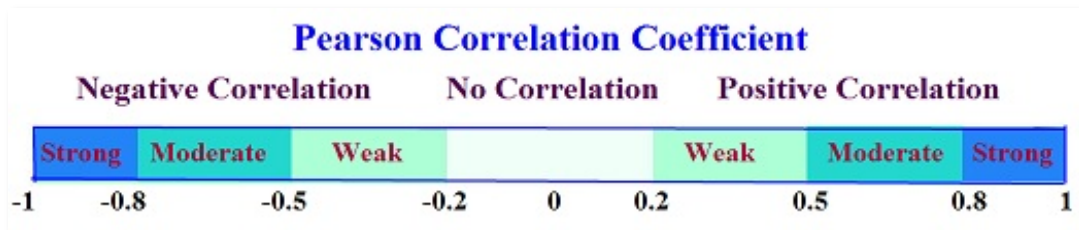
Example 10.9: Finding the Correlation Coefficient

Find the correlation coefficient from your regression analysis.

$$r = \sqrt{0.9523} = -0.9759 \quad (\text{because slope } b_1 < 0)$$

There are three important information which can be interpreted from the correlation coefficient:

- (1) **Direction** of the relationship is determined by the sign of r (positive or negative).
- (2) **Strength** is determined by the value of r using the Pearson correlation coefficient scale.



- (3) **Linearity:** The correlation coefficient is not a measure of just any relationship; it is specifically a measurement of the strength and direction of the [linear relationship](#) between x and y . Two variables can be highly correlated (e.g., they are quadratically dependent, not introduced in STAT 3090) without having a strong r value, so it is important that our interpretation include the word **linear**.

Therefore, we can interpret the r as follows:

There is a _____ linear relationship between _____ and _____.

(strength of relationship) (positive / negative) (the explanatory variable x) (the response variable y)

Example 10.10: Interpreting the Correlation Coefficient

Interpret the correlation coefficient in context of your regression analysis.

$$r = -0.9759$$

There is a strong negative linear relationship between FICO score and interest rate on a 36-month car loan.

MY NOTES:

10.5.4 RESIDUAL ANALYSIS

Just as in our other inferential methods of making confidence intervals and performing hypothesis tests, there are certain assumptions or conditions that must be met for our estimated regression model to be valid.

- (1) There is linear relationship between the independent and dependent variables
- (2) The error terms, ε_i , are normally distributed
- (3) The error terms, ε_i , are independent
- (4) The error terms, ε_i , have a mean of 0
- (5) The error terms, ε_i , have a constant variance

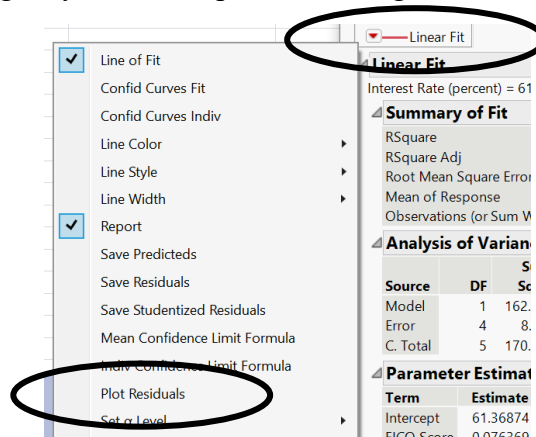
where ε_i are the random errors (see Formula 10.1).

We need to check all these five conditions by data set or model. If any of them is violated, the linear regression model may not be suitable for your data set. In that situation, you may need to apply some advanced models such as quadratic model or autoregressive model which will not be introduced in STAT 3090.

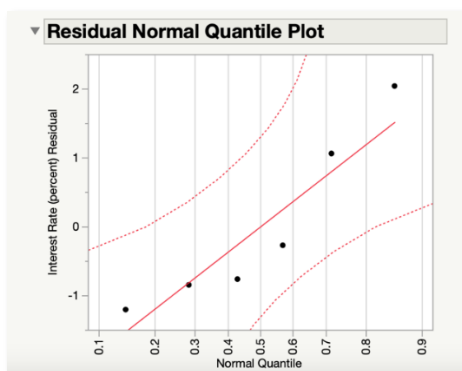
To check the first condition, we can make a [scatterplot](#) of the data set to see whether there's a linear trend. You can also do this after calculating the linear regression equation, since a correlation coefficient r close to 1 or -1 is also able to verify such linear relationship as we have already learned.

