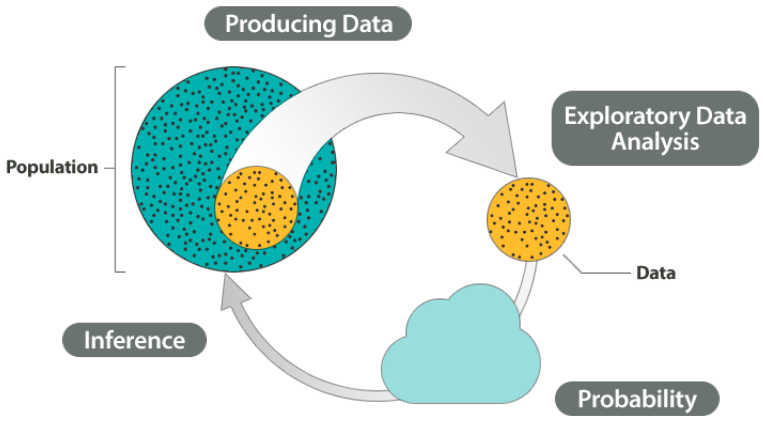
MA336 Worksheet

Dr. Ye

Last updated on 11/29/2022

1 Statistical Studies



1.1 Basic statistical concepts

- **Data** consists of information from observation, counts, measurements, responses or experiments.
- A **population** is the collection of all objects that are of interest.
- A **parameter** is a number that is a property of the population.
- A **sample** is a subset of a population.
- A **statistic** is a number, such as a percentage, that represents a property of a sample.
- In statistics, a **variable** is a characteristic, or attribute of interest that we gather about individuals or objects.
- **Categorical variables** (or qualitative variables) represent attributes, labels or nonnumerical entries, such as names, and colors.
- **Quantitative variables** represent numerical measurements or counts, such as weights and number of students in each class.

Example 1.1. Identify the population, parameter, sample, and statistic in the following study: *To learn the percentage of students go to school by public transportation, 500 students at a college were surveyed. 50% say they go to school by public transportation*

Example 1.2. Identify the type variables.
(1) Age (2) Hair color (3) GPA (4) temperature (5) Education attainment

Exercise 1.1. Identify the population, sample, the variable of study, the type of the variable, the population parameter and the sample statistics.

An administrator wishes to estimate the passing rate of a certain course. She takes a random sample of 50 students and obtains their letter grades of that course. Among the 50 students, 32 students earned a grade C or better.

1.2 Types of statistical studies

A statistical study can usually be categorized as an **observational study** or an **experiment** by the mean of study.

- An observational study observes individuals and measures variables of interest. The main purpose of an observational study is to describe a group of individuals or to investigate an association between two variables.
- An experiment intentionally manipulates one variable in an attempt to cause an effect on another variable. The primary goal of an experiment is to provide evidence for a cause-and-effect relationship between two variables.

Example 1.3. Which type of study will answer the question best.

(1) What proportion of all college students in the United States have taken classes at a community college?

(2) Does use of computer-aided instruction in college math classes improve test scores?

Exercise 1.2. Identify the type of statistical study:

(1) *A study took random sample of adults and asked them about their bedtime habits. The data showed that people who drank a cup of tea before bedtime were more likely to go to sleep earlier than those who didn't drink tea.*

(2) *Another study took a group of adults and randomly divided them into two groups. One group was told to drink tea every night for a week, while the other group was told not to drink tea that week. Researchers then compared when each group fell asleep.*

1.3 Questions about population

Type of Research Question	Examples
Make an estimate about the population (often an estimate about an *average* value or a *proportion* with a given characteristic)	What is the *average* number of hours that community college students work each week? What *proportion* of all U.S. college students are enrolled at a community college?

Type of Research Question	Examples
Test a claim about the population (often a claim about an *average* value or a *proportion* with a given characteristic)	Is the *average* course load for a community college student greater than 12 units? Do the *majority* of community college students qualify for federal student loans?
Compare two populations (often a comparison of population averages or proportions with a given characteristic)	In community colleges, do female students have a *higher* GPA than male students? Are college athletes *more* likely than non-athletes to receive academic advising?
Investigate a correlation between two variables in the population	Is there a *correlation* between the number of hours high school students spend each week on Facebook and their GPA? Is academic counseling *associated* with quicker completion of a college degree?

Exercise 1.3. Give an example of a research question that involves

(1) estimating a characteristic about all students at QCC.

(2) testing a claim of a characteristic about all students at QCC.

(3) investigating a correlation between two variables about

all students at QCC.

1.4 Question on cause-and-effect

A research question that focuses on a cause-and-effect relationship is common in disciplines that use experiments, such as medicine or psychology.

- Does cell phone usage increase the risk of developing a brain tumor?
- Does drinking red wine lower the risk of a heart attack?

In a study of a relationship between two variables, one variable is the **explanatory variable**, and the other is the **response variable**.

Example 1.4. Determine if the question is a cause-and-effect question? What are the explanatory and response variables?

(1) Does use of computer-aided instruction in college math classes improve test scores?

(2) Does tutoring correlate with improved performance on exams?

Exercise 1.4. A researcher studies the medical records of 500 randomly selected patients. Based on the information in the records, he divides the patients into two groups: those given the recommendation to take an aspirin every day and those with no

such recommendation. He reports the percentage of each group that developed heart disease.

Determine whether the study supports the conclusion that taking aspirin lowers the risk of heart attacks.

Exercise 1.5. Does higher education attainment lead to higher salary?

(1) Determine if the question is a cause-and-effect question?

(2) What are the explanatory and response variables?

(3) If a student want to study this question, what type of statistical study can be used? What kind of conclusion can be drawn?

1.5 Sampling plans

To make accurate inference, the sample must be representative of the population.

- A **sampling plan** describes exactly how we will choose the sample.
- A sampling plan is **biased** if it systematically favors certain outcomes.
- In **random Sampling**, every individual or object has an equal chance of being selected.

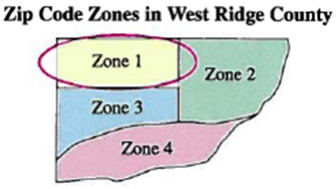
1.6 Methods of random sampling

- **Simple random sample**: groups of the same size are randomly selected. Table of random numbers, calculator and

computer programs are often used to generate random numbers.

- **Stratified random sample**: The population is first split into homogeneous groups. Then a same proportion or a same number of subjects from each group are selected randomly.

- **Cluster sample**: The population is first split into groups. Then some groups are selected randomly.



- **Systematic sample**: First, a starting number is chosen randomly. Then take every n-th piece of the data.

Exercise 1.6. Classify each of the sampling procedures below.

(1) A high school educational researcher interviews 50 high school female teachers and 50 high school male teachers.

(2) The names of 25 employees being chosen out of a hat from a company of 250 employees.

(3) A researcher starts surveying every 10th customer at a local coffee shop starting at a random time between 8 am and 9 am.

(4) A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on the aver-

age.

1.7 Bad sampling

Biased sampling

- Online polls. These are examples of a voluntary response sample.
- Mall surveys. These are an example of a convenience sample.

Undercoverage

- It occurs when some groups in the population are left out of the process of choosing a sample. For example, random survey math students to estimate the average GPA or a college.

Example 1.5. Suppose that you want to estimate the proportion of students at your college that use the library.

Which sampling plan will produce the most reliable results?

(1) Select 100 students at random from students in the library.

(2) Select 200 students at random from students who use the Tutoring Center.

(3) Select 300 students who have checked out a book from the library.

(4) Select 50 students at random from the college.

Exercise 1.7. Identify the flaw(s) in the study

Students at an elementary school are given a questionnaire that they are asked to return after their parents have completed it. One of the questions asked is, "Do you find that your work schedule makes it difficult for you to spend time with your kids after school?" Of the parents who replied, 85% said "no". Based on these results, the school officials conclude that a great majority of the parents have no difficulty spending time with their kids after school.

1.8 Elements of experimental design

To establish a cause-and-effect relationship, we want to make sure that the explanatory variable is the only thing that impacts the response variable. We therefore want to get rid of all other factors that might affect the response. These factors are called **confounding variables**.

- **Control** reduces the effects of confounding variables. Three control strategies are **control groups**, **placebos**, and **blinding**.
  - A **control group** is a baseline group that receives no treatment or a neutral treatment.
  - A neutral treatment that has no "real" effect on the dependent variable is called a **placebo**, and a participant's positive response to a placebo is called the **placebo effect**.
  - **Blinding** is the practice of not telling participants whether they are receiving a placebo. **Double-blinding** is the practice of not telling both the participants and the researchers which group receiving a treatment or a placebo.
- **Randomization** ensures that this estimate is statistically valid.

  - With random assignment, we can be fairly confident that any differences we observe in the response of treatment groups is due to the explanatory variable.
- **Replication** reduces variability in experimental results and increases their significance.
  Although randomization helps to insure that treatment groups are as similar as possible, the results of a single experiment, applied to a small number of objects or subjects, should not be accepted without question.
  Any good experiment should be reproducible, and in particular, replication should yield similar results.

Example 1.6. In the study of the relation between a type fertilizer and tomato size, the amount of sunshine will be a confounding variable. It contributes to the growth of tomato.

Exercise 1.8. *A company tested their new golf ball by having 20 professional golfers each hit 100 shots with the company's new ball and 100 shots with the golfer's current ball (in a random order). The labels were removed, so the golfers didn't know which balls were which. The golfers, on average, hit their shots significantly farther with the new ball.*
The company cites this study in an advertisement claiming that this new ball will help all golfers hit farther shots.

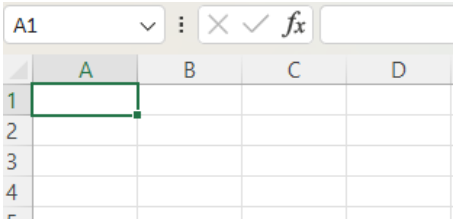
Is the company's claim appropriate? Why?

Exercise 1.9. *Over the years it has been said that coffee is bad for you. When looking at the studies that have shown that coffee is linked to poor health, you will see that people who tend to drink coffee don't sleep much, tend to smoke, don't eat healthy, and tend to not exercise.*

Can you say that the coffee is the reason for the poor health or are there any confounding variables?

1.9 Lab 1 - Introduction to Excel

Exercise 1.10. In the follow figure, highlight the cell C3, the array A1:B2, and the icon for inserting function.



Exercise 1.11. Describe how to use the Excel autofill function to generate a sequential array.

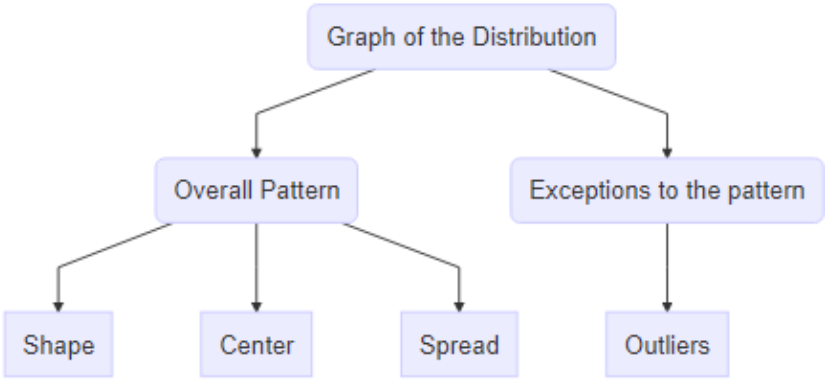
Exercise 1.12. Write down two Excel functions that can be used to generate random numbers and describe the difference.

Exercise 1.13. Describe approach on how to use the paste special option to convert a row array into a column array.

2 Summarizing Data Graphically

2.1 Distribution of Quantitative Data

- In data analysis, one goal is to describe **patterns** (known as the **distribution**) of the variable in the data set and create a useful summary about the set.
- To describe patterns in data, we use descriptions of **shape**, **center**, and **spread**. We also describe exceptions to the pattern. We call these exceptions **outliers**.



2.2 Dot Plots

- A **dot plot** includes all values from the data set, with one dot for each occurrence of an observed value from the set.

Example 2.1. The data set contains 15 petal lengths of iris flower. Create a dot plot to describe the distribution of petal lengths.

1.4, 1.4, 1.3, 1.5, 1.4, 1.7, 1.4, 1.5, 1.4, 1.5, 1.5, 1.6, 1.4, 1.1, 1.2

Exercise 2.1. The data set contains the heights of 20 Black Cherry Trees. Create a dot plot to describe the distribution of the heights.

64, 69, 71, 72, 74, 74, 75, 76, 76, 77, 78, 80, 80, 80, 80, 81, 82, 85, 86, 87

2.3 Histograms

- A **histogram** divides values of a variable into *equal-sized* intervals called **bins** (classes in some books) and uses a rectangular bar to show the **frequency (count)** of observations in each interval.
- A **frequency distribution** is a table which contains bins, frequencies and/or **relative frequencies** which are proportions (percentage) defined by the formula

$$\text{Relative frequency} = \frac{\text{Class frequency}}{\text{Sample size}}.$$

- Each bin has a **lower bin limit**, which is the left endpoint of the interval, and an **upper bin limit**, which is the right endpoint of the interval.
- The **bin width** is the distance between the lower (or upper) bin limits of two consecutive bins.
- The difference between the maximum and the minimum data entries is called the **range**.
- The **midpoint** of a bin is the half of the sum of the lower and upper limits of the bin.

Example 2.2. The following data set show the mpg (mile per gallon) of 30 cars. Construct a frequency table and frequency histogram for the data set using 7 bins.

21, 21, 22.8, 21.4, 18.7, 18.1, 14.3, 24.4, 22.8, 19.2, 17.8, 16.4, 17.3, 15.2, 10.4, 10.4, 14.7, 32.4, 30.4, 33.9, 21.5, 15.5, 15.2, 13.3, 19.2, 27.3, 26, 30.4, 15.8, 19.7

Remark.

- Avoid histograms with large bin widths and small bin widths. See Histogram 2 of 4 in Concepts in Statistics for an interactive demonstration
- When bin width is no given, we may first determine the number of bins. If the number of bins is k, then we choose a number with the same or one more decimal place that is greater than $\frac{\text{range}}{k}$, but no more than $\frac{\text{range}}{k-1}$ as the bin width. This is to avoid that the last bin limit is too much bigger than the max. To determine the number of bins, there are some "rules of thumb". For example, the Rice rule takes the bin number $k = \lceil 2n^{1/3} \rceil$, where $\lceil 2n^{1/3} \rceil$ is the roundup of $2n^{1/3}$.
  See the Statistic How To page for more discussion on choosing bin width.
- The convenient starting point should not be too much smaller than the min. The starting points together with the bin width affects the shape of the histogram. It'd be better to experiment with different choices of the starting point and the bin width.
- The area of a bar represents the relative frequency for the bin. There should be *no space* between any two bars.

Exercise 2.2. The following data set show the petal length of 20 irises. Construct a frequency table and frequency histogram for the data set using 6 bins.

1.4, 5.4, 1.2, 4.5, 6.1, 1.5, 4.7, 1.4, 5.6, 5.2, 1.3, 6.3, 5.1, 5.6, 5, 6.7, 1.4, 1.6, 1.5, 1.5

2.4 Common Descriptions of Shape Distribution

- **Right skewed** (or reverse J-shaped): A right-skewed distribution has a lot of data at lower variable values. (Example: the histogram example.)
- **Left skewed** (or J-shaped): A left skewed distribution has a lot of data at higher variable values with smaller amounts of data at lower variable values.
- **Symmetric with a central peak (or bell-shaped)**: A central peak with a tail in both directions. A bell-shaped distribution has a lot of data in the center with smaller amounts of data tapering off in each direction. (Example: the petal length example.)
- **Uniform**: A rectangular shape, the same amount of data for each variable value.

Exercise 2.3. Statistics are used to compare and sometimes identify authors. The following lists shows a simple random sample that compares the letter counts for three authors.

Terry: 7, 9, 3, 3, 3, 4, 1, 3, 2, 2

Davis: 3, 4, 4, 4, 1, 4, 5, 2, 3, 1

Maris: 2, 3, 4, 4, 4, 6, 6, 6, 8, 3

Create a dot plot for each sample and describe the shape of the distribution of each sample.

