

Ch 1: Stats Starts Here

1.1 What is (are) Statistics?

- **Statistics** (the discipline) is a way of reasoning, along with a collection of tools and methods, designed to help us understand the world with the use of data.

or

- **Statistics** (plural) are particular calculations made from data.

In this course, we will cover statistical theory, including the **collection**, the **organization** and the **interpretation** of data.

1.2 What are Data?

- Data are values with a context.
- They can be numbers, record names, or other labels.

Why we need to collect data?

- to find answers to the questions that cannot be answered otherwise
 - Ex: what is the long-term effect of a specific medication?

OR

- an experiment usually results in different outcomes (variability) and we want to study the reason(s) for the variability
 - Ex: different concentrations of chemical in water samples at a sewage treatment plant.

Statistical methods are used in:

- Personal life: what route should you take to school?
- Politics and Economics: What proportion of voters think the government is doing a good job?
- Biology: Does water quality based on water specimens?
- Medicine: Is a new migraine drug more effective than the old drug?
- Psychology: Would men or women tend to get jealous more easily?
- Social Science: Are there any changes in the proportion of women working outside their homes
- Engineering: What is the chance that a machine fails in warranty period?
- Dentistry: Does fluoride in salt increase the hardness of your teeth?
- Business: How does the stock market affect gas prices?

The process of making a data driven decision:

- a. State the question!
- b. Decide how to collect the data and which statistical methods have to be applied to answer the question or to make a decision.
- c. Collect the data = Draw a sample from the population of interest.
- d. Describe and summarize the data. Find the statistics.
- e. Use appropriate statistical tools, to find answer to question. Use formal data analysis and make statistical inference. In other words, generalize the finding from the sample to the entire population, if possible.

NOTE: In order to understand and apply inferential procedures, we need to understand Probability Theory.

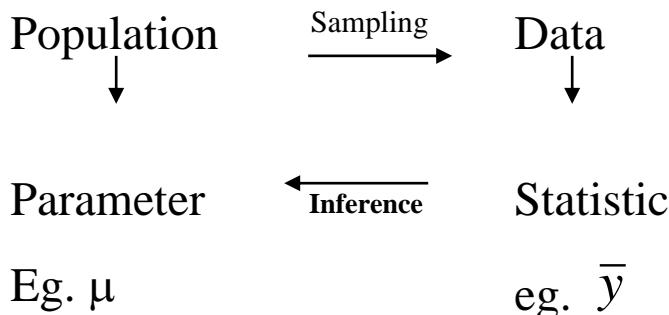
Goal of this class

- Learn appropriate methods for data collection
- Learn how to formulate, carry out, and interpret simple statistical analyses
- Learn how to follow basic statistical arguments
- Develop a critical attitude toward quantitative claims
- Develop statistical reasoning
- Summary: Become capable of making decisions based on data

Definition:

- The entire collection of individuals is called the *population of interest*.
- A *sample* is a subset of the population, selected for study in some prescribed manner.
- A **parameter** is a number that describes a characteristic of the population (often unknown)
- A **statistic** is a number that describes a characteristic of the sample (are known once the data are observed)
- We typically use Greek letters to denote parameters and Latin letters to denote statistics.

Name	Statistic	Parameter
Mean	\bar{y}	μ (mu, pronounced "meeoo," not "moo")
Standard deviation	s	σ (sigma)
Correlation	r	ρ (rho)
Regression coefficient	b	β (beta, pronounced "baytah' ,
Proportion	\hat{p}	p (pronounced "pee' ,



Example: I want to estimate the proportion of male students for this Stat 151 Lecture by randomly selecting 10 students in this class.

Population of Interest:

Sample:

Parameter:

Statistic:

Example:

An investigator wants to estimate the average height of Canadian females by measuring the height of 1000 randomly picked Canadian females.

Population of Interest:

Sample:

Parameter:

Statistic:

1.3 Variables

Recall: **Data** – can be numbers, record names, or other labels.

- Not all data represented by numbers are numerical data (eg. 1 = male, 2 = female)
- To provide context, the data need
 - o five “W’s”: **Who**, **What**, **When**, **Where**, and **Why** (if possible)
 - o One “H”: **How**

Who: Who are you interested in?

- The **Who** of the data tells us the individual cases about which (or whom) we have collected data
- **Subjects or participants** – people on whom we experiment
 - o The entire set of subjects is the **population**.
 - o The set of subjects you observe is your **sample**.
- **Respondents** – individuals who answer a survey
- **Experimental units** – animals, plants, and inanimate subjects

What: what characteristic (or variable) are you measuring?