As we mentioned before, since the estimated error (residual) can capture the properties of error terms, for the conditions (2)-(5) (sometimes also (1)), we can use [graphical methods](#) for a **residual analysis** to determine if these conditions are met. JMP can make several diagnostic plots once we have an estimated linear regression equation.

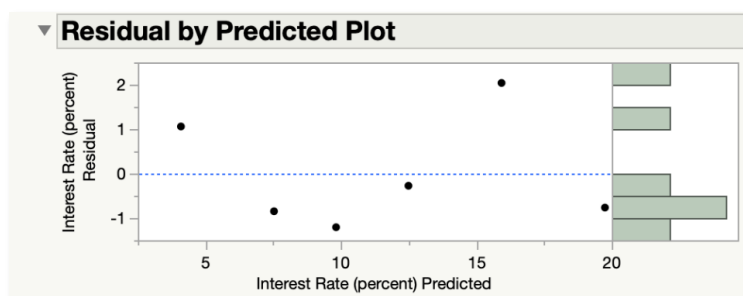
For the second condition, we create a [normal quantile plot](#) (also called normal probability plot) of the residuals. To have JMP draw this graph, select the **triangle** next to **Linear Fit** and choose **Plot Residuals**. JMP will give you several plots including a residual normal quantile plot.



In a residual normal quantile plot, if the residuals roughly align on a straight line without going beyond the curved boundaries, the second condition is met. For example, as it is shown below, the FICO scores meet this condition.



For conditions (3)-(5), we will make a scatterplot of the residuals vs the predicted y values. This graph is also included when you ask JMP to **Plot Residuals**. In this plot, if the data points are randomly scattered around the line $y = 0$ without showing any special patterns, the conditions (3)-(5) are met. For example, as it is shown below, the FICO scores meet these conditions.



Below are some special patterns in the residual vs predicted plot that can indicate there is a problem with your linear regression model.

The model has a non-linear relationship that was not accounted for. This linear model had $r^2 = 0.8225$	The errors do not have constant variance - note the fan shape. This linear model had $r^2 = 0.8526$.	There is an outlier that is potentially skewing the estimates. This linear model had $r^2 = 0.5519$.

10.5.5 VALUE OF SLOPE β_1

Another way to evaluate if the model is useful is to determine whether the slope in the linear model is significantly different from 0. Recall that the simple linear regression model in Formula 10.1:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where the y -intercept β_0 and slope β_1 are **parameters** from population, which can be estimated by b_0 and b_1 respectively.

The ideal is simple - if β_1 is not significantly different from zero, there's no evidence to indicate that y changes with x linearly (recall the equation with slope equals zero is a flat line), and therefore our linear model is a kind of useless.

Then, how can we determine whether the slope β_1 is “significantly different from zero”? Either a confidence interval or a hypothesis test is a good approach. We will introduce the confidence interval approach here, and you can explore the hypothesis test approach on your own.

To determine whether the population slope β_1 is significantly different from zero, we can construct a $(1 - \alpha)100\%$ confidence interval for it. If the confidence interval contains 0, then β_1 is *not significantly different* from 0 at $\alpha\%$ level; if the confidence interval doesn't contain 0, then β_1 is *significantly different* from 0 at $\alpha\%$ level.

Formula 10.8: Confidence Interval for Slope β_1

Consider a simple linear regression model estimated by the least squares regression line $\hat{y} = b_0 + b_1 x$ for an explanatory variable x and response variable y . A **$(1 - \alpha)100\%$ confidence interval for β_1** , the population slope of the regression model, is

$$b_1 \pm (t_{\alpha/2, n-2})(s_{b_1}),$$

where b_1 is the point estimate of the slope from the LSRL, s_{b_1} is the standard error of the slope, and $t_{\alpha/2, n-2}$ is a t -critical value with $n - 2$ degrees of freedom.

The **standard error of the slope**, s_{b_1} , is the standard deviation of the distribution of sample slopes, and the **variance of the slope** is simply $s_{b_1}^2$.

The calculation of s_{b_1} is quite tedious which will not be covered in STAT 3090, but fortunately, JMP can do this job for us. You can find s_{b_1} by looking at the **Parameter Estimates** table in the **Std Error** column and the row that corresponds with your **explanatory variable** x . For example, the table below shows that the s_{b_1} of FICO scores is 0.008544.

Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	61.36874	5.597944	10.96	0.0004*
FICO Score	-0.076369	0.008544	-8.94	0.0009*

While similar to the confidence intervals we constructed previously, there is a slight variation in the interpretation of confidence intervals for the population slope β_1 .