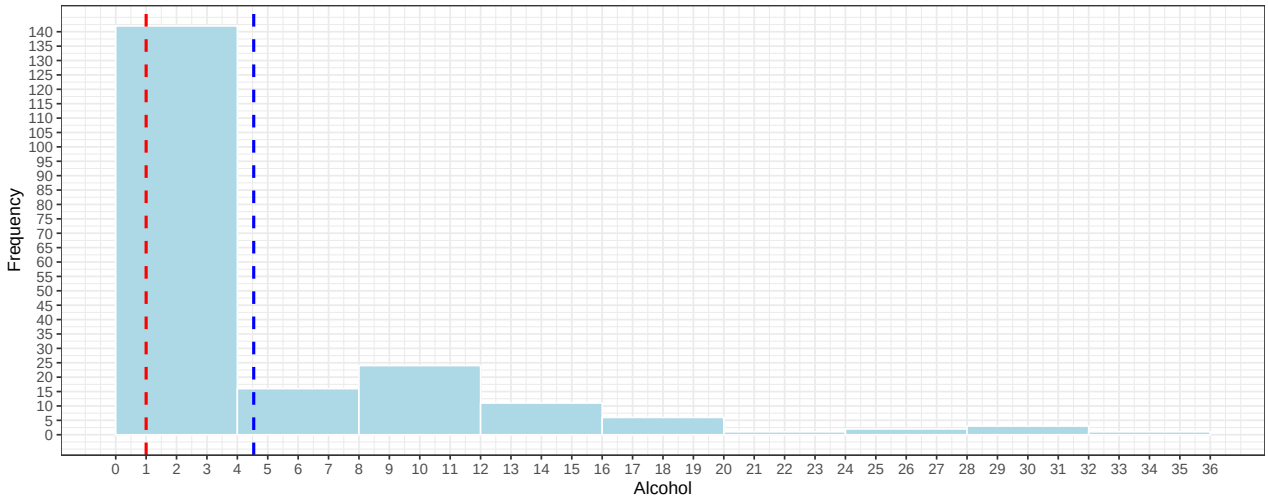
2.5 Measure of Centers

- **Mean**: The mean is the average, this is the quotient of the total sum by the total number.
- **Median**: The median is a value that separate the data into the lower half and the upper half. To calculate the median, sort the data first. If the number of data values is odd, the median is the middle value. Otherwise, the median is the mean of the middle two values.
- **Mode**: The mode is the value that has the most occurrence in the data set.
- Use the *mean* as a measure of center only for distributions that are *reasonably symmetric* with a central peak. When outliers are present, the mean is not a good choice.
- Use the *median* as a measure of center for all *other cases*.
- We need to use a graph to determine the shape of the distribution. So graph the data first.

Exercise 2.4. A student survey was conducted at a major university. The following histogram shows distribution of alcoholic beverages consumed in a typical week.

1. What is the typical number of drinks a student has during a week?

2. Do the data suggest that drinking is a problem in this university?



2.6 Pie Charts

- A **pie chart** is a pie with sectors represents categories and the area of each sector is proportional to the frequency of each category.
  - The **frequency** of a category is the number of occurrences of elements in the category.
  - The proportion of a frequency to the size of the population or the sample is also called the **relative frequency**.

Example 2.3. The counts of majors of 100 students in a sample are shown in the table. Use a pie chart to organize the data.

Grade	Frequency (Counts)
Art	30
Engineering	50
Science	20

Exercise 2.5. The following data table summarize passengers on Titanic. Using a pie chart to describe the data table.

Class	Passengers
1st	325
2nd	285
3rd	706
Crew	885

2.7 Lab 2: Summarizing Data Graphically

2.7.1 Create Frequency Tables

In Excel, to create a frequency table for a data array, we need a bin array which is used to split the date set into smaller intervals. The values in a bin array in Excel are (upper) boundaries of intervals. With a data array and a bin array, we can use the Excel function FREQUENCY (data_array, bins_array) to create a frequency table.

Suppose the data set is in column A and the bin array is in column B. Here is how to create a frequency table using the function FREQUENCY (data_array, bins_array):

(1) In column C, right to the smallest value of the bin array enter =FREQUENCY(

(2) select the data values

(3) in the formula bar, enter the symbol comma ,

(4) select the bin array
(5) in the formula bar, enter).
Hit the Enter, you will get a frequency table.

Remark. (1) In this formula, the values in a bin array should be first $k-1$ upper class limits (or the last $k-1$ lower class limits), where k is the number of bins. In Excel, if the bin array consists of 30, 40, and 50, then the bins will be $(-\infty, 30]$, $(30, 40]$, $(40, 50]$, $(50, \infty)$.

(2) In older version of Excel, you may have to highlight cells for frequencies first, enter the FREQUENCY function secondly, and then hit Ctrl+Shift+Enter (or Cmd+Shift+Enter on Mac).

Exercise 2.6. Create a frequency table for the following data using the bin width 10.

31, 32, 32, 33, 35, 36, 37, 37, 38, 38, 39, 40, 40, 40, 42, 42, 43, 43, 45, 45, 46, 47, 48, 48, 51, 55, 55, 56, 60, 60, 61, 66

2.7.2 Creating Charts in Excel

Excel has many built-in chart functions. To create a charts,
(1) Select the data array/table
(2) Under the Insert tab, click on an appropriate chart in the Charts command set.
The appearance of chart can be changed after being created.

Exercise 2.7. The counts of majors of 50 students in a sample are shown in the table. Use a pie chart to organize the data.

Grade	Count
Art	10
Engineering	25
Mathematics	15

2.7.3 Create a Histogram in Excel

(1) Select the data

(2) On the Insert tab, in the Charts group, from the Insert Statistic Chart dropdown list, select Histogram:

Note: The histogram contains a special first bin which always contains the smallest number. This is different from many textbooks.

To **format the histogram chart** is similar to format a Pie chart. For example, you can change bin width from Format Axis.

(1) Right-click on the horizontal axis and choose Format Axis in the popup menu:

(2) In the Format Axis pane, on the Axis Options tab, you may try different options for bins.

Remark.
- Excel using a different convention to create histogram. The first bin is a closed interval and other bins are left open and right closed intervals.
- Select the **Overflow bin** checkbox and type the number, all values above this number will be added to the last bin.
- Select the **Underflow bin** checkbox and type the number, all values below and equal to this number will be added to the first bin.
- Histograms show the shape and the spread of quantitative data. For categorical data, discrete by its definition, bar charts are usually used to represent category frequencies.

2.7.4 Create Histogram Charts in Excel using the Analysis ToolPak

Suppose your data set is in Column A in Excel.

- In the cell B1, put the *first lower bin limit*, which is a number slightly less than the minimum but has more decimal places than the data set.
- Create upper bin limits in column C.
- In Data menu, look for the Data Analysis ToolPak (if not, go to File → Options → Add-ins → Manage Excel Add-ins, check Analysis ToolPak). In the popup windows, find Histogram.
- In the input range, select your data set. In the bin range, select upper bins.
- Check Chart Output and hit OK. You will see the frequency table and histogram in Sheet 2.
- Change the gap between bars. Right click a bar and choose Format Data Series... and change the Gap Width to 2% or 1%.

2.7.5 How to Create a Dotplot in Excel

- If you have a raw data set, follow the same procedure a creating a histogram but with a bin width equal the same accuracy of the data. For example, if you data set consists of integers, then choose 1 as the bin-width.
- Change the format of bars in the histogram.
  - Right click a bar and select Format Data Series....
  - Find Fill & Line and select both Picture or texture fill and Stack and Scale with.
  - Click the button Online... and input *dot* in search bing and hit enter.
  - Select a picture you like and you will get a dot-plot.

Exercise 2.8. Use Excel to complete the following tasks:

(1) Create a random sample of 30 two-digit integers.

(2) Create a histogram with 6 bins for the sample.

(3) Describe the shape of the distribution of the sample of 30 two-digit integers.

3 Measure of Center and Variation

3.1 Median, Quartiles, Interquartile Range and Outliers

- The three **quartiles**, Q_1, Q_2, and Q_3 are numbers in an ordered data set that divide the data set into four equal parts. The second quartile is known as the **median**.
- **Interquartile Range (IQR for short)** is the measure of variation when using the median to measure center. It is defined as the difference of the third and the first quartiles: IQR = $Q_3 - Q_1$.
- When the center and the spread are measured by the median and the IQR, a value in the data is considered an **outlier** if the value is
  - less than the lower fence $\text{fence}_{lower} = Q_1 - 1.5 \cdot \text{IQR}$ or
  - greater than the upper fence $\text{fence}_{upper} = Q_3 + 1.5 \cdot \text{IQR}$.

  Note: An outlier in this definition is also called a **mild outlier**. An outlier that is less than the extreme lower fence extreme $\text{fence}_{lower} = Q_1 - 3 \cdot \text{IQR}$ or greater than the extreme upper fence extreme $\text{fence}_{upper} = Q_3 + 3 \cdot \text{IQR}$ is also called **extreme outlier**.
- The minimum, Q_1, Q_2, Q_3 and maximum are known as the "**five-number summary**" of the data set.
- The difference of maximum and minimum is called the **range**.

Example 3.1. Find the median, quartiles, IQR and outliers (if they exist) of the sample height of 15 trees.

70, 65, 63, 72, 81, 83, 66, 75, 80, 75, 79, 76, 76, 69, 75

Exercise 3.1. Find the five-number summary, the IQR and the Range for the following set of data.

2, 7, 7, 7, 10, 11, 14, 17, 18, 20

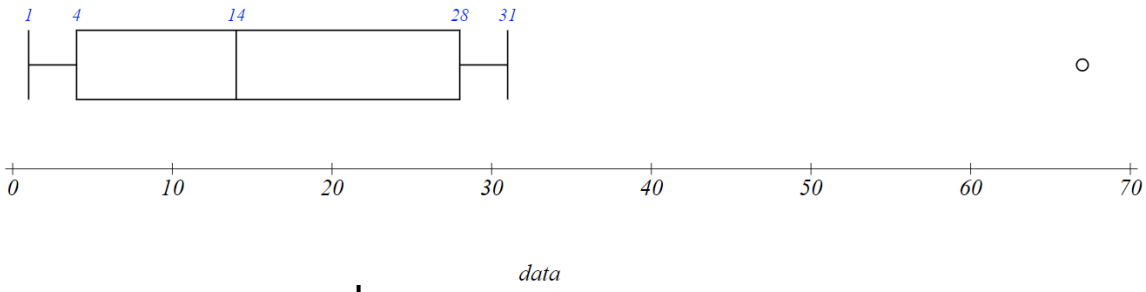
3.2 Box Plot

- A **box plot** shows a "five-number summary" of the data set. It contains a box, two whiskers and dots (for outliers).
- To create the boxplot for a distribution,
  - Draw a box from Q_1 to Q_3.
  - Draw a vertical line in the box at the median.
  - Extend a tail from Q_1 to the smallest value that is not an outlier and from Q_3 to the largest value that is not an outlier.
  - Indicate outliers with a solid dot.

Example 3.2. Create the boxplot for the ages of 32 best actor oscar winners (1970–2001).

31, 32, 32, 33, 35, 36, 37, 37, 38, 38, 39, 40, 40, 40, 42, 42, 43, 43, 45, 45, 46, 47, 48, 48, 51, 55, 55, 56, 60, 60, 61, 76

Exercise 3.2. Based on the boxplot below, identify the 5 number summary (minimum, lower quartile (Q1), median (Q2), upper quartile (Q3), maximum)



3.3 Mean and Standard Deviation

- Sigma notation: in math, we denote the sum of values x_1, x_2, ..., x_n of a variable x by $\sum_{i=1}^{n} x_i$ or simply by $\sum x$.
- The **population mean** is $\mu = \frac{\sum x}{N}$, where N is the **population size**, i.e the number of elements in the population.
  The notation μ reads as mu.
- The **sample mean** is $\bar{x} = \frac{\sum x}{n}$, where n is the **sample size**.
  The notation $\bar{x}$ reads as x–bar.

Example 3.3. Find the mean city mpg for a sample of 10 cars.
18, 21, 20, 21, 16, 18, 18, 18, 16, 20

3.4 Weighted Mean

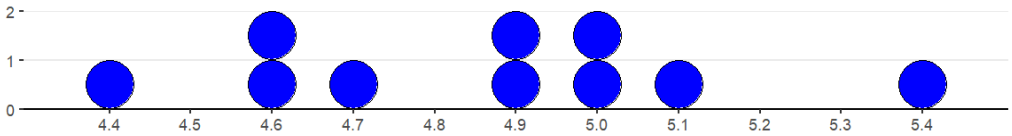
- The weighted mean of a set of numbers $\{x_1, \ldots, x_n\}$ with weights w_1, w_2, ..., w_n is defined as
$$\frac{\sum w_i x_i}{\sum w_i}.$$
- The mean of a frequency table is weighted mean $\overline{x} = \frac{\sum fx}{n}$, where x is an element with frequency f and n is the sample size.

Example 3.4. In a course, the overall grade is determined in the following way: the homework average counts for 10%, the quiz average counts for 10%, the test average counts 50% , and the final exam counts for 30%. What's the overall grade of the student who earned 92 on homework, 95 on quizzes, 90 on tests and 93 on the final.

Exercise 3.3. Find the average petal width for a sample of 10 iris followers.

0.2, 2.1, 0.2, 1.7, 2.3, 0.3, 1.2, 0.2, 1.8, 2.3

Exercise 3.4. Find the mean and median from the dot plot of sepal length for a sample of 10 iris flowers.



Exercise 3.5. In a student's chemistry class, the final grade is based on six categories.

The categories, grades, and weights are shown in this table.

Category	Grade	Weight %
Test 1	46	20
Test2	61	20
Test 3	45	15
Homework	72	9
Semester Project	53	8
Final Exam	77	28

Compute a weighted average to determine the student's overall final grade in the course. Record the overall final grade below as a percentage. Round accurately to two decimal places.

3.5 Measure of Variation about Population Mean

- The **deviation** of an entry x in a population data set is the difference $x - \mu$, where μ is the mean of the population.
- The **population variance** of a population of N entries is defined as

$$\text{VAR.P} = \sigma^2 = \frac{\sum (x-\mu)^2}{N}.$$

- The **population standard deviation** is

$$\text{STDEV.P} = \sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}.$$

3.6 Measure of Variation about Sample Mean

- The **deviation** of an entry x in a sample data set is the difference $x - \bar{x}$, where $\bar{x}$ is the mean of the sample.
- The **sample variance** and **sample standard deviation** are defined similarly

$$\text{VAR.S} = s^2 = \frac{\sum (x-\bar{x})^2}{n-1}, \qquad \text{STDEV.S} = s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}},$$

  where n is the sample size.
- **Rounding rule:** for mean, variance and standard deviation, we keep at least one more digit than the accuracy of the data set.

Note: To measure the spread, one may also use the **mean absolute deviation**

$$MAD = \frac{\sum |x-\bar{x}|}{n}.$$

However, the standard deviation has better properties in applications.

Example 3.5. Find the mean and standard deviation ages of a sample of 32 best actor oscar winners (1970–2001).

31, 32, 32, 33, 35, 36, 37, 37, 38, 38, 39, 40, 40, 40, 42, 42, 43, 43, 45, 45, 46, 47, 48, 48, 51, 55, 55, 56, 60, 60, 61, 76

Exercise 3.6. A *sample* of GPAs from ten students random chosen from a college are recorded as follows.

1.90, 3.00, 2.53, 3.71, 2.12, 1.76, 2.71, 1.39, 4.00, 3.33

Find the standard deviation of this sample.

3.7 Mean and Standard Deviation under Linear Transformation

- When we increase values in a data set by a fixed number c, the standard deviation of a data set won't change. However, the mean increases by c too.
- When we multiple values in a data set by a factor k, the mean and the standard deviation both scale by the factor k.

https://tinyurl.com/6vrp7ze8

3.8 Effect of Changes of Data on Statistical Measures

https://tinyurl.com/2n3r7xj2

Exercise 3.7. A sample of the highest temperature of 10 days has a standard deviation $5°\mathrm{C}$ in Celsius.

(1) If we want to know the standard deviation in Fahrenheit, do we need to recalculate using the sample?

(2) What is the standard deviation in Fahrenheit.

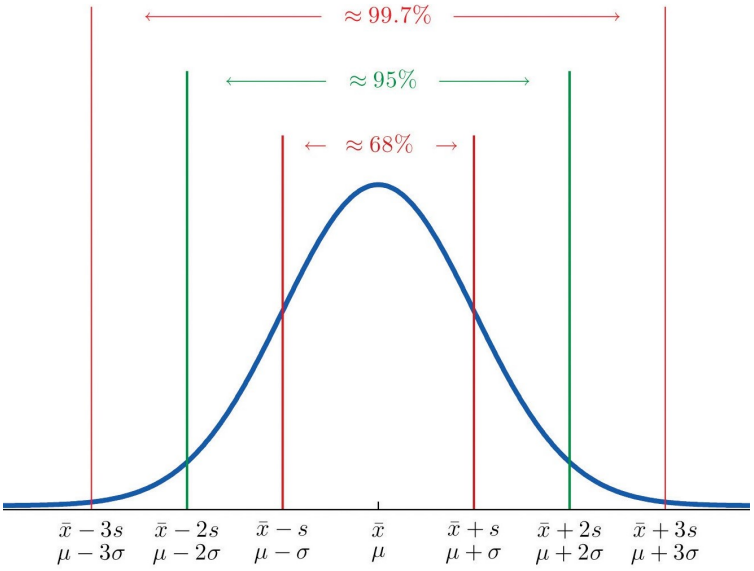
3.9 The Empirical Rule

If a data set has an **approximately bell-shaped** distribution, then

(1) approximately 68% of the data lie within one standard deviation of the mean.

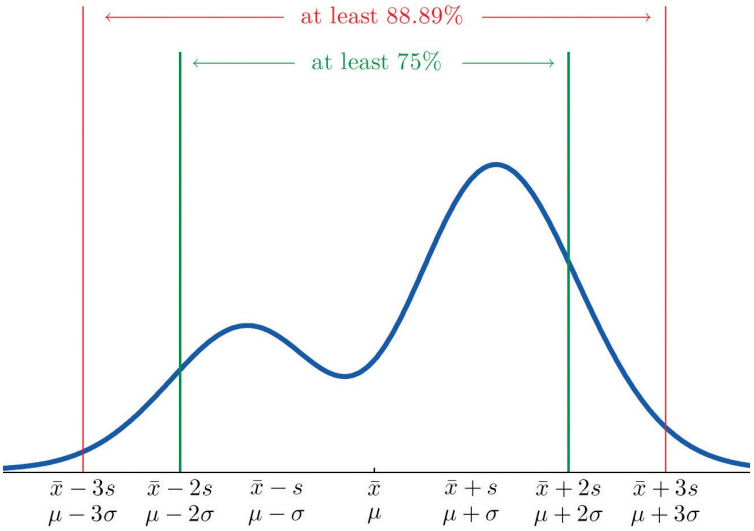
(2) approximately 95% of the data lie within two standard deviations of the mean.

(3) approximately 99.7% of the data lies within three standard deviations of the mean.



3.10 Chebyshev's Theorem

For any numerical data set, at least $1-1/k^2$ of the data lie within k standard deviations of the mean, where k is any positive whole number that is at least 2.



Example 3.6. A population data set with a bell-shaped distribution has mean $\mu = 6$ and standard deviation $\sigma = 2$. Find the approximate proportion of observations in the data set that lie:

(1) between 4 and 8;

(2) below 4.

Example 3.7. A sample data set has mean $\overline{x} = 6$ and standard deviation $s = 2$. Find the minimum proportion of observations in the data set that must lie between 2 and 10.

Exercise 3.8. The maintenance department at the main campus of a large state university receives daily requests to replace fluorecent lightbulbs. The distribution of the number of daily requests is bell-shaped and has a mean of 60 and a standard deviation of 9. Using the 68-95-99.7 rule, what is the approximate percentage of lightbulb replacement requests numbering between 60 and 78?

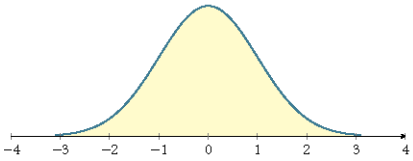
Exercise 3.9. A sample data set has mean $\overline{x} = 10$ and standard deviation $s = 3$. Find the minimum proportion of observations in the data set that must lie between 1 and 19.

3.11 More Practice

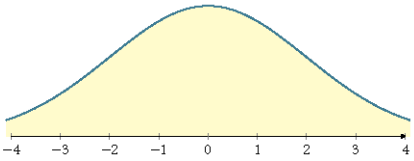
Exercise 3.10. A teacher decide to curve the final exam by adding 10 points for each student. Which of the following statistic will

NOT change: (1) median, (2) mean, (3) interquartile range, (4) standard deviation? **Please explain your conclusion.**

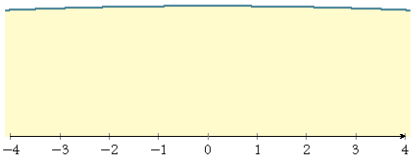
Exercise 3.11. Which distribution of data has the SMALLEST standard deviation? Please explain your conclusion.



A.



B.



C.

3.12 Lab 3: Centers and Variations

3.12.1 Mean, Median, Quartiles and Standard Deviation

- To find the median, you may use the function `MEDIAN()`.
- To find quartiles, you may use the function `QUARTILE.EXC()`. **Note:** this function calculates first and third quartiles with 25% and 75% weights. The results may be different from the results calculated by hand discussed in this course.

- To find the mean, you may use the function AVERAGE().
- To find the **population** standard deviation, you may use the function STDEV.P().
- To find the **sample** standard deviation, you may use the function STDEV.S().

3.12.2 How to Create a Boxplot in Excel

- Select your data—either a single data series, or multiple data series.
- Click Insert → Insert Statistic Chart → Box and Whisker to create a boxplot.

For more information, see Create a box and whisker chart in Excel 365

Exercise 3.12. Consider the following sample that consists of speeds of 20 cars.

19, 4, 17, 22, 23, 8, 20, 19, 10, 10, 13, 13, 15, 12, 20, 14, 9, 20, 12, 11

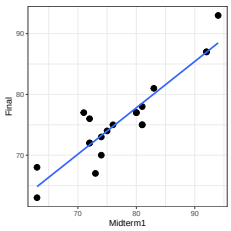
(1) Use Excel to find the mean, median, quartiles and standard deviation of the sample.

(2) Create a box-plot for the sample.

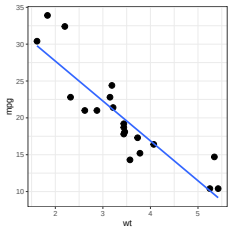
4 Linear Relationship

4.1 Scatterplots

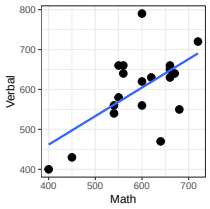
- Correlation refers to a relationship between two quantitative variables:
  - the independent (or explanatory) variable, usually denoted by x.
  - the dependent (or response) variable, usually denoted by y.
- To describe the relationship between two quantitative variables, statisticians use a scatterplot.
- In a scatterplot, we describe the overall pattern with descriptions of direction, form, and strength.
- **Positive relationship**: the response variable (y) increases when the explanatory variable (x) increases.
- **Negative relationship**: the response variable (y) decreases when the explanatory variable (x) increases.
- **Forms of relationship:**



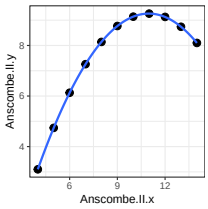
Positive Relation



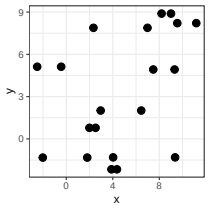
Negative Relation



Linear form

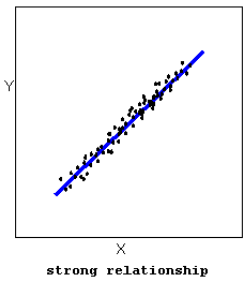


Curvilinear form

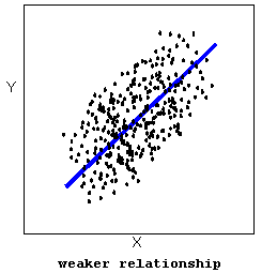


No obvious relationship

- The strength of the relationship is a description of how closely the data follow the form of the relationship.

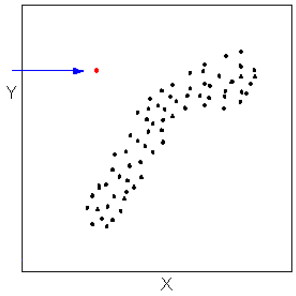


strong relationship

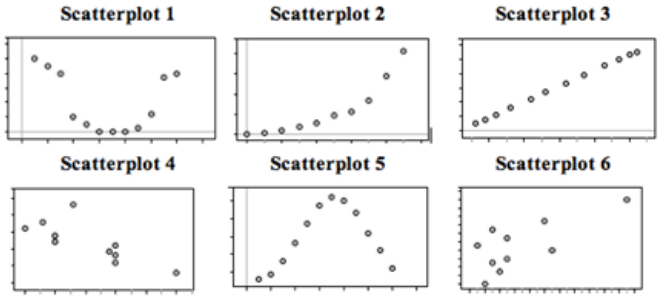


weaker relationship

- Outliers are points that deviate from the pattern of the relationship.



Exercise 4.1. Match scatterplots with the descriptions of relations between given variables.



A : X = month (January = 1), Y = rainfall (inches) in Napa, CA in 2010 (Note: Napa has rain in the winter months and months with little to no rainfall in summer.)

B : X = month (January = 1), Y = average temperature in Boston MA in 2010 (Note: Boston has cold winters and hot summers.)

C : X = year (in five-year increments from 1970), Y = Medicare costs (in $) (Note: the yearly increase in Medicare costs has gotten bigger and bigger over time.)

D : X = average temperature in Boston MA (°F), Y = average temperature in Boston MA (°C) each month in 2010

E : X = chest girth (cm), Y = shoulder girth (cm) for a sample of men

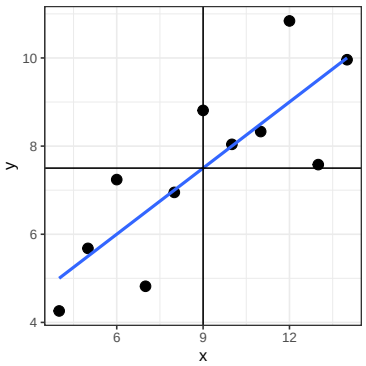
F : X = engine displacement (liters), Y = city miles per gallon for a sample of cars (Note: engine displacement is roughly a measure of engine size. Large engines use more gas.)

4.2 The Correlation Coefficient

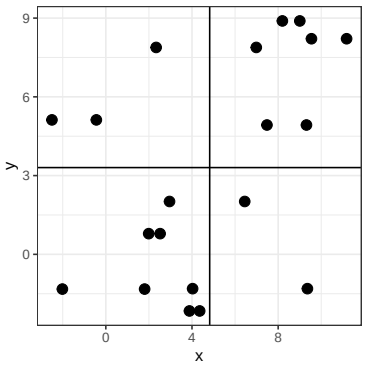
- The correlation coefficient r is a numeric measure that measures the strength and direction of a linear relationship between two quantitative variables.

$$r = \frac{\sum \left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)}{n-1},$$

  where n is the sample size, x is a data value for the explanatory variable, $\bar{x}$ is the mean of the x-values, s_x is the standard deviation of the x-values, and similarly, for the notations involving y.
- The expression $z = \frac{x-\bar{x}}{s_x}$ is known as the standardized variable (or z-score) which
  - doesn't depend on the unit of the variable x,
  - has mean 0 and standard deviation 1.
- In Excel, the correlation coefficient can be calculated using the function CORREL().
- **Rounding Rule:** Round to the nearest thousandth for r, m and b.
- Geometric explanation of the definition of r.



$r = 0.816$



$r = 0.420$

- The correlation coefficient r is between -1 and 1.
- The closer the absolute value $|r|$ is to 1, the stronger the linear relationship is.
- The correlation is symmetric in x and y, that is CORREL(x, y)=CORREL(y, x).

- The correlation does not change when the units of measurement of either one of the variables change. In other words, if we change the units of measurement of the explanatory variable and/or the response variable, it has no effect on the correlation (r).
- The correlation by itself is not enough to determine whether a relationship is linear. It's important to graph data set before analyzing it.
  https://en.wikipedia.org/wiki/Anscombe%27s_quartet
- The correlation is heavily influenced by outliers. Try the simulation in Linear Relation (4 of 4) in Concepts in Statistics
- The reason that $|r|$ is less than 1 is from the Cauchy-Schwarz inequality: $(\sum XY)^2 \leq \sum X^2 \sum Y^2$.

Exercise 4.2. Open the linked website and try to guess the correlation coefficient.
https://istats.shinyapps.io/guesscorr/

Example 4.1. Use the data on Midterm 1 and Final from a sample of 10 students.
- Draw a scatter plot for the data table.
- Is it appropriate to study the relationship using a linear model.
- Find and interpret the correlation coefficient.

Midterm1	Final
72	72
93	88
81	82
82	82
94	88
80	77
73	78
71	77
81	76
81	76
63	68

Exercise 4.3. Use the data shown below to answer the following questions.

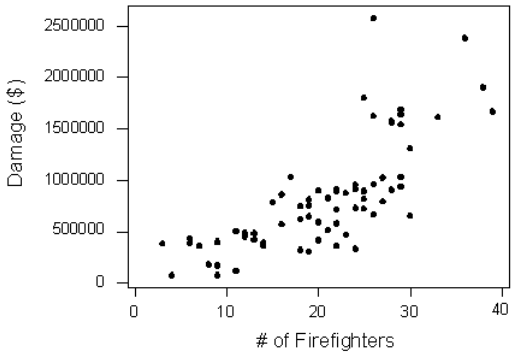
- Draw a scatter plot for the data table.
- Is it appropriate to study the relationship using a linear model.
- Find and interpret the correlation coefficient.

x	y
4	14.86
5	15.65
6	17.94
7	18.63
8	17.12
9	21.11
10	19.7
11	21.99

4.3 Correlation v.s. Causation

- Correlation is described by data from observational study. Observational studies cannot prove cause and effect which requires controlled study and rigorous inferences.
- Correlation may be used to make a prediction which is probabilistic.
- In a linear relationship, an r-value that is close to 1 or -1 is insufficient to claim that the explanatory variable causes changes in the response variable. The correct interpretation is that there is a statistical relationship between the variables.
- A **lurking variable** is a variable that is not measured in the study, but affects the interpretation of the relationship between the explanatory and response variables.

Example 4.2. The scatterplot below shows the relationship between the number of firefighters sent to fires (x) and the amount of damage caused by fires (y) in a certain city.



Can we conclude that the increase in firefighters causes the increase in damage?

Exercise 4.4. Over a period of a few years, the population of Denver increased. It was observed that during this period the correlation between the number of people attending church and the number of people receiving traffic tickets was $r = 0.92$. Does going to church cause people to get traffic tickets? Is there a lurking variable that might cause both variables to increase?

4.4 The Regression Line

- The line that best summarizes a linear relationship is **the least squares regression line**. The regression line is the line with the smallest sum of squares of the errors (**SSE**).
- We use the least-squares regression line to predict the value $\hat{y}$ for a value of the explanatory variable x.
- The regression line is unique and passes though $(\overline{x}, \overline{y})$. The equation is given by
$$\hat{y} = m(x - \overline{x}) + \overline{y} = mx + b,$$
where the slope is
$$m = \frac{\sum(x - \overline{x})(y - \overline{y})}{\sum(x - \overline{x})^2} = r\frac{s_y}{s_x}$$
and the y-intercept is $b = \overline{y} - m\overline{x}$.

- The **error of a prediction** is

$$\text{Error} = \text{Observed} - \text{Predicted} = y - \hat{y}.$$

- A prediction beyond the range of the data is called **extrapolation**.

Example 4.3. The following sample is taken from data about the Old Faithful geyser.

(1) Study the linear relationship. Is it positive? What's the strength? What's the direction?

(2) Find the regression line, and the predicated value and the error if the eruption time is 1.8 minutes.

eruptions	**waiting**
3.917	84
1.75	62
4.200	78
4.80	84
1.750	47
1.60	52
4.700	83
4.25	79
2.167	52
1.80	51

Exercise 4.5. Research was conducted on the amount of training for 5K and the time a contestant took to run the race. The researcher recorded the number of miles during training (a 1-month period) and the time to complete the 5K. The results are below.

Miles Trained	20	69	102	29	46	68
Time(Minutes)	37.26	38.3	39.96	29.95	26.56	43.78

(1) Find the correlation coefficient.

(2) Find the equation of regression line.

(3) Predict the time in the 5K if someone trained 29 miles.

(4) Find the residual (the prediction error) for 102 miles trained.

4.5 Assessing the Fit of a Regression Line

- The prediction error is also called a **residual**. Another way to express the previous equation for error is
$$y = \hat{y} + \text{residual}.$$
- **Residual plots** are used to determine if a linear model is appropriate.
A random pattern (or no obvious pattern) indicates a good fit of a linear model. See Assessing the Fit of a Line (2 of 4) in Concepts in Statistics for examples.
- A "typical" error used to measure the fit of the regression is the **residual standard errors** (or **standard error of the regression**), calculated by the Excel function STEYX(), is
$$s_e = \sqrt{\frac{SSE}{n-2}},$$
where $SSE = \sum(y - \hat{y})^2$ is the sum of square errors.
- The smaller s_e is, the more accurate the prediction is.
- The fit of a regression line can also be measured by the proportion of the variation in the response variable that is explained by the least-squares regression line. This proportion is known as the **coefficient of determination**.

– The **total variance** is $SSD = \sum(y - \bar{y})^2$
– The **explained variance** is $SSR = \sum(\hat{y} - \bar{y})^2$.
– The **coefficient of determination** is

$$r^2 = \frac{SSR}{SSD} = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2}.$$

Remark.
- The r in the coefficient of determination is the correlation coefficient. Equivalently, $r = \pm\sqrt{r^2}$.
- The smaller the standard error, the larger the coefficient of determination:

$$r^2 = 1 - \frac{SSE}{SSD} = 1 - \frac{(n-2)s_e^2}{SSD}.$$

- $n - 2$ is the degrees of freedom. We lose two degrees of freedom because we estimate the slope and the y-intercept.
- In a linear regression model $Y = \beta_0 + \beta_1 X + \epsilon$, even we have β_0 and β_1 from the population, we still need estimate the standard deviation of error.

Example 4.4. Find the standard error and coefficient of determination for the data of midterm1 and final.

Midterm1	72	93	81	82	94	80	73	71	81	81
Final	72	88	82	82	88	77	78	77	76	76

Exercise 4.6. A researcher measures the wrist circumference and height of a random sample of individuals. The data are displayed below.

Wrist Size (in)	5.5	5.6	5.8	5.9	6.1	6.3	6.4	6.5	6.6	6.8
Height (in)	60.4	59.7	66.3	63.5	60.4	66.9	65.6	70.9	59.7	64.9

(1) Find the equation of the best-fit line $y = mx + b$.

(2) Find the correlation coefficient.

(3) Predict the height of a person with a wrist circumference of 6 inches using the best-fit line.

(4) Calculate the residual for the point (6.5,70.9).

(5) When predicting heights, what is a "typical" error of this linear model?

(6) What proportion of variability in heights can be explained by this linear model? (Write your answer in decimal.)

(7) Find the correlation coefficient if the heights are measured in feet.

4.6 Lab 4: Linear Regressions

- To create a scatter plot, first select the data sets, and then look for Insert Scatter(X, Y) in the menu Insert → Charts.
- The correlation coefficient r can be calculated by the Excel function correl().
- The slope of a linear regression can be calculated by the Excel function SLOPE().
- The y-intercept of a linear regression can be calculated by the Excel function INTERCEPT().
- The coefficient of determination can be calculated by first finding r, then applying the formula r^2.
- The standard error of the regression (residual standard error) can be calculated by the Excel function STEYX().

Exercise 4.7. A researcher measures the wrist circumference and height of a random sample of individuals. The data and the scatterplot are displayed below.

Wrist Size (in)	5.7	5.9	6	6.2	6.3	6.5	6.7	7.1	7.3	8	8.2	8.4
Height (in)	62.8	68.3	68.7	59.1	61.2	67.6	69.7	70.6	75.2	80.8	78.2	80.9

(1) Create a scatterplot for the data.

(2) Find the equation of the best-fit line $y = mx + b$.

(3) Predict the height of a person with a wrist circumference of 6 inches using the best-fit line.

(4) Calculate the residual for the point (5.9,68.3).

(5) When predicting heights, what is a "typical" error of this linear model?

(6) What proportion of variability in heights can be explained by this linear model?

(7) Find the correlation coefficient if the heights are measured in feet.

5 Two-way Tables

5.1 Two-way Frequency Tables

- As we organize and analyze data from two categorical variables, we make use of two-way tables.
- Information in a **two-way frequency table**:
  - Values of the two variables are displayed in the left column and the top row.
  - The body of table consists of frequency counts associated to pairs of values of the two variables.
  - The right column and the bottom row, which are called margins of the table, consists of row totals and column totals respectively.
- A number in a margin are called **marginal frequency**.
- A numbers in the body of the table is called **joint frequency**.

Example 5.1. The following table summarize responses of a random sample of 1,200 U.S. college students as part of a larger survey.

	About Right	Overweight	Underweight	Row Totals
Female	560	163	37	760
Male	295	72	73	440
Column Totals	855	235	110	1,200

5.2 Two-Way Relative Frequency Tables and Probability

- A **two-way relative frequency table** is obtained from a two-way frequency table by converting frequencies in a two-way table to relative frequencies.
- **Marginal probability**

$$P(X) = \frac{\text{Marginal frequency in } X}{\text{Total}}$$

- **Conditional probability**

$$P(X|Y) = \frac{\text{Joint frequency}}{\text{Marginal Frequency in } Y}$$

$$P(Y|X) = \frac{\text{Joint frequency}}{\text{Marginal Frequency in } X}$$

- **Joint probability**

$$P(X \text{and } Y) = \frac{\text{Joint frequency}}{\text{Total}}$$

- Note that $P(X \text{ and } Y) = P(X) \cdot P(Y|X) = P(Y) \cdot P(X|Y)$.

Example 5.2. The following table shows joint and marginal probabilities of body image and gender.

	About Right	Overweight	Underweight	Row Totals
Female	$\frac{560}{1200} = 46.67\%$	$\frac{163}{1200} = 13.58\%$	$\frac{37}{1200} = 3.08\%$	$\frac{760}{1200} = 63.33\%$
Male	$\frac{295}{1200} = 24.58\%$	$\frac{72}{1200} = 6.00\%$	$\frac{73}{1200} = 6.08\%$	$\frac{440}{1200} = 36.67\%$
Column Totals	$\frac{855}{1200} = 71.25\%$	$\frac{235}{1200} = 19.58\%$	$\frac{110}{1200} = 9.17\%$	$\frac{1200}{1200} = 100.00\%$

Example 5.3. The following table shows probabilities of randomly select male or female who has a certain body image.

	About Right	Overweight	Underweight	Row Totals
Female	$\frac{560}{760} = 73.68\%$	$\frac{163}{760} = 21.45\%$	$\frac{37}{760} = 4.87\%$	$\frac{760}{760} = 100.00\%$
Male	$\frac{295}{440} = 67.05\%$	$\frac{72}{440} = 16.36\%$	$\frac{73}{440} = 16.59\%$	$\frac{440}{440} = 100.00\%$

Example 5.4. The following table summarizes the full-time enrollment at a community college.

	Arts-Sci	Bus-Econ	Info Tech	Health Science	Graphics Design	Culinary Arts	Row Totals
Female	4,660	435	494	421	105	83	6,198
Male	4,334	490	564	223	97	94	5,802
Column Totals	8,994	925	1,058	644	202	177	12,000

(1) What proportion of the total number of students are male students?

(2) If we select a male student at random, what is the probability that he is in the Info Tech program?

(3) If a student is selected at random, what is the probability that the student is a male and in the Info Tech program?

(4) How are those three probabilities related?

Exercise 5.1. This table relates the weights and heights of a group of individuals participating in an observational study.

Weight/Height	Tall	Medium	Short
Obese	18	28	14
Normal	20	51	28
Underweight	12	25	9

(1) Find the total for each row and column

(2) Find the probability that a randomly chosen individual from this group is Short.

(3) Find the probability that a randomly chosen individual from this group is Obese and Short.

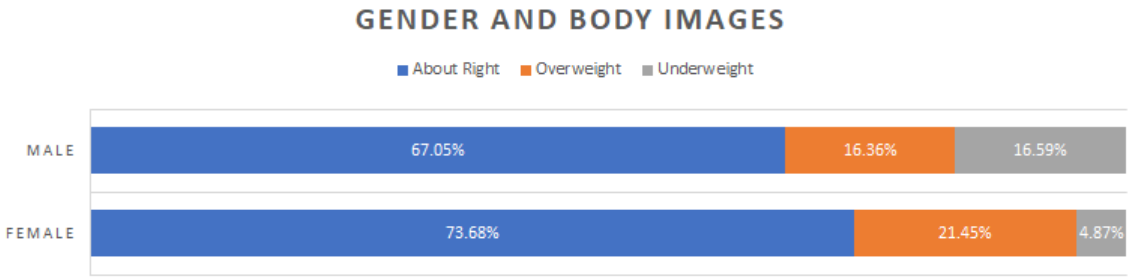
(4) Find the probability that a randomly chosen individual from this group is Underweight given that the individual is Tall.

5.3 Test of (No) Association

- To understand association between categorical variables, we may think conversely. How do we test no association?
- If the conditional probabilities are nearly equal for all categories, there may be no association between the variables. Conversely, if the conditional probabilities are different enough, we are confidence to say there is an association.
- In general, the bigger the differences in the conditional probabilities, the stronger the association between the variables.
- Two variables X and Y are **independent** if $P(X \text{ and } Y) = P(X) \cdot P(Y)$. Equivalently, $P(X|Y) = P(X)$ and $P(Y|X) = P(Y)$.

Example 5.5. Is body image related to gender?

	About Right	Overweight	Underweight	Row Totals
Female	560	163	37	760
Male	295	72	73	440
Column Totals	855	235	110	1,200



5.4 Percentage Reduction of Risk

- When calculating the probability of a negative outcome, we often refer to the probability as a **risk**.
- In general, we are interested in determining how much a new treatment reduces the risk compared to a reference risk
- The **percentage reduction of risk** is

$$\text{percentage reduction of risk} = \frac{\text{new treatment risk} - \text{reference risk}}{\text{reference risk}}.$$

Example 5.6. Researchers in the Physicians' Health Study (1989) designed a randomized double-blind experiment to determine whether aspirin reduces the risk of heart attack. Here are the final results.

	Heart Attack	No Heart Attack	Row Totals
Aspirin	139	10,898	11,037
Placebo	239	10,795	11,034
Column Totals	378	21,693	22,071

Does aspirin lower the risk of having a heart attack?

5.5 Hypothetical Two-way Tables

A **hypothetical two-way table**, also known as a hypothetical 1000 two-way table, is a two-way table constructed from given probability conditions with 1000 or higher as the total frequency. It can be used to answer complex probability questions.

Example 5.7. A pregnant woman often opts to have an ultrasound to predict the gender of her baby. Assume the following facts are known:

- Fact 1: 48% of the babies born are female.
- Fact 2: 90% of girls were correctly identified.
- Fact 3: 75% of boys were correctly identified.

Use the above facts to answer the following questions.

(1) If the examination predicts a girl, how likely the baby will be a girl?

(2) If the examination predicts a boy, how likely the baby will be a boy?

Exercise 5.2. The table below is based on a 1988 study of accident records conducted by the Florida State Department of Highway Safety.

	Nonfatal	Fatal	Row Totals
Seat Belt	412,368	510	412,878
No Seat Belt	162,527	1,601	164,128
Column Totals	574,895	2,111	577,006

Does wearing a seat belt lower the risk of an accident resulting in a fatality?

Exercise 5.3. A large company has instituted a mandatory employee drug screening program. Assume that the drug test used is known to be 99% accurate. That is, if an employee is a drug user, the test will come back positive ("drug detected") 99% of the time. If an employee is a non-drug user, then the test will come back negative ("no drug detected") 99% of the time. Assume that 2% of the employees of the company are drug users.

If an employee's drug test comes back positive, what is the probability that the test is wrong and the employee is in fact a non drug user?

5.6 Create Stacked Bar Chart in Excel

To create a a stacked bar chart of a two-way table

- First select the data table.
- Look for and click `Insert Column or Bar Chart` in the menu `Insert` → `Charts`.

- In the dropdown menu, choose the third option in 2-D Column (`100% Stacked Column`) or the third option 2-D Bar (`100% Stacked Bar`).
- To switch row/column, in the output graph, right click the row axis or the column axis, and chose the option `Select Data...` to make a switch.

Exercise 5.4. The following table summarize results from a study on program selection and gender.

	Arts-Sci	Bus-Econ	Info Tech	Health Science	Graphics Design	Culinary Arts	Row Totals
Female	4,660	435	494	421	105	83	6,198
Male	4,334	490	564	223	97	94	5,802
Column Totals	8,994	925	1,058	644	202	177	12,000

Use Excel to answer the following question about the study.

(1) Is there an association between gender and program selection? Why or why not?

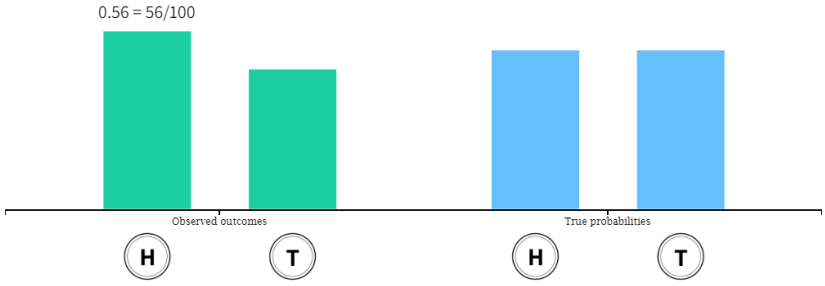
(2) If they are associated, is the association strong or week?

6 Basics of Probability

6.1 Experiments, Sample Spaces, and Events

- An **experiment** is a procedure that can be infinitely repeated and has a well-defined set of outcomes.
- An **outcome** is the result of a single trial (individual repetition) of an experiment.
- A **chance experiment** (or **random experiment**) is an experiment that has more than one possible outcome and whose outcomes cannot be predicted with certainty.
- The **sample space** of a chance experiment is the set of all possible outcomes.
- An **event** is a subset of the sample space.

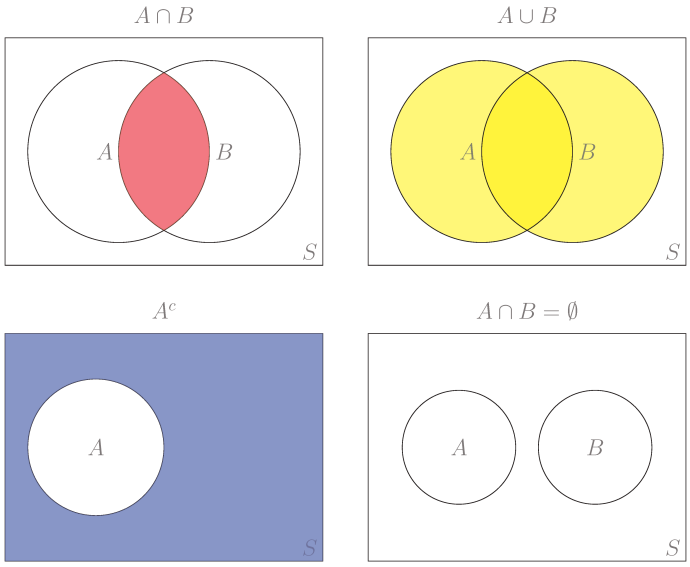
Example 6.1. A classic example of chance experiment is to toss a fair coin. The following figure shows observed outcomes from an experiment of tossing a coin 100 times as well as the true probabilities of getting a head and a tail.
(See https://seeing-theory.brown.edu)



6.2 Complement, Intersection and Union

- The **complement** E^c of event E is the set of all outcomes in a sample space that are **NOT** included in event E.
- The **intersection** $A \cap B$ of two events A and B is the set of all outcomes in the sample space that are shared by A and B.
- The **union** $A \cup B$ of two events A and B is the set of all outcomes in the sample space that are either in A or B.
- Two events A and B are **mutually exclusive** if there intersection $A \cap B$ is empty.

Venn Diagrams for Complement, Intersection, and Union



6.3 Classical Definition of Probability

- A **probability** $P(E)$ is the measures of how likely an outcome in the event E will occur in a chance experiment.
- **Equally likely** means that each outcome of an experiment occurs with equal chance.
- When the outcomes in the sample space of an chance experiment are ***equally likely***, the probability of an event E is

$$P(E) = \frac{\text{number of outcomes in } E}{\text{number of outcomes in } S}$$

- Chance experiment that involves tossing fair coins, rolling fair dice and drawing a card from a well-mixed deck of cards have equally likely outcomes.
- Note that many chance experiment do not have equally likely outcomes. For example, the majors of students in a class are not equally likely outcomes.

Example 6.2. Imagine flipping one fair coin (which means the chance of a head and the chance of a tail are the same). What is the probability of getting the head.

6.4 Empirical Probability

- An **empirical (or a statistical) probability** is the relative frequency of occurrence of outcomes from observations in repeated experiments:

$$P(E) = \frac{\text{number of occurrence of event } E}{\text{total number of observations}} = \frac{\text{frequency in } E}{\text{total frequency}}.$$

Example 6.3. A statistics class has 5 math majors and 20 other majors. If a students was randomly select from the class, what's the probability that the selected students is a math major?

Exercise 6.1. A group of people were asked if they had run a red light in the last year. 332 responded "yes", and 164 responded "no".

Find the probability that if a person is chosen at random, they have run a red light in the last year.

6.5 Theoretical Probability

- Theoretical probability is an expected value that can be calculated by mathematical theory and assumptions.
- When all outcomes in the sample space are equally likely, the probability of a desired event E, known as a **theoretical probability**, is calculated by

$$P(E) = \frac{\text{number of desired outcomes for event } E}{\text{number of all possible outcomes}}.$$

- **Tree diagrams** are often used for counting all possible outcomes.

Example 6.4. Find the probability of getting two heads when flipping a fair coins twice.

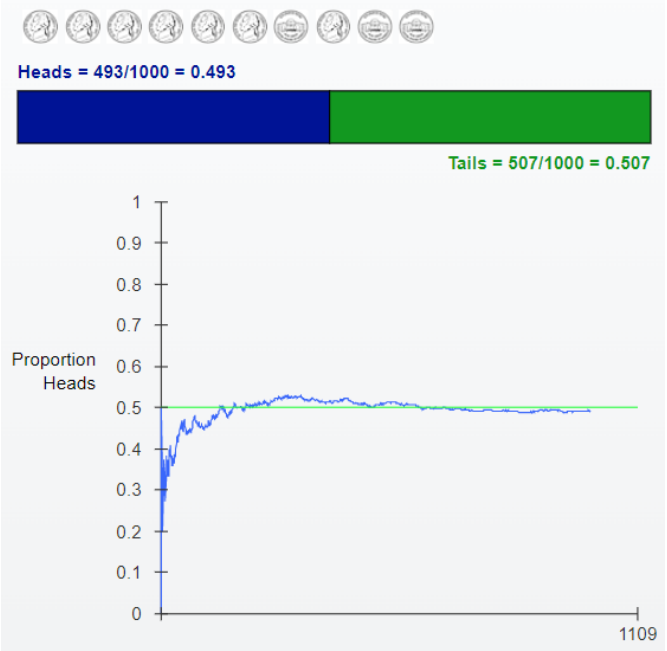
Exercise 6.2. Flipping a fair coin twice, find the probabilities of getting exactly one head.

6.6 Law of Large Numbers

Law of Large Numbers: As an experiment is repeated over and over, that is the number of trials getting larger and larger, the empirical probability of an event approaches the theoretical probability of the event. (Wiki: Law of large numbers.)

By the law of large number, we can say that the probability of any event is the **long-term relative frequency** of that event.

Example 6.5. The following figure shows the probability of simulating coin flipping 1000 times.



Source: http://digitalfirst.bfwpub.com/stats_applet/stats_applet_10_prob.html

6.7 Fundamental Properties (that Define Probability)

- **Property 1:** For an event E, the probability $P(E)$ is ranged from 0 to 1:
$$0 \le P(E) \le 1.$$
- **Property 2:** If S is the sample space, then $P(S) = 1$.
- **Property 3:** The probability of an event $E = \{e_1, e_2, \cdots e_k\}$ of distinct outcome is equal to the sum of probabilities of individual outcomes:
$$P(E) = P(e_1) + P(e_2) + \cdots + P(e_k)$$
where $P(e_i)$ is the probability of getting the outcome e_i.

Remark. When an event E consists of infinitely many outcomes, the right hand side of the equality in Property 3 will be an infinite sum.

- **Easy consequence 1:** If events A and B are mutually exclusive, then
$$P(A \cup B) = P(A) + P(B).$$
- **Easy consequence 2:** The probability $P(E)$ of an event E and the probability $P(E^c)$ of the complement event E^c satisfies the identity:
$$P(E) + P(E^c) = 1.$$
Equivalently,
$$P(E^c) = 1 - P(E) \quad \text{or} \quad P(E) = 1 - P(E^c).$$

Example 6.6. A six-sided fair die is rolled. Denote by E the event of getting a number less than 3.

(1) Find the probability $P(E)$ of the event E.

(2) Find the probability $P(E^c)$ of the complement E^c of the event E.

(3) Verify that $P(E) + P(E^c) = 1$.

Example 6.7. Two six-sided fair dice were rolled. Find the probability of getting two numbers whose sum is at least 4.

Exercise 6.3. Two six-sided fair dice were rolled. Find the probability of getting two numbers whose sum is at most 10.

Exercise 6.4. A bag of M&M's has 4 red, 6 green, 2 blue, and 3 yellow M&M's. What is the probability of randomly picking:

(1) a yellow?

(2) a blue or green?

(3) an orange?

6.8 The Addition Rule

When outcomes in the sample spaces are equally likely,

- the **probability of the intersection** of two events is

$$P(A \cap B) = \frac{\text{numbers of elements in } A \cap B}{\text{number of elements in the sample space } S}.$$

- the **probability of the union** of two events is

$$P(A \cup B) = \frac{\text{numbers of elements in } A \cup B}{\text{number of elements in the sample space } S}.$$

In general, the probability of the union of two events from a chance experiment is defined by the basic rules and the addition rule.

Addition Rule: the probability of the union of two events A and B is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Example 6.8. A card was randomly drew from a deck of 52 cards.

	Ace	2	3	4	5	6	7	8	9	10	Jack	Queen	King
Clubs													
Diamonds													
Hearts													
Spades													

(1) What's the probability of getting a heart?

(2) What's the probability of getting a face?

(3) What's the probability of getting a heart face?

(4) What's the probability of getting a heart or a face?

(5) What's the probability of getting a club and spade?

Exercise 6.5. A jar contains 10 red marbles numbered 1 to 10

and 8 blue marbles numbered 1 to 8. A marble is drawn at random from the jar. Find the probability of the given event, please show your answers as reduced fractions.

(1) The marble is red.

(2) The marble is odd-numbered.

(3) The marble is blue or even-numbered.

(4) The marble is blue or even.

6.9 The Conditional Probability

- The **conditional probability** of A given B, written as $P(A \mid B)$, is the probability that event A will occur given that the event B has already occurred.
- In the case that the chance experiment has equally likely outcomes, the conditional probability is,
$$P(A \mid B) = \frac{\text{numbers of elements in } A \cap B}{\text{number of elements in } B}.$$
- In general, we may use fundamental rules of probability and the multiplication rule to calculate the conditional probability.

Example 6.9. A fair die is rolled.

(1) Find the probability that the number rolled is a five, given that it is odd.

(2) Find the probability that the number rolled is odd, given that it is a five.

6.10 The Multiplication Rule

- **Multiplication Rule:** the probability of the intersection of two events A and B satisfies the following equality
$$P(A \cap B) = P(B)P(A \mid B) = P(A)P(B \mid A).$$
- The multiplication rule gives a formula for conditional probability:
$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \qquad P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

Example 6.10. The probability that a student borrows a statistics book from the library is 0.3. The probability that a student borrows a biology book is 0.4. Given that a student borrowed a biology book, the probability that he/she borrows a statistics book is 0.6.

(1) Find the probability that a student borrows a statistics book and a biology book.

(2) Find the probability that a student barrows a statistics boor or a biology book.

6.11 Independent Events

- Two events A and B are **independent** if
$$P(A \mid B) = P(A) \quad \text{or} \quad P(B) = P(B \mid A).$$

Equivalently,

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

- **Fundamental Counting Principle:** if there are m ways of doing something and n ways of doing another thing independently, then there are $m \cdot n$ ways of performing both actions *in order*.

Example 6.11. Consider flipping a fair coin and rolling a fair six-sided die together.

(1) What's the probability that the coin shows a head?

(2) Given that a head occurs, what's the probability that the die shows a number bigger than 4?

(3) What's the probability of getting a head and a number bigger than 4?

(4) Verify that flipping a head and rolling a number bigger than 4 are independent events.

6.12 Sampling with Replacement or without Replacement

- **With replacement:** If each member of a population is replaced after it is picked, then that member has the possibil-

ity of being chosen more than once. When sampling is done with replacement, then events are considered to be independent, meaning the result of the first pick will not change the probabilities for the second pick.

- **Without replacement:** When sampling is done without replacement, each member of a population may be chosen only once. In this case, the probabilities for the second pick are affected by the result of the first pick. The events are considered to be dependent.

Example 6.12. Two cards were randomly drawn from a standard deck of 52 cards *with replacement*. Find the probability of getting exactly one club card.

Example 6.13. Two cards were randomly drawn from a standard deck of 52 cards without replacement, which means the first card will not be put back.

(1) Find the probability that getting two spades.

(2) Find the probability that getting exactly one spade card.

Exercise 6.6. A hacker is trying to guess someone's password. The hacker knows (somehow) that the password is 11 characters long, and that each character is either a lowercase letter, (a, b, c, etc), an uppercase letter (A, B, C, etc) or a numerical digit (0, 1,

2, 3, 4, 5, 6, 7, 8, or 9). Assume that the hacker makes random guesses.

What is the probability that the hacker guesses the password on his first try?

Exercise 6.7. A special deck of 16 cards has 4 that are blue, 4 yellow, 4 green, and 4 red. The four cards of each color are numbered from one to four. A single card is drawn at random. Find the following probabilities.

(1) The probability that the card drawn is red.

(2) The probability that the card is red, given that it is not green.

(3) The probability that the card is red, given that it is neither red nor yellow.

(4) The probability that the card is red, given that it is not a four.

Exercise 6.8. Use the following probabilities of event A and B $P(A) = 0.33$, $P(B) = 0.47$ and $P(A \text{ and } B) = 0.20$ to find the probability $P(B \mid A^c)$.

Exercise 6.9. A box contains 10 pens, 6 black and 4 red. Two pens are drawn without replacement, which means that the first one is not put back.

(1) What is the probability that both pens are red?

(2) What is the probability that at most one pen is red?

(3) What is the probability that at least one pen is red?

7 Discrete Random Variables

7.1 Random Variables

- A **random variable**, usually written X, is a variable whose values are numerical quantities of possible outcomes a random experiment.
- A **discrete random variable** takes on only a finite or countable number of distinct values.
- A **continuous random variable** takes on values which form an interval of numbers.

Example 7.1. • Rolling a fair dice, the number of dots on the top faces is a discrete random variables takes on the possible values: 1, 2, 3, ,4, 5, 6.
• Flipping a fair coin 10 times, the number of heads is a discrete random variable takes on the possible values: 1, 2, 3, ..., 10.

Example 7.2. • The height of an randomly select 10 year-old boy in US is normally between 129 cm and 157 cm. So the height is a continuous random variable.
• The measure the voltage at an randomly electrical outlet normally is between 118 and 122. So the measure of voltage is a continuous random variable.

Exercise 7.1. Classify each random variable as either discrete or continuous.

(1) The number of boys in a randomly selected three-child family.

(2) The temperature of a cup of coffee served at a restaurant.

(3) The number of math majors in randomly selected group of 10 students.

(4) The amount of rain recorded in a small town one day.

7.2 Probability Distributions

- The **probability distribution** of a discrete random variable X is defined by the probability $P(X = x)$ associated with each possible value x of the variable X. The function $p_X(x) = P(X = x)$ is called the **probability mass function**.
- A probability distribution of a discrete random variable is usually characterized by a table of all possible values X together with probabilities $P(X)$, or a probability histogram, or a formula.
- A random variable X (discrete and continuous) always has a **cumulative distribution function**: $F_X(x) = P(X \leq x)$ (= $\sum_{x_i \leq x} P(x_i)$ if X is discrete).

7.3 Basic Properties of Probability Distributions

- Basic rules of probability:
  - $0 \leq P(X = x) \leq 1$.
  - the sum of all the probabilities is 1, that is $P(X \leq x_{max}) = 1$.
  - In particular, $0 \leq F_X(x) \leq 1$.
  - The cumulative distribution function $F_X(x)$ is non-decreasing.
- The probability distribution can be recovered from its cumulative distribution function. Indeed, for a *discrete* random variable X, we have

$$P(X = x_i) = P(X \leq x_i) - P(X \leq x_{i-1}),$$

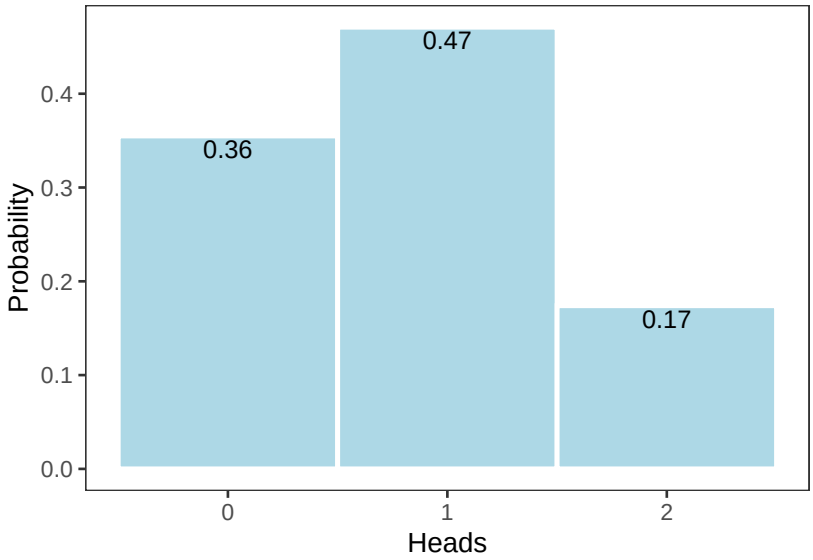
where $P(X \leq x_i) = \sum_{k=1}^{i} P(X = x_k)$.

Example 7.3. Let X be the number of heads that are observed when tossing two fair coins.

(1) Construct the probability distribution for X.

(2) Find $P(X \leq 1)$ and $P(X \leq 2)$.

Example 7.4. The probability distribution of an unfair coin is characterized by the following histogram. Find the probability of getting at most 1 head.



Exercise 7.2. A pair of fair 6-sided dice were rolled. Let X denote the sum of the number of dots on the top faces.

(1) Construct the probability distribution of X.

(2) Find the probability that X takes an odd value.

Exercise 7.3. The number X of days in the summer months that

a construction crew cannot work because of the weather has the probability distribution

x	6	7	8	9	10	11	12	13	14
P(x)	0.03	0.08	0.15	0.2	0.19	0.16	0.1	0.07	0.02

(1) Find the probability that no more than ten days the construction crew cannot work in the summer.

(2) Find the probability that from 8 to 12 days the construction crew cannot work in the summer.

(3) Find the probability that no days at all the construction crew cannot work in the summer.

7.4 Mean and Standard Deviation of a Discrete Random Variable

Let X be a discrete random variable and $p_X(x) = P(X = x)$ the probability mass function.

- The **expected value** $E(X)$ (also called **mean** and denoted by μ) of the discrete random variable X is the number
$$\mu = E(X) = \sum x p_X(x).$$
- The **variance** $\mathrm{Var}(X)$ (also denoted by σ^2) of the discrete random variable X is the number
$$\sigma^2 = \mathrm{Var(X)} = \sum (x - E(X))^2 p_X(x).$$

- The **standard deviation** σ of a discrete random variable X is the square root of its variance:

$$\sigma = \sqrt{\sum (x - E(X))^2 p_X(x)}.$$

Example 7.5. One thousand raffle tickets are sold for \$2 each. Each has an equal chance of winning. First prize is \$500, second prize is \$300, and third prize is \$100. Find the expected value of the net gain, and interpret its meaning.

Example 7.6. The waiting time (rounded to multiples of 5) in the cafeteria at a Community College has the following probability distribution. Find the expected waiting time and the standard deviation.

x (minutes)	5	10	15	20	25
$P(X = x)$	0.13	0.25	0.31	0.21	0.1

Example 7.7. The probability distribution of an unfair die is given in the following table.

x	1	2	3	4	5
$P(X = x)$	0.18	0.12	?	0.14	0.23

(1) Find $P(X = 3)$.

(2) Find the mean, variance and standard deviation of this probability distribution.

Exercise 7.4. Seven thousand lottery tickets are sold for \$5 each. One ticket will win \$2,000, two tickets will win \$750 each, and five tickets will win \$100 each. Let X denote the net gain from the purchase of a randomly selected ticket.

(1) Construct the probability distribution of X.

(2) Compute the expected value $E(X)$ of X. Interpret its meaning.

(3) Compute the standard deviation σ of X.

7.5 Binomial Distribution

- A **binomial experiment** is a probability experiment satisfying:

  (1) The experiment has a fixed number n of independent trials.

  (2) Each trial has only two possible outcomes: a success (S) or a failure (F).

  (3) The probability p of a success is the same for each trial.

- The discrete random variable X counting the number of successes in the n trials is the **binomial random variable**. We say X has a **binomial distribution** with parameters n and p and write it as $X \sim B(n, p)$.
- For $X \sim B(n, p)$, the **probability of getting exactly x successes in n trials** is

$$P(X = x) = B(x, n, p) = {}_nC_x p^x (1-p)^{n-x} = \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x},$$

  where $n! = n(n-1)\cdots 1$, read as n factorial, for $n > 0$ and $0! = 1$.
- The notation ${}_nC_x = \frac{n!}{(n-x)!x!}$ is read as n choose x, which is the number of ways to choose x objects from a set of n objects.

Example 7.8. A card is randomly selected from a standard deck and replaced. This experiment is repeated a total of 5 times.

- Find the probability of getting exactly 3 clubs.
- Find the probability of getting at least 3 clubs.

Exercise 7.5. Let X be a binomial random variable with parameters $n = 5$, $p = 0.2$. Find the probabilities

(1) $P(X = 3)$,

(2) $P(X < 3)$,

(3) $P(X > 3)$.

Exercise 7.6. A manufacturing machine has a 4% defect rate.

If 6 items are chosen at random, what is the probability that at least one will have a defect?

7.6 Mean and Standard Deviation of Binomial Distribution

- The mean of a binomial distribution of n trials is

$$\mu = \sum xP(X = x) = \sum x \cdot \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x} = np.$$

- The variance of a binomial distribution of n trials is

$$\sigma^2 = \sum (x-np)^2 P(X = x) = \sum x^2 P(X = x) - (np)^2 = np(1-p).$$

- The standard deviation of a binomial distribution of n trials is

$$\sigma = \sqrt{np(1-p)}.$$

- We consider an event E **unusual** if the probability $P(E) \leq 5\%$.

Example 7.9. The probability that an egg in a retail package is cracked or broken is 0.02.

(1) Find the average number of cracked or broken eggs in a one dozen carton.

(2) Find the standard deviation.

(3) Is getting at least two broken eggs unusual?

Exercise 7.7. Adverse growing conditions have caused 5% of grapefruit grown in a certain region to be of inferior quality. Grapefruit are sold by the dozen.

(1) Find the average number of inferior quality grapefruit per box of a dozen.

(2) A box that contains two or more grapefruit of inferior quality will cause a strong adverse customer reaction. Find the probability that a box of one dozen grapefruit will contain two or more grapefruit of inferior quality.

Exercise 7.8. CCA has stated in 2017 that 48% of its students are first generation college students.

Suppose you sample 5 CCA students and ask if they are first generation college students or not, counting the number of first generation students.

(1) Create a binomial probability distribution (table) for this situation.

(2) Find the mean of the binomial distribution.

(3) Find the standard deviation of this binomial distribution.

7.7 Extra Practice Problems

Exercise 7.9. Find the mean and the standard deviation of the probability distribution.

x	P(x)
0	0.2
1	0.2
2	0.1
3	0.5

Exercise 7.10. A company tracks the number of sales new employees make each day during a 100-day probationary period. The results for one new employee are shown at the right.

Sales per day x	Number of days f
0	16
1	19
2	15
3	21
4	9
5	10
6	8
7	2

(1) Find the probability of each outcome.

(2) Construct a probability distribution table.

(3) Find the mean of the probability distribution.

(4) Find the variance and standard deviation.

Exercise 7.11. A poll is given, showing 35% are in favor of a new building project.

If 5 people are chosen at random, what is the probability that exactly 2 of them favor the new building project?

7.8 Lab: Binomial Distribution

Let X be a binomial random variable with parameters n and p, that is $X \sim B(n, p)$. In Excel, $P(X = x)$ is given by
`BINOM.DIST(x, n, p, FALSE)`
and $P(X \leq x)$ is given by
`BINOM.DIST(x, n, p, TRUE)`.
You may click input function f_x and then search `binom` to find the function.

Exercise 7.12. A type of surgery has a 90% chance of success. The surgery is performed on 7 patients.

Use excel to answer the following questions.

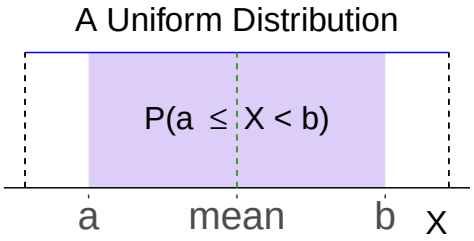
(1) Find the probability of the surgery being successful on exactly 5 patients.

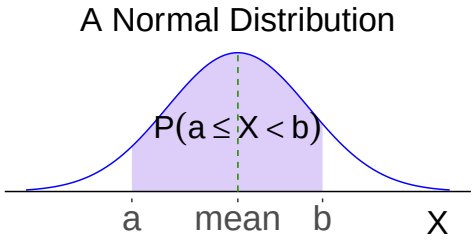
(2) Find the probability of the surgery being successful on at least 4 patients.

8 Continuous Random Variables

8.1 Probability Distribution of a Continuous Random Variable

- The probability distribution of a continuous random variable X is characterized by its **probability density function** $f(X)$ satisfying that the probability $P(a \le X \le b)$ equals the area above the interval $[a, b]$ but under the graph of the density function $f(X)$ which is also called a **density curve**.





8.2 Properties of Probability Distribution of a Continuous Random Variable

- The probability density function f is nonnegative, that is $f(X) \ge 0$.
- The total area under a density curve is 1.
- The cumulative probability $P(X \le b)$ of a random variable X equals the area under the density curve to the left side of b.
- By the addition rule of probability, we have

$$P(a \le X \le b) = P(X \le b) - P(X \le a)$$

$$P(X \ge b) = 1 - P(X \le b)$$

- As a line segment has no area, we have $P(X \le a) = P(X < a)$ as well as $P(X \ge b) = P(X > b)$

Example 8.1. Let X be the amount of time that a commuter must wait for a train. Suppose X has a probability density function

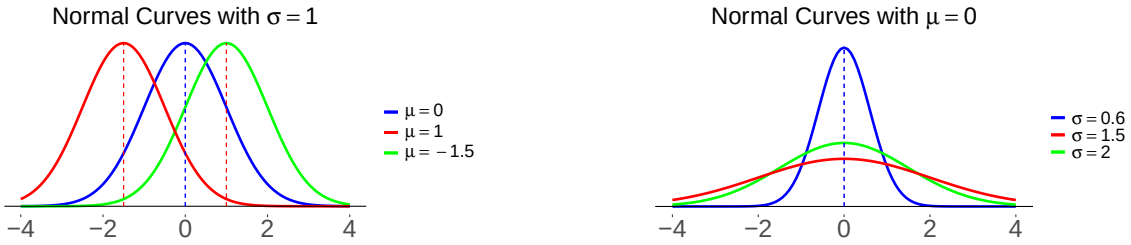
$$f(X) = \begin{cases} 0.1, & 0 \le X \le 10 \\ 0, & \text{otherwise} \end{cases}$$

What is the probability that the commuter's waiting time is less than 4 minutes?

8.3 Normal Distribution

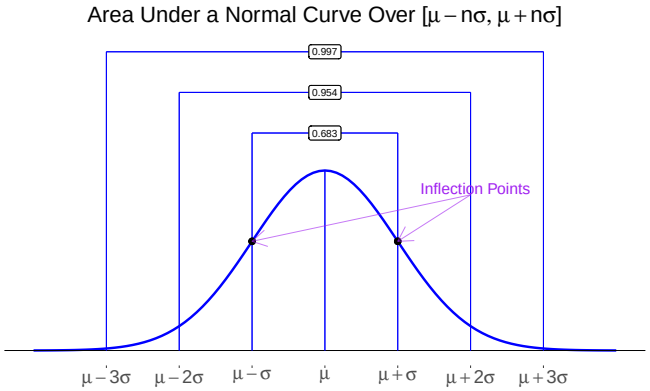
- A **normal distribution** has a **density function**
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$
where μ is the mean, σ is the standard deviation, $\pi \approx 3.14159$ and $e \approx 2.71828$. The graph of f is called a **normal curve**.
- We write $X \sim N(\mu, \sigma^2)$ for a normal random variable X with the mean μ and the standard deviation σ.
- A normal distribution has the following properties:
  - *The mean, median, and mode are equal.*
  - The normal curve is *bell shaped and* ***symmetric*** with respect to the mean.
  - The *total area* under the curve and above the x-axis is 1.
  - The normal curve *approaches, but never touches, the x-axis* as x goes to $\pm\infty$.
  - Between $\mu-\sigma$ and $\mu+\sigma$, the graph *curves downward.* On the left side of $\mu-\sigma$ or the right side of $\mu+\sigma$, the graph *curves upward.* A point at which the curve changes the direction of curving is called an **inflection point**.

Normal Curves with Different Means and Standard Deviations



8.4 The Empirical Rule for Normal Distributions

For any normal distribution, the proportion of data values within 1, 2, and 3 standard deviations away from the mean are approximately 68.3%, 95.4% and 99.7% respectively.



Example 8.2. Suppose that foot length of a randomly chosen adult male is a normal random variable with the mean $\mu = 11$ and the standard deviation $\sigma = 1.5$.

(1) How likely is a male's foot length to be smaller than 9.5 inches

(2) How likely is a male's foot length to be bigger than 8 inches

8.5 Standard Normal Distribution

- A normal distribution is called a **standard normal distribution** if the mean is $\mu = 0$ and the standard deviation is $\sigma = 1$.
- A random normal variable can be **standardized** by the following formula $z = \frac{x-\mu}{\sigma}$. We call the value z the Z-**score** of x.

In Excel, the Z-score of x can be calculated using the function `STANDARDIZE()`.

- Standardization preserves probability:

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right).$$

- The probability $P(Z < z)$ of a standard normal random variable Z can be found using the Excel function `NORM.S.DIST(z, TRUE)`.
- The probability $P(X < x)$ of a normal random variable X can be calculated using the Excel function `NORM.DIST(x, mean, sd, TRUE)`.

Example 8.3. Let X be a norma random variable with the mean $\mu = 8$ and the standard deviation $\sigma = 2$.

(1) Find the Z-score for the value $X = 13$.

(2) Find the X-value for the Z-score $z = -0.6$.

Example 8.4. Let Z be a standard normal random variable.

(1) Find $P(Z < 1.21)$.

(2) Find $P(Z \geq 1.21)$.

(3) Find $P(0 < Z \leq 1.21)$.

Example 8.5. The heights of 25-year-old women in a certain region are approximately normally distributed with mean 62 inches and standard deviation 4 inches. Find the probability that a randomly selected 25-year-old woman is more than 67 inches tall.

8.6 Cutoff Value for a Given Tail Area

- The k-th percentile for a random variable X is the value x_k that cuts off a left tail with the area $k/100$, that is $P(X < x_k) = \frac{k}{100}$, where $0 \leq k \leq 100$.
- Let c be a nonnegative number less than or equal to 1. The $(100c)$-th percentile for the standard normal distribution is usually denoted as $-z_c$, that is $P(Z < -z_c) = c$. By symmetry, z_c is the value such that $P(Z > z_c) = c$, that is $P(Z < z_c) = 1 - c$.
- For a normal random variable X with the mean μ and standard deviation σ, the cutoff value x^* with a **tail area** c, can be calculated using the standardization formula, that is,

$$x^* = z^* \cdot \sigma + \mu,$$

where z^* is the cutoff z-score with the tail area c, that is $z^* = -z_c$ given that c is the left-tail area and $z^* = z_c$ given that c is the right tail area.

Example 8.6. Let X be the normal random variable with mean 6 and standard deviation 3. Suppose the value x^* cuts off a left-tail area 0.05. Find the value x^*.

Example 8.7. Scores on a standardized college placement examination are normally distributed with mean 60 and standard deviation 13. Students whose scores are in the top 5% will be placed in a Calculus II course. Find the minimum score needed to be placed in a Calculus II course.

Exercise 8.1. A man arrives at a bus stop at a random time (that is, with no regard for the scheduled service) to catch the next bus. Buses run every 30 minutes without fail, hence the next bus will come any time during the next 30 minutes with evenly distributed probability (a uniform distribution). Find the probability that a bus will come within the next 10 minutes.

Exercise 8.2. (1) Let Z be a standard normal random variable. Find the probabilities:

1. $P(Z < 1.58)$ 2. $P(-0.6 < Z < 1.67)$ 3. $P(Z > 0.19)$.

(2) Let X be a normal random variable with $\mu = 5$ and $\sigma = 2$.

Find the probabilities:

1. $P(-2 < X < 8)$ 2. $P(X > -1)$ 3. $P(X < 4)$.

Exercise 8.3. The lifetimes of the tread of a certain automobile tire are normally distributed with mean 37,500 miles and standard deviation 4,500 miles. Find the probability that the tread life of a randomly selected tire will be between 30,000 and 40,000 miles.

Exercise 8.4. A manufacturer knows that their items have a normally distributed lifespan, with a mean of 9.6 years, and standard deviation of 3 years.

The 4% of items with the shortest lifespan will last less than how many years?

Exercise 8.5. The life of a particular battery is known to follow a normal distribution , with a mean of 1133 hours and a standard deviation of 105 hours.

(1) What percent of batteries last less than 1033 hours?

(2) The 86 th percentile is represented by what number of hours of battery life?

(3) What is the probability that a randomly selected battery will last more than 1362 hours?

Exercise 8.6. Let Z be a normal random variable with $\mu = 0$ and $\sigma = 1$. Let X be a normal random variable with $\mu = 4.3$ and $\sigma = 1.7$.

Determine the values $P(Z > 1) + P(X < 6)$ and explain how do you find the value.

8.7 Lab: Normal Distributions

- Let Z be a standard normal random variable. In Excel, $P(Z < z)$ is given by `NORM.S.DIST(z, TRUE)`.
- Let X be a normal random variable with mean μ and standard deviation σ, that is $X \sim N(\mu, \sigma^2)$. In Excel, $P(X < x)$ is given by `NORM.DIST(x, mean, sd, TRUE)`.
- When a cumulative probability $p = P(X < x)$ of a normal random variable X is given, we can find x using `NORM.INV(p, mean, sd)`.
- When a cumulative probability $p = P(Z < z)$ of a standard normal random variable Z is given, we can find z using `NORM.S.INV(p)`.

Exercise 8.7. Let Z be a standard normal random variable. Find each of the following probabilities. **Write down the Excel function you used to do the calculation.**

(1) $P(Z < 0.96)$

(2) $P(Z > -1.43)$

(3) $P(-0.47 < Z \leq 2.31)$

(4) $P(Z < 1.23 \text{ or } Z > 2.13)$

Exercise 8.8. Let X be a normal random variable with mean 52 and standard deviation 7. Find each of the following probabilities. **Write down the Excel function you used to do the calculation.**

(1) $P(X < 62)$

(2) $P(X > 35)$

(3) $P(41 < X < 58)$

(4) $P(X < 51 \text{ or } Z > 67)$

9 Sampling Distributions

- When using sample statistics to estimate population parameter, there will be a chance error

  Population Parameter = Sample Statistic + Chance Error.
- To understand the chance error, we need to know how sample statistics distribute. Consider samples of the same size n randomly chosen from the population with replacement.
- The probability distribution of a sample statistic is called a **sampling distribution**.
- The sampling distribution varies as the sample size changes. In general, A larger sample size will result a smaller standard deviation of the sampling distribution.
- The standard deviation of a sampling distribution is also called the **standard error**.

9.1 Central Limit Theorem for Mean

Theorem 9.1 (The Central Limit Theorem)**.** *As the sample size n increases, the sampling distribution of the sample mean, from a population with the mean μ and the standard deviation σ, will approach to a normal distribution with the mean $\mu_{\overline{X}} = \mu$ and the standard deviation $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$.*

Remark.
- In terms of standardization, the central limit theorem says that the random variable $\overline{Z} = \dfrac{\overline{x} - \mu}{\sigma/\sqrt{n}}$ has an approximately standard normal distribution.
- For most distributions (**not highly skewed**), when sample size $n > 30$, the sampling distribution of the sample mean $\overline{X}$ can be approximated reasonably well by a normal distribution. The larger the sample size, the better the approximation will be.
- When the population is normally distributed, the sampling distribution of the sample means will be normally distributed for any sample size.
- If the population distribution is highly skewed, relying on CLT can be risky.

See the discussion on intuitive explanation: https://bit.ly/3dtf0q0

Example 9.1. Randomly draw samples of size 2 with replacement from the numbers 1, 3, 4.

(1) Find the sampling distribution of sample means.

(2) Find the mean, and standard deviation of the sample means.

(3) Find the mean, and standard deviation of the population.

(4) How are the means of the population and the sampling distribution related.

(5) How are the standard deviations of the population and the sampling distribution related.

Example 9.2. Suppose the mean length of time that a caller is placed on hold when telephoning a customer service center is 23.8 seconds, with standard deviation 4.6 seconds. Find the probability that the mean length of time on hold in a random sample of 1,000 calls will be within 0.5 second of the population mean.

Example 9.3. Suppose speeds of vehicles on a particular stretch of roadway are normally distributed with mean 36.6 mph and standard deviation 1.7 mph.

(1) Find the probability that the speed X of a randomly selected vehicle is between 35 and 40 mph.

(2) Find the probability that the mean speed $\overline{X}$ of 10 randomly selected vehicles is between 35 and 40 mph.

9.2 Sampling Distribution of a Sample Proportion

The proportion of a specific characteristic in a data set can be viewed as the mean of the data set by identifying the specific characteristic with 1 and others with 0.

Example 9.4. Consider the following data set

1, 0, **1**, **1**, 0, 0, **1**, 0, **1**, **1**

(1) What proportion of the numbers are in bold?

(2) What's the mean of the data set?

(3) Is there any relation between the proportion and the mean? If so, describe it.

- In general, if a population consisting of 1s and 0s, then the proportion p of 1s is the same as the mean. The standard deviation is
$$\sigma = \sqrt{(1-p)^2 p + (0-p)^2(1-p)} = \sqrt{p(1-p)}.$$

9.3 Central Limit Theorem for Proportion

For a sampling distribution of sample proportion, we write $\hat{P}$ for the random variable of sample proportions.

Theorem 9.2. ***Central Limit Theorem for Proportion:***
For large samples, the distribution of sample proportions $\hat{P}$ is approximately normal, with the mean $\mu_{\hat{P}} = p$ and the standard deviation $\sigma_{\hat{P}} = \sqrt{\frac{p(1-p)}{n}}$, where p is the population proportion and n is the sample size.

- Because a sample proportion is always between 0 and 1, and 99.7% of sample proportions lie within 3 standard deviation away from the population proportion. When using the central limit theorem for proportion, we require the sample size n satisfying the following condition: the interval
$$\left[p - 3\sqrt{\frac{p(1-p)}{n}}, p + 3\sqrt{\frac{p(1-p)}{n}}\right]$$
lies wholly in the interval $[0, 1]$.
- In practice, if n satisfies the following two inequalities: $np \geq 10$ and $n(1-p) \geq 10$, then we consider n is large enough for assuming that the sampling distribution of the sample proportion is approximately normal.
- When the population proportion p is unknown, to apply the central limit theorem for proportion, we require the sample size n satisfying the same conditions with p replaced by the

sample proportion $\hat{p}$. That is, the sample size n should satisfies $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$.

Example 9.5. Suppose that in a population of voters in a certain region 53% are in favor of a particular law. Nine hundred randomly selected voters are asked if they favor the law.

Find the probability that the sample proportion computed from a random sample of size 900 will be at least 2% above true population proportion.

Example 9.6. Suppose that in 36% of all car accidents involve injury. Find the probability that the injury rate in a random sample of 250 car accidents is between 30% and 45%.

Exercise 9.1. An unknown distribution has a mean of 28 and a standard deviation 6. Samples of size $n = 40$ are drawn randomly from the population. Find the probability that the sample mean is between 27 and 30.

Exercise 9.2. The numerical population of grade point averages at a college has mean 2.61 and standard deviation 0.5. If a random sample of size 100 is taken from the population, what is the probability that the sample mean will be between 2.51 and 2.71?

Exercise 9.3. An airline claims that 72% of all its flights to a certain region arrive on time. In a random sample of 30 recent arrivals, 19 were on time. You may assume that the normal distribution applies.

(1) Compute the sample proportion.

(2) Assuming the airline's claim is true, find the probability of a sample of size 30 producing a sample proportion so low as was observed in this sample.

Exercise 9.4. In a mayoral election, based on a poll, a newspaper reported that the current mayor received 45% of the vote. If this is true, what is the probability that a random sample of 100 voters had less than 35% voting for the current mayor?

9.4 More Practice on Sampling Distributions

Exercise 9.5. A population has the mean 73.5 and the standard deviation 2.5.

(1) Find the mean and standard deviation of $\overline{X}$ for samples of size 40.

(2) Find the probability that the mean of a sample of size 40 will be less than 72.

Exercise 9.6. A normally distributed population has the mean 57.7 and the standard deviation 12.1.

(1) Find the probability that a single randomly selected element X of the population is less than 45.

(2) Find the mean and standard deviation of $\overline{X}$ for samples of size 16.

(3) Find the probability that the mean of a sample of size 16 drawn from this population is less than 45.

Exercise 9.7. Suppose the mean amount of cholesterol in eggs

labeled "large" is 186 milligrams, with the standard deviation 7 milligrams. Find the probability that the mean amount of cholesterol in a sample of 144 eggs will be within 2 milligrams of the population mean.

Exercise 9.8. Suppose that 8% of all males suffer some form of color blindness. Find the probability that in a random sample of 250 men at least 10% will suffer some form of color blindness.

Exercise 9.9. An airline claims that 72% of all its flights to a certain region arrive on time. In a random sample of 30 recent arrivals, 19 were on time. You may assume that the normal distribution applies.

(1) Compute the sample proportion.

(2) Assuming the airline's claim is true, find the probability of a sample of size 30 producing a sample proportion so low as was observed in this sample.

Exercise 9.10. A particular fruit's weights are normally distributed, with a mean of 663 grams and a standard deviation of 38 grams. If you pick 13 fruits at random, then 9% of the time, their mean weight will be greater than how many grams?

9.5 Lab: Normal Distributions

- Let X be a normal random variable with mean μ and standard deviation σ, that is $X \sim \mathcal{N}(\mu, \sigma^2)$. In Excel, $P(X < x)$ is given by `NORM.DIST(x, mean, sd, TRUE)`.
- Recall the mean of a data set can obtained by the Excel function `AVERAGE()`.
- Given the population mean μ and standard deviation σ, if the sample size n is bigger than 30 and the sample mean is $\bar{x}$. The probability of getting another sample of the same size but smaller mean can be obtained by the following Excel function: `NORM.DIST(` $\bar{x}, \mu, \sigma$ `/sqrt(n),TRUE)`.

Exercise 9.11. CNNBC recently reported that the mean annual cost of auto insurance is 1035 dollars. Assume the standard deviation is 109 dollars. Assume the annual cost is normally distributed.

(1) Find the probability that a single randomly selected policy has a mean value between 1028.6 and 1044.8 dollars.

(2) Find the probability that a random sample of size $n = 79$ has a mean value between 1028.6 and 1044.8 dollars.

10 Confidence Intervals

10.1 Point Estimation

- When estimating a population parameter, we may consider the statistic of a random sample as an estimate. But we expect some chance error.
- Estimating an unknown parameter by a single number calculated from a sample is called a **point estimation**. The single number (statistic) from the sample is called a **point estimate**.
- Point estimate gives no indication of how reliable the estimate is or how large the error is.

Example 10.1. From a box of 20 pencils of two colors, black and blue, 10 pencils were randomly drawn. 6 out of the 10 pencils are black. What proportion of black pencils are in the box.

10.2 Interval Estimation

- To increase the chance, we estimate an unknown parameter using intervals that are obtained by adding chance errors to a point estimate.
- Estimating an unknown parameter using an interval of values which likely contains the true value of the parameter is called a **interval estimation**. The interval is called an **interval estimate**.
- The reliability of an interval estimate is measured by the probability $1 - \alpha$ that the interval estimate will capture the true value of the parameter. This probability $1 - \alpha$ is called the **confidence level**.
- The 90%, 95% and 99% levels of confidence are frequently used in statistical study. The 95% level of confidence is usually the

standard choice of confidence level for scientific polls published in the media and online.

Example 10.2. Recall that the **standard error** of a statistic, denoted by SE, is the standard deviation of the sampling distribution.

A randomly selected 100 students at a college have an average GPA 3.0. How likely does the interval $[3.0-2\cdot\text{SE}, 3.0+2\cdot\text{SE}]$ contain the average GPA μ of that college?

10.3 Confidence Interval

- When the sampling distribution of a statistic is approximately symmetric, we take interval estimates in the following form [Statistic − E, Statistic + E], where the value E is called the **marginal error** or **margin of error**.
- Given a confidence level $100(1-\alpha)\%$, the marginal error E is the value such that $100(1-\alpha)\%$ of the intervals [Statistic − E, Statistic+E] contains the true parameter μ_{par}. Equivalently, the marginal error E is the value such that $100(1-\alpha)\%$ of statistics are in the interval $[\mu_{\text{par}} - \text{E}, \mu_{\text{par}} + \text{E}]$.
- Denote by X the random variable for the sample statistic. Then E is determined the following probability equation
$$P(\mu_{\text{par}} - \text{E} < X < \mu_{\text{par}} + \text{E}) = 1-\alpha.$$
If the distribution of X is symmetric, then the marginal error E is the value such that
$$P(X - \mu_{\text{par}} < \text{E}) = 1 - \alpha/2.$$
- Because the parameter μ_{par} is unknown. If we standardize the random variable X by $Z = \frac{X-\mu_{\text{par}}}{\text{SE}}$, we get
$$P\left(-\tfrac{\text{E}}{\text{SE}} < Z < \tfrac{\text{E}}{\text{SE}}\right) = 1-\alpha,$$
where the random variable Z has a mean 0 and standard deviation 1.

- The above probability equation suggests the following formula

$$\text{Marginal Error} = \text{Critical value} \cdot \text{Standard Error},$$

where the **critical value** is the value $z_{\alpha/2}$ so that $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$.
- Let X be a point estimate, we call the interval $[X - z_{\alpha/2}\text{SE}, X + z_{\alpha/2}\text{SE}]$ a **confidence interval** at the $100(1-\alpha)\%$ level of confidence.

A Visualization of Confidence Intervals for Mean
https://rpsychologist.com/d3/CI/

10.4 Confidence Intervals for Mean with Known Population SD

- Suppose the population standard deviation σ is given. By the central limit theorem, if $n > 30$ or the population distribution is approximately normal, then the sampling distribution is approximately normal with the standard error $\sigma/\sqrt{n}$.
At the confidence level $1-\alpha$, the marginal error for a population mean is $E = z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$ and the confidence interval is

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right],$$

where the **critical value** $z_{\alpha/2}$ satisfies that $P(Z < z_{\alpha/2}) = 1 - \alpha/2$ for the standard normal variable Z.
- In Excel,
$z_{\alpha/2}$=NORM.S.INV((1+confidence level)/2).
- The marginal error $E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ can also be obtained by the Excel function
CONFIDENCE.NORM(1-confidence level, σ, n).

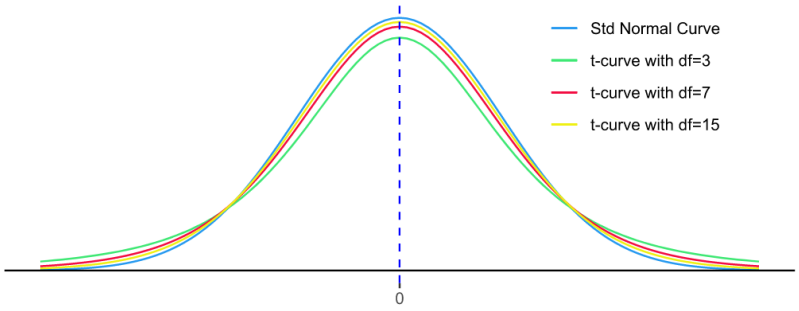
Example 10.3. A sample of size 15 drawn from a normally distributed population with the standard deviation 6. Find the critical value $z_{\alpha/2}$ needed in construction of a confidence interval:
(1) when the level of confidence is 90%;

(2) when the level of confidence is 98%.

Example 10.4. A random sample of 50 students from a college gives a mean GPA 2.51. Suppose the standard deviation of GPAs of all students at the college is 0.43. Construct a 99% confidence interval for the mean GPA of all students at the college.

10.5 Student's t-Distribution

- When the population standard deviation is unknown, we may replace σ by the sample standard deviation s and use $s/\sqrt{n}$ as an estimate to the standard error for the sampling distribution of the sample mean.
- When we use the estimated standard error $s/\sqrt{n}$ to build a confidence interval, the normal distribution may NOT be appropriate for calculating the critical value. Because the t-statistic $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$ is not normal in general.
- The random variable $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$ has approximately a **Student's t-distribution with the degree of freedom** $n - 1$ if the random variable X is approximately normal.
- When the random variable X is not approximately normal but not highly skewed, people usually assume that the t-statistic is normal if the sample size n is large enough, say, $n \geq 30$.



- The t-distributions form a family of curves, called t-**curves**, parameterized by the degrees of freedom.
- The t-distribution has the following important properties.
  - Similar to the standard normal curve, it is symmetric about 0 and the total area under a t-curve is 1.
  - The t-distribution has slightly more variation (i.e. t-curves are slightly "fatter") than the standard normal distribution.
  - When the degree of freedom increases, the t-distribution becomes closer to the standard normal distribution.

10.6 Confidence Intervals for a Mean with Unknown Population SD

- Suppose the population is approximately normally distributed or not highly skewed with large sample size. At the confidence level $1-\alpha$, the margin of error is $E = t_{\alpha/2}\frac{s}{\sqrt{n}}$, and the confidence interval for a population mean μ is
$$\left[\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right],$$
where $t_{\alpha/2}$ is the critical value such that $P(T < t_{\alpha/2}) = 1-\alpha/2$ for a Student t-distribution with degree of freedom $n-1$.
- In Excel, the critical value $t_{\alpha/2}$ can be calculated by `T.INV((1+confidence level)/2, n-1)` or `T.INV.2T(1-confidence level, n-1)`, where n is the sample size.
- The marginal error $E = t_{\alpha/2}\frac{s}{\sqrt{n}}$ can also be obtained by the Excel function `CONFIDENCE.T(1-confidence level, s, n)`.

Example 10.5. A sample of size 15 drawn from a normally distributed population. Find the critical value $t_{\alpha/2}$ needed in construction of a confidence interval:

(1) when the level of confidence is 99%;

(2) when the level of confidence is 95%.

Example 10.6. A sample of size 16 is randomly drawn from a normally distributed population. The sample has a mean 79 and standard deviation 7. Construct a confidence interval for that population mean at the 90% level of confidence.

Example 10.7. The data blow shows numbers of hours worked from 40 randomly selected employees from several grocery stores in the county.

30, 26, 33, 26, 26, 33, 31, 31, 21, 37, 27, 20, 34, 35, 30, 24, 38, 34, 39, 31,
22, 30, 23, 23, 31, 44, 31, 33, 33, 26, 27, 28, 25, 35, 23, 32, 29, 31, 25, 27

Construct 99% confidence interval for the mean worked time.

10.7 Choose Between Normal Distribution and t-Distribution

- Population is approximately normally distributed.
  - the population standard deviation σ is known: use the normal distribution.
  - the population standard deviation σ is *unknown*: use the *t-distribution*.
- Population distribution unknown but **not highly skewed**. If the **sample size is large** enough, i.e. $n > 30$, then
  - the population standard deviation σ is known: use normal distribution.
  - the population standard deviation σ is *unknown*: either one can be used but the *t-distribution* is more accurate.
- **Warning:** When the population distribution is unknown and the sample size is small, neither the t-distribution nor the normal distribution is reliable.
- For small samples, there is method called "The Shapiro–Wilk test" which can be used to determine if we may assume the sampling distribution is approximately normal.
- Even when $n > 30$, a visual inspection (using histogram for example) of the normality is necessary.

10.8 Practice

Exercise 10.1. Decide whether the following statements are true or false. Explain your reasoning.

- The statement, "the 95% confidence interval for the population mean is (350, 400)" means that 95% of the population values are between 350 and 400.
- For a given standard error, lower confidence levels produce wider confidence intervals.
- If you increase sample size, the width of confidence intervals will increase.
- If you take large random samples over and over again from the same population, and make 95% confidence intervals for the population average, about 95% of the intervals should contain the population average.

Exercise 10.2. A sample of 34 watermelons' have a mean weight of 64 ounces. Assume the population standard deviation is 12.7 ounces. Based on this, what is the maximal margin of error associated with a 90% confidence interval for the true population mean watermelon weight.

Exercise 10.3. If a school district takes a random sample of 70 Math SAT scores and finds that the average is 426, and knowing that the population standard deviation of Math SAT scores is intended to be 100. Find a 99% confidence interval for the mean math SAT score for this district.

Exercise 10.4. In a survey, 32 people were asked how much they spent on their child's last birthday gift. The results were roughly bell-shaped with a mean of \$39 and standard deviation of \$7. Find the margin of error at a 80% confidence level.

Exercise 10.5. A statistics student is curious about drinking habits of students at his college. He wants to estimate the mean number of alcoholic drinks consumed each week by students at his college. He plans to use a 90% confidence interval. He surveys a random sample of 71 students. The sample mean is 3.93 alcoholic drinks per week. The sample standard deviation is 3.78 drinks.

Exercise 10.6. Four hundred randomly selected working adults in a certain state, including those who worked at home, were asked the distance from their home to their workplace. The average distance was 8.84 miles with standard deviation 2.70 miles.

Construct a 98% confidence interval for the mean distance from home to work for all residents of this state.

Exercise 10.7. City planners wish to estimate the mean lifetime of the most commonly planted trees in urban settings. A sample of 16 recently felled trees yielded mean age 32.7 years with standard deviation 3.1 years. Assuming the lifetimes of all such trees are normally distributed, construct a 99.8% confidence interval for the mean lifetime of all such trees.

Exercise 10.8. Assuming the the population is normally distributed, find the 90% confidence interval for the population mean using the following sample.

45.8, 56.8, 65, 67.5, 30.4, 43.9, 59.7, 51.3

10.9 Lab: Confidence Intervals

10.9.1 Excel Functions for t-Distributions

Suppose a Student's t-distribution has the degree of freedom df = $n - 1$.

- Find a probability for a given t-value.
  - The area of the left tail of the t-value may be calculated by the function `T.DIST(t,df,true)`.
  - The area of the right tail of the t-value may be calculated by the function `T.DIST.RT(t,df)`.
  - The area of two tails of the t-value (here $t > 0$) may be calculated by function `T.DIST.2T(t,df)`.
- Find the critical value for a given probability p.
  - When the area of the left tail is given, the function `T.INV(p,df)` may be used.
  - When the area of both tails is given, the function `T.INV.2T(p,df)` may be used. This function is good for construction confidence interval.

10.9.2 Excel Functions for Marginal Errors

- If the population standard deviation σ is given and the sampling distribution is approximately normal, the marginal error can be obtained by the Excel function

```
CONFIDENCE.NORM(1-confidence level, population SD, sample size)
```

- If the population standard deviation σ is NOT given and the sampling distribution is approximately normal, the marginal error can be obtained by the Excel function, the marginal error can be obtained by the Excel function

```
CONFIDENCE.T(1-confidence level, sample SD, sample size)
```

Exercise 10.9. A sample of size 28 randomly selected to estimate a population mean with a confidence interval. The population is approximately normally distributed.

Find the critical value that corresponds to a confidence level of 80%.

Exercise 10.10. A sample of 22 backpacks' have a mean weight of 77 ounces. Assume the population standard deviation is 11.3 ounces. Based on this, what is the maximal margin of error associated with a 95% confidence interval for the true population mean backpack weight.

Exercise 10.11. In a survey, 21 people were asked how much they spent on their child's last birthday gift. The results were roughly bell-shaped with a mean of \$34 and standard deviation of \$6. Find the margin of error at a 98% confidence level.

Exercise 10.12. Assuming the the population is normally distributed. Find the 99.5% confidence interval for the population mean using the following sample.

97.8, 90.4, 76.9, 88.6, 97.1, 87.8, 88, 69.9, 75.3, 81.6

10.10 Confidence Interval for a Proportion

- Recall that the standard error of sample proportions is $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$, where n is the sample size and p is the population proportion. As a consequence, when estimating the population proportion p, we only have a point estimate $\hat{p}$ (phat) to use. For the standard error, we use the estimation
$$\sigma_{\hat{p}} \approx \hat{\sigma}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$
- Based on the central limit theorem, if p is not close to 0 or 1 and n is large enough, at the $100(1-\alpha)\%$ level, the margin of error for p is defined as
$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$
In Excel,

$z_{\alpha/2}$=NORM.S.INV((1 + confidence level)/2). [1]

The marginal error can also be obtained by

CONFIDENCE.NORM(1-confidence level, SQRT(phat*(1-phat)/n, n).

- The confidence interval for p is defined by
$$[\hat{p} - E, \hat{p} + E] = \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right],$$
where the critical value $z_{\alpha/2}$ satisfies that $P(Z < z_{\alpha/2}) = 1 - \alpha/2$ for the standard normal variable Z.
- In practical, the sample size n is considered large enough if $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$.
- The above defined confidence interval is known as the normal approximation (or Wald's) confidence interval. It is popular in introductory statistics books. However, it is unreliable when the sample size is small or the sample proportion is close to 0 or 1. Indeed, if the sample proportion is 0 or 1, the confidence interval defined here will have zero length.

Example 10.8. In a random sample of 100 students in college, 65 said that they come to college by bus.

(1) Give a point estimate of the proportion of all students who come to college by bus.

(2) Construct a 99% confidence interval for that proportion.

Example 10.9. Foothill College's athletic department wants to calculate the proportion of students who have attended a women's basketball game at the college. They use student email addresses, randomly choose 220 students, and email them. Of the 145 who responded, 22 had attended a women's basketball game.

Calculate and interpret the approximate 90% confidence interval for the proportion of all Foothill College students who have attended a women's basketball game.

10.11 Factors Affect the Width of Confidence Intervals

- The width of a confidence interval, equals twice the standard error, gives a measure of precision of the estimation.
- Recall, for population proportion and mean,

$$\text{Marginal Error} = \text{Critical Value} \cdot \frac{\text{(estimated) Population SD}}{\sqrt{\text{Sample Size}}}$$

- The formula tells us the precision of a confidence interval is affected by the confidence level, the variability, and the sample size.
  - Larger the confidence levels give larger critical values and errors.
  - Populations (and samples) with more variability gives larger errors.
  - Larger sample sizes give smaller errors.

10.12 Sample Size Determination

- In practice, we may desire a marginal error of E. With a fixed confidence level $100(1-\alpha)\%$, the larger the sample size the smaller the marginal error.
- When estimating population proportion, if we can produce a reasonable guess $\hat{p}$ for population proportion, then an appropriate minimum sample size for the study is determined by
$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \cdot \hat{p}(1-\hat{p}).$$
- When estimating population mean, if we can produce a reasonable guess σ for the population standard deviation, then an appropriate minimum sample size is given by
$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E}\right)^2.$$

Example 10.10. Suppose you want to estimate the proportion of students at QCC who live in Queens. By surveying your classmates, you find around 70% live in Queens. Use this as a guess to determine how many students would need to be included in a random sample if you wanted the error of margin for a 95% confidence interval to be less than or equal to 2%.