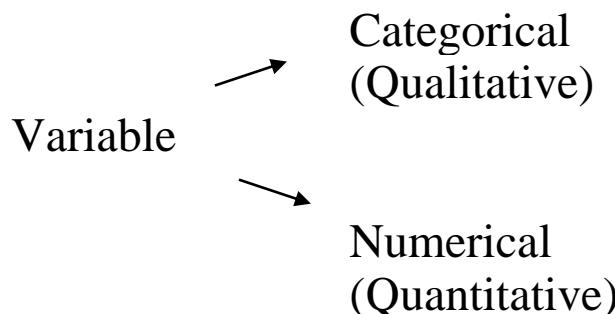
- **Variables** are characteristics recorded about each individual

- A variable can take different values for different individuals.
- Some variables have units that tell how each value has been measured and tell the scale of the measurement

The International System of Units links together all systems of weights and measures by international agreement. There are seven base units from which all other physical units are derived:

- | | |
|-----------------------|----------|
| • Distance | Meter |
| • Mass | Kilogram |
| • Time | Second |
| • Electric current | Ampere |
| • Temperature | Kelvin |
| • Amount of substance | Mole |
| • Intensity of light | Candela |

Types of Variables



Definition:

1. A **categorical** variable places a subject into one of several groups or categories (or levels).

- Usually we determine the counts of cases that fall into each category
- Two types:
 - i. **Nominal**: the levels have no order
 - ii. **Ordinal**: the levels have some order

Example:

- a) gender (M or F) →
- b) hair color (blonde, white, black, red, etc...) →
- c) nationality (American, Canadian, Chinese, French, German, Japanese, etc...) →
- d) letter grade (A+, A, A-, B+, B, B-, C+, C, C-, D+, D, F) →
- e) car manufacturer (Dodge, Ford, Honda, Others) →
- f) opinion (strongly agree, agree, neutral, disagree, strongly disagree) →
- g) education level (high school diploma, undergraduate degree, graduate degree) →

2. A ***quantitative*** variable measures a numerical quantity or amount in each subject.

➤ Two types:

- i. **Discrete**: can only take on distinct values
- ii. **Continuous**: can take on any value in a given interval

Example 2: A medical study.

Data from a medical study contain values of many variables of each of the people who were the subjects of the study. Which of the following variables are categorical and which are quantitative?

- a) Age (years)
- b) Smoker (yes or no)
- c) Systolic blood pressure (millimeters of mercury)
- d) Level of calcium in the blood (micrograms per milliliter)
- e) Drug effectiveness (1=strongly agree, 2=agree, 3=neutral, 4=disagree, 5=strongly disagree)

Example:

Each student has his or her own characteristics, eg. age, student ID, height, weight, gender, hair color, nationality, grade (A+, A, A-, ... D, F), number of siblings, etc...

Student	Age	Student ID	Height	Gender	Grade
Alice	19	00001	165	F	B+
Boris	20	00002	170	M	A-
Catherine	18	00003	158	F	A

- This data table clearly shows the context of the data.
- Specifically, it tells us the “What” (column titles) and “Who” (row titles) for these data.
- NOTE: student ID is known as an **identifier variable**. It is not a quantitative variable as it doesn’t have units. It is a categorical variable with one individual in each category.
The university assigns you a unique student ID, so that they can identify you on their system.
 - Example: SIN, ISBN

Why: Why are you doing this? What is the reason for collecting data?

- The questions of interest shape what we think about and how we treat the variable

NOTE: we need the Who, What, and Why to analyze data.

When and Where: gives us some nice info about the context

- Example: values recorded at the U of A in 1960s may mean something different than similar values recorded last year. (eg. average price of a pen that students use)
- Example: salary in Canada vs salary in 3rd world country

How: How are the data being collected? How is the variable being measured? (Ch 9-11)

- Data must be gathered properly.
- Improper data collection methodology may lead to wrong conclusions.
- Critically important for appropriate analysis and validity of inferences.
- Ex: results from voluntary internet surveys are often useless

Example:

One of the reasons that the Monitoring the Future (MTF) project was started was “to study changes in the beliefs, attitudes, and behavior of young people in the United States.” Data are collected from 8th, 10th, and 12th graders each year. To get a representative nationwide sample, surveys are given to a randomly selected group of students. In Spring 2004, students were asked about alcohol, illegal drug, and cigarette use. Describe the W’s, if the information is given. If the information is not given, state that it is not specified.

- **Who:** 8th, 10th, and 12th graders
- **What:** alcohol, illegal drug, and cigarette use
- **Why:** “to study changes in the beliefs, attitudes, and behavior of young people in the United States”
- **When:** Spring 2004
- **Where:** United States
- **How:** survey

Ch 2: Displaying and Describing Categorical Data

2.1 One Categorical Variable

It is very unlikely that we can draw conclusion about a variable simply by looking at raw data. Thus, it would be beneficial to summarize the raw data into a more manageable form in order to draw any