We are _____% confident that the population slope of the regression line using
(confidence level * 100)

_____ to predict _____ falls between _____ and _____.
(explanatory variable x) (response variable y) (lower bound) (upper bound)

Example 10.11: Confidence Interval for the Slope β_1

Consider the slope of the equation $\hat{y} = 61.36874 - 0.076392x$ from your regression analysis.

Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	61.36874	5.597944	10.96	0.0004*
FICO Score	-0.076369	0.008544	-8.94	0.0009*

- A. Identify the standard error of the slope. Use appropriate symbol.

$$s_{b_1} = 0.008544$$

- B. Calculate a 90% confidence interval for β_1 . Recall there are 6 data points ($n = 6$).

$$b_1 = -0.076369, t_{.05,4} = 2.132, s_{b_1} = 0.008544$$

$$b_1 \pm (t_{\alpha/2, n-2})(s_{b_1}) = -0.076369 \pm (2.132)(0.008544) = [-0.0946, -0.0582]$$

- C. Interpret the confidence interval found in Part B.

We are 90% confidence that the population slope of the regression line using FICO score to predict interest rate is between -0.0946 and -0.0582 .

- D. Based on your confidence interval, would you say the slope is significantly different than zero at 10% level, indicating a linear relationship?

Yes, since zero does not fall in the 90% confidence interval.

10.6 USING THE SIMPLE LINEAR REGRESSION MODEL

If you determine that the model fits well, we can use the regression equation to predict future values of the response variable based on a given value of the explanatory variable.

For the FICO scores example, we can see that the model fits well since:

- (1) $s = 1.424$, which is fairly small compared to our y -values ranging from 5 to 19,
- (2) $r^2 = 0.9523$ or 95.23%, which is close enough to 1,
- (3) $r = -0.9759$ indicating there is a strong negative linear relationship between FICO scores and the interest rate,
- (4) The residual analysis shows that the assumptions for a linear regression model have been met,
- (5) Zero does not lie within the CI for slope: $[-0.0946, -0.0582]$.

Therefore, we can use the model for prediction.

Example 10.12: Using the LSRL for Prediction

Find the predicted interest rate on a 36-month auto loan for a person with a FICO score of 640 using the equation $\hat{y} = 61.36874 - 0.076392x$.

$$\hat{y} = 61.36874 - 0.076392(640) = 12.478 \text{ percentage points}$$

Note: We should only use our model to predict values of the response variable that fall within the range of observed values of the explanatory variable. Trying to predict the value of y for a value of x that falls outside of this range is called extrapolation and should be avoided, as we cannot be certain of the behavior of data that fall outside of the range of observations in our sample.

MY NOTES:

PRACTICE EXERCISES FROM SECTION 10

1. A company wishes to predict sales volume, y , based on its advertising expenditures, x . The sales manager used previous years' data to estimate the following linear equation. Both sales volume and advertising expenditures are measured in thousands of dollars.

$$\hat{y} = 49.25 + 0.51x, \quad 0 \leq x \leq 100$$

- (a) What is the dependent (response) variable in this model?

Sales Volume

- (b) What is the independent (explanatory) variable in this model?

Advertising Expenditures

- (c) What is the estimated sales volume for this company when the marketing department spends \$40,000 on advertising?

$$\hat{y} = 49.25 + 0.51(40) = 69.95 \text{ thousand dollars} = \$69,950$$

- (d) If the company has a target sales volume of \$100,000, how much should the sales manager allocate for advertising in the budget based on this model?

$$100 = 49.25 + 0.51x \Rightarrow x = 99.5 \text{ thousand dollars} = \$99,500$$

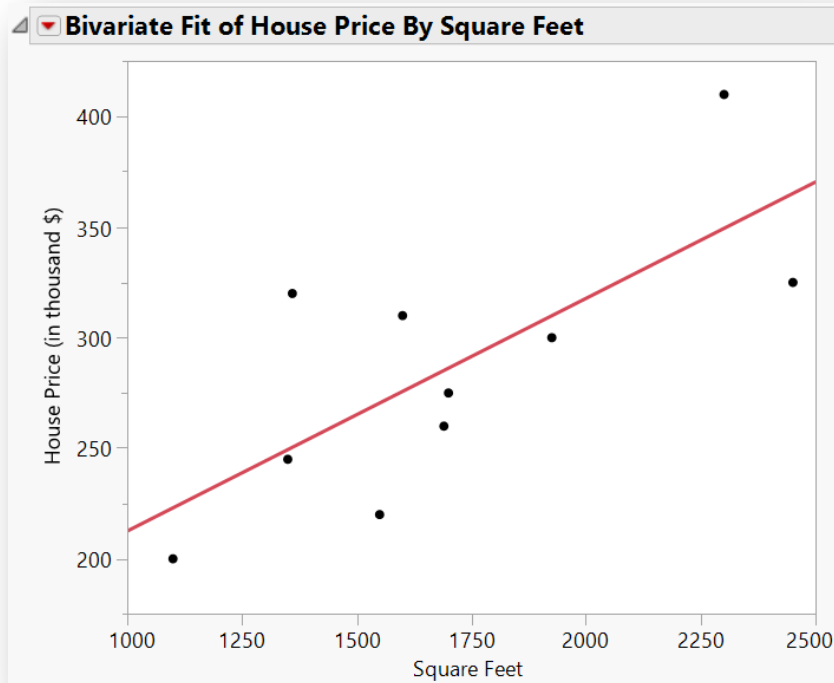
- (e) Interpret the value of the estimated slope in this situation.