Example 10.11. Find the minimum sample size necessary to construct a 99% confidence interval for the population mean with a margin of error $E = 0.2$. Assume that the estimated population standard deviation is $\sigma = 1.3$.

Exercise 10.13. Out of 400 people sampled, 92 had kids. Based on this, construct a 90% confidence interval for the true population proportion of people with kids.

Exercise 10.14. To understand the reason for returned goods, the manager of a store examines the records on 40 products that were returned in the last year. Reasons were coded by 1 for "defective," 2 for "unsatisfactory," and 0 for all other reasons, with the results shown in the table.

0	0	0	0	2	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	2	0	2	0	0	0	0	0	0	2	0	0

(1) Give a point estimate of the proportion of all returns that are because of something wrong with the product, that is, either defective or performed unsatisfactorily.

(2) Construct an 80% confidence interval for the proportion of all returns that are because of something wrong with the product.

Exercise 10.15. You want to obtain a sample to estimate a population mean. Based on previous evidence, you believe the population standard deviation is approximately $\sigma = 41.5$. You would like to be 99% confident that your esimate is within 0.5 of the true population mean. How large of a sample size is required?

Exercise 10.16. Suppose a mobile phone company wants to determine the current percentage of customers aged 50+ who use text messaging on their cell phones. How many customers aged 50+ should the company survey in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of customers aged 50+ who use text messaging on their cell phones.

10.13 Lab: Confidence Interval for Proportion

- When a cumulative probability $p = P(Z < z)$ of a standard normal random variable Z is given, we can find z using `NORM.S.INV(p)`.
- If a sample of size n has the proportion $\hat{p}$ and the sampling distribution is approximately normal, the marginal error for the proportion can be obtained by the Excel function

```
CONFIDENCE.NORM(1-confidence level, SQRT(phat*(1-phat)), n)
```

Exercise 10.17. Suppose 250 randomly selected people are surveyed to determine if they own a tablet. Of the 250 surveyed, 98 reported owning a tablet. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of people who own tablets.

Exercise 10.18. A software engineer wishes to estimate, to within 5 seconds, the mean time that a new application takes to start up, with 95% confidence. Estimate the minimum size sample required if the standard deviation of start up times for similar software is 12 seconds.

Exercise 10.19. The administration at a college wishes to estimate, to within two percentage points, the proportion of all its entering freshmen who graduate within four years, with 90% confidence. Estimate the minimum size sample required.

11 Concepts of Hypothesis Testing

11.1 The Basic Idea of Hypothesis Testing

- The testing procedure starts with an initial assumption that the statement on population parameter is true.
- We test this initial assumption using a random sample. If the initial assumption is really the truth, then the test statistic from a random sample shouldn't be too far away from the center of the sampling distribution. Conversely, if the test statistic is too far away from the center, then we should not believe in the initial assumption.
- To determine how far is too far away, we need to specify a threshold, a prior probability, or equivalently a critical value.
- If the test statistic is at least extreme as the critical value, then the testing is significant enough to allow us to reject the initial assumption. Otherwise, we cannot draw a definite conclusion.
- The prior probability measures the chance that the initial assumption was wrongly rejected.

11.2 Two Hypotheses

- A statistical **hypothesis** is a statement about a population parameter.
- A **hypothesis test** is a process that uses sample statistics to test a **hypothesis**.
- To test a population parameter, we choose a pair of hypotheses, the null hypothesis and the alternative hypothesis which are contradictory to each other.
- The **null hypothesis**, denoted by H_0, is the statement about the population parameter that is assumed to be true.
- The **alternative hypothesis**, denoted H_a, is a statement about the population parameter that is contradictory to the null hypothesis.

Mathematical Symbols Used in H_0 and H_a

H_0	H_a
equal =	not equal ≠ or greater than > or less than <
greater than or equal to ≥	less than <
less than or equal to ≤	more than >

Example 11.1. Identify the Null and the Alternative Hypotheses

(1) Test a statement that the population mean is 1.

(2) Test a statement that the population mean is more than 3.

(3) Test a statement that the population mean is no more than 3.

11.3 The Logic of Hypothesis Testing

The logic of hypothesis testing and two types of error can be summarized in the following table.

ACTION	H_0 is Actually True	H_0 is Actually False
Do not reject H_0	Correct Outcome	Type II error
Reject H_0	Type I Error	Correct Outcome

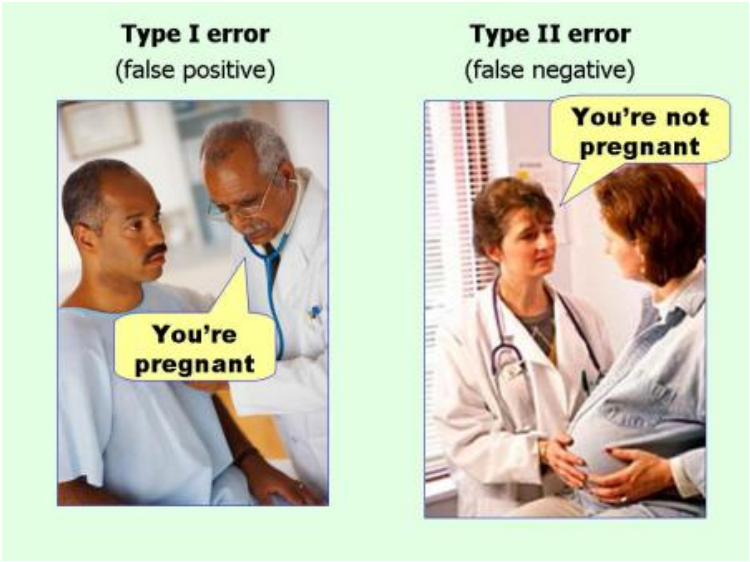
The interpretation of hypothesis testing is summarized in the following table.

Action	If the claim to be tested is in H_0	If the claim to be tested is in H_a
Reject H_0	There is enough evidence to reject the claim	There is enough evidence to support the claim
Fail to Reject H_0	There is not enough evidence to reject the claim	There is not enough evidence to support the claim

11.4 Type of Errors in Hypothesis Testing

- Rejecting the null hypothesis when it is indeed true is called a **type I error**. The maximum allowable probability of making a type I error is called the **level of significance**, denoted by α. In other words,
$$\alpha = P(\text{Type I error}) = P(\text{reject a true } H_0).$$
- Failing to reject the null hypothesis when the it is false is called a **type II error**. The probability of a type II error is usually denoted by β. The **power of a hypothesis test**, equals $1-\beta$, is the probability of rejecting the null hypothesis when it is false.

Example 11.2. Examples of Type I and Type II errors



In the above picture, bubble texts are in H_a

11.5 Type of Tests

- If H_a has the form $\mu \neq \mu_0$ the test is called a **two-tailed test**.
- If H_a has the form $\mu < \mu_0$ the test is called a **left-tailed test**.
- If H_a has the form $\mu > \mu_0$ the test is called a **right-tailed test**.
- Each of the last two forms is also called a **one-tailed test**.

11.6 Observed Significance

- To make a decision, one may also compare probabilities. The **observed significance** (***P*-value**) of a test statistic is the probability of obtaining a sample statistic at least as extreme as the (observed) test statistic, given that the null hypothesis were true.
- *P*-Value as Tail area

Sign in H_a	$\neq$	$<$
P-value	Double of the tail area	Left tail area

- Making decision by comparing the *P*-value with the significance level α:
  - reject H_0 if $p \leq \alpha$
  - fail to reject H_0 if $p > \alpha$.

Example 11.3. Given the following testing hypotheses

$H_0 : p = 0.50$ vs. $H_a : p \neq 0.50, n = 360, \hat{p} = 0.56,$

find the *P*-value for the test and make a decision at the 5% level of significance.

Example 11.4. It is believed that a stock price for a particular company will grow at a rate of \$5 per week with a standard deviation of \$1. An investor believes the stock won't grow as quickly. The changes in stock price is recorded for ten weeks and are as follows: \$4, \$3, \$2, \$3, \$1, \$7, \$2, \$1, \$1, \$2. State the null and alternative hypotheses, find the p-value, and identify the Type I and Type II errors.

11.7 Hypothesis Testing Procedure

(1) Check if the population distribution is approximately normal or sample size is large enough and determine if a Z-test or T-test can be performed. For proportion, Z-test may be used. For mean, if σ is known, the Z-test may be used. If σ is unknown, the T-test may be used.

(2) State the null and alternative hypothesis. The null hypothesis always contains the equal sign (and possibly together with a less than or greater than symbol, depending on H_a.)

(3) Set a significance level α. Commonly used levels are $\alpha = 0.01$, $\alpha = 0.05$ and $\alpha = 0.1$.

(4) Calculate the standardized test statistic: the Z-test statistic or the T-test statistic.

(5) Calculate the P-value according to the type of the test.

Sign in H_a	Type of Test
$\neq$	Two-tailed
$<$	Left-tailed
$>$	Right-tailed

(6) Make a test decision about the null hypothesis H_0. We reject H_0 if the P-value less than the significance level α.

(7) State an overall conclusion.

Example 11.5. Residences on a certain street claim that the mean speed of automobiles run through the street is greater than the speed limit of 25 miles per hour. A random sample of 100 automobiles has a mean speed of 26 miles per hour. Assume the population standard deviation is 4 miles per hour. Is there enough evidence to support the claim of the residences at the significance level $\alpha = 0.05$?

Example 11.6. A certain manufacturer claims that average numbers of candies in a certain sized bag that they produce is 20. To test the claims, you collected a random sample of 10 bags and find the mean is 18 and the standard deviation is 2.7. Assume the numbers of candies are normally distributed. At the significance level $\alpha = 0.05$, does your analysis support the manufacturer's claim?

Example 11.7. An instructor would like to know if the students enrolled in a math course in the current semester performed better than students in the last semester. The mean final exam from last semester is 75.5. The final exam scores of 40 randomly selected 40 students were obtained

93	88	69	74	76	81	78	77	74	63	67	81	80	82	68	88	76	69	75	78
75	77	94	87	74	88	63	75	94	88	91	77	76	68	80	88	68	83	72	72

Do the data provide evidence that the students in this semester performed significantly better on the final than last semester?

Example 11.8. Suppose you want to determine if a coin is fair. You toss the coin 50 times and observe 16 heads and 34 tails. At the significant level 0.01, do you think that the coin is fair? If not, does the coin favor the head or tail?

Example 11.9. Globally the long-term proportion of newborns who are male is 51.46%. A researcher believes that the proportion of boys at birth changes under severe economic conditions. To test this belief randomly selected birth records of 5,000 babies born during a period of economic recession were examined. It was found in the sample that 52.55% of the newborns were boys. Determine whether there is sufficient evidence, at the 10% level of significance, to support the researcher's belief.

Remark. In some books, the standard error of the sample distribution of sample proportions assuming that $p = p_0$ is calculated using the approximation

$$\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

An arguable explanation is that using the above value for SE will be consistent with the approach to a hypothesis testing using a confidence interval in the case that a two-tailed test is preformed.

Example 11.10. The mean work week for engineers in a start-up company is believed to be about 60 hours. A newly hired engineer hopes that it's shorter. She asks ten engineering friends in start-ups for the lengths of their mean work weeks. Based on the results that follow, should she count on the mean work week to be shorter than 60 hours?

Data (length of mean work week): 70; 45; 55; 60; 65; 55; 55; 60; 50; 55.

11.8 Practice

Exercise 11.1. Decide whether the following statements are true or false. Explain your reasoning.

- In case of a left-tailed test, we reject the null hypothesis if the sample statistic is significantly smaller than the hypothesized population parameter.
- A P-value of 0.08 is more evidence against the null hypothesis than a P-value of 0.04.
- The statement, "the P-value is 0.03", is equivalent to the statement, "there is a 3% probability that the null hypothesis is true".
- Even though you rejected the null hypothesis, it may still be true.
- Failing to reject null hypothesis means the null hypothesis is true.
- That the P-value of a sample statistic is $p = 0$ means the null hypothesis cannot be true.

Exercise 11.2. Determine both Type I and Type II errors for the following scenario:

Assume a null hypothesis, H_0, that states the percentage of adults with jobs is at least 88%. Identify the Type I and Type II errors from these four statements.

(1) Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%

(2) Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%.

(3) Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%.

(4) Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%.

Exercise 11.3. Determine if the following statements are true or false. Please explain your reasoning.

- The P-value of the test statistic is $p = 0.06$. At the significance level $\alpha = 0.01$, the null hypothesis H_0 should be rejected.
- A two-tailed test has larger probability of getting a type I error that a one-tailed test.
- That a test statistic falls in the rejection region means the P-value is smaller than the significance level.

Exercise 11.4. Suppose we're conducting a hypothesis testing for a population mean. Find the P-value for each of the following testing scenario with the given sample size n and the test statistics t.

(1) $H_0 : \mu = 25$ vs. $H_a : \mu < 25$, $n = 30$, $t = -2.43$.

(2) $H_0 : \mu = 35$ vs. $H_a : \mu > 35$, $n = 50$, $t = 2.13$.

(3) $H_0 : \mu = -7.9$ vs. $H_a : \mu \neq -7.9$, $n = 40$, $t = -1.99$.

Exercise 11.5. It's a Boy Genetics Labs claim their procedures improve the chances of a boy being born. The results for a test of a single population proportion are as follows:

- $H_0 : p = 0.50$, $H_a : p > 0.50$
- $\alpha = 0.01$
- $p -$ value $= 0.025$

Interpret the results and state a conclusion in simple, non-technical terms.

Exercise 11.6. A college football coach thought that his players could bench press a mean weight of 275 pounds. It is known that the standard deviation is 55 pounds. Three of his players thought that the mean weight was more than that amount. They asked 30 of their teammates for their estimated maximum lift on the bench press exercise. The mean of their maximum lift is 286.2.

Conduct a hypothesis test using a 2.5% level of significance to determine if the bench press mean is more than 275 pounds.

Exercise 11.7. In a college report, it says the mean age of students is 23.4 years old. An instructor thinks that the mean age is younger than 23.4. He randomly surveyed 50 students and found that the sample mean is 21.5 and the standard deviation is 1.9. At the significance level $\alpha = 0.025$, is there enough evidence to support the instructor's estimation?

Exercise 11.8. A teacher believes that 85% of students in the class will want to go on a field trip to the local zoo. She performs a hypothesis test to determine if the percentage is the same or different from 85%. The teacher samples 50 students and 39 reply that they would want to go to the zoo. For a 1% level of significance, would the data support the teacher's believe?

Exercise 11.9. The average number of days to complete recovery from a particular type of knee operation is 123.7 days. From his experience a physician suspects that use of a topical pain medication might be lengthening the recovery time. He randomly selects the records of seven knee surgery patients who used the topical medication. The times to total recovery were:

128, 135, 121, 142, 126, 151, 123

Assuming a normal distribution of recovery times, perform the relevant test of hypotheses at the 10% level of significance.

Would the decision be the same at the 5% level of significance?

11.9 Lab: Excel Functions for Normal Distributions

- Let Z be a standard normal random varaible. In Excel, $P(Z < z)$ is given by `NORM.S.DIST(z,TRUE)`.
- Let X be a normal random variable with mean μ and standard deviation σ, that is $X \sim \mathcal{N}(\mu, \sigma^2)$. In Excel, $P(X < x)$ is given by `NORM.DIST(x,mean,sd,TRUE)`.
- When a cumulative probability $p = P(X < x)$ of a normal random variable X is given, we can find x using `NORM.INV(p,mean,sd)`.
- When a cumulative probability $p = P(Z < z)$ of a standard normal random variable Z is given, we can find z using `NORM.S.INV(p)`.

11.10 Lab: Excel Functions for T-Distributions

Suppose a Student's T-distribution has the degree of freedom $\text{df} = n - 1$.

How to find a probability for a given T-value?

- The area of the left tail of the T-value may be calculated by the function `T.DIST(t,df,true)`.
- The area of the right tail of the T-value may be calculated by the function `T.DIST.RT(t, df)`.
- The area of two tails of the T-value ($t > 0$) may be calculated by function `T.DIST.2T(t,df)`.

Exercise 11.10. Joon believes that 50% of first-time brides in the United States are younger than their grooms. She performs a hypothesis test to determine if the percentage is the same or different from 50%. Joon samples 100 first-time brides and 53 reply that they are younger than their grooms. For the hypothesis test, find the p-value.

Exercise 11.11. The average McDonald's restaurant generates \$3.7 million in sales each year with a standard deviation of 0.7. Taylor wants to know if the average sales generated by McDonald's restaurants in Missouri is greater than the worldwide average. He surveys 24 restaurants in Missouri and finds the following data (in millions of dollars):

2.0	3.1	3.7	2.6	4.0	3.9	3.4	3.5	3.5	3.6	4.1	2.0
4.4	2.3	3.8	2.6	1.9	4.8	2.7	2.8	2.8	3.1	3.9	2.6

Find the p-value.

Exercise 11.12. The average number of days to complete recovery from a particular type of knee operation is 123.7 days. From his experience a physician suspects that use of a topical pain medication might be lengthening the recovery time. He randomly selects the records of seven knee surgery patients who used the topical medication. The times to total recovery were:

128, 135, 121, 142, 126, 151, 123

Assuming a normal distribution of recovery times, perform the relevant test of hypotheses at the 10% level of significance.

Would the decision be the same at the 5% level of significance?