$b_1 = 0.51$: For every additional \$1,000 spent on advertising, the estimated sales volume will increase by an average of \$510.

- (f) Interpret the value of the estimated y -intercept in this situation.

$b_0 = 49.25$: When the company does not spend any money on advertising, the average sales volume is \$49,250.

2. A local real estate agency would like to be able to predict the price of a house (y) in the Clemson area using the size of the house in square feet (x). Data are collected from a sample of ten houses from 1100 ft² to 2450 ft² for sale in the Clemson area and a LSRL is fitted to the data. Use the following JMP output to answer the questions below.



Summary of Fit

RSquare	0.538403
RSquare Adj	0.480704
Root Mean Square Error	43.67053
Mean of Response	286.5
Observations (or Sum Wgts)	10

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	17795.577	17795.6	9.3311
Error	8	15256.923	1907.1	Prob > F
C. Total	9	33052.500		0.0157*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	107.16167	60.31146	1.78	0.1135
Square Feet	0.1053382	0.034484	3.05	0.0157*

- (a) Define the least squares regression line equation. (You can find the y -intercept and slope in the Parameter Estimates table.) Be sure to include the range of observed values of the explanatory variable.

Estimated House Price = $107.16167 + 0.1053382 * (\text{Square Feet})$,
for $1100 \leq \text{Square Feet} \leq 2450$

$$\text{or } \hat{y} = 107.1617 + 0.1053x, 1100 \leq x \leq 2450$$

- (b) Identify the slope of the LSRL using the appropriate symbol and interpret its value.

$b_1 = 0.1053382$. For every square-foot increase in house size, the price of the house increases on average by 0.1053 thousand dollars (or \$105.30).

- (c) Identify the y -intercept of the LSRL and interpret its value.

$b_0 = 107.16167$. Since a house size of zero square feet is outside the scope of the problem (i.e. data was only collected for houses between 1100 ft² and 2450 ft²), there is not a valid interpretation for slope in this context.

- (d) Find and interpret the standard error found by your regression analysis.

$s = 43.67053$ thousand dollars

For any given house size in square feet, the true house price will vary from the prediction on average by 43.67 thousand dollars (or \$43,670).

- (e) Find the estimated variance of errors from your regression analysis.

$$s^2 = (43.67053)^2 = 1907.1152$$

- (f) Find and interpret the coefficient of determination.

$$r^2 = 0.538403$$

Approximately 53.84% of the variation in house price can be explained by house size (in square feet).

- (g) Find and interpret the correlation coefficient.

$$r = \sqrt{0.538403} = 0.7338 \quad (\text{positive since } b_1 > 0)$$

There is a moderate positive linear relationship between house size (in square feet) and house price (in thousand dollars).

- (h) Identify the SSE for the least squares regression line in your analysis.

$$SSE = 15256.923$$

- (i) Identify the standard error of the slope.

$$s_{b_1} = 0.034484$$

- (j) Calculate and interpret a 95% confidence interval for the population slope β_1 .

$$b_1 \pm (t_{.025,8})(s_{b_1}) = 0.1053 \pm 2.306(.034484) = [0.0258, 0.1848]$$

We are 95% confident that the population slope of the regression line using house size (in square feet) to predict house price (in thousand dollars) is between 0.0258 and 0.1848.

- (k) Based on your confidence interval, would you say that the slope is significantly different than zero, indicating a linear relationship? Justify your answer.

Yes, since zero does not fall within the 95% confidence interval.

- (l) Use the LSRL to predict the price of a house that is 2,100 square feet.

$$\hat{y} = 107.1617 + 0.1053(2100) = 328.2917 \text{ thousand dollars}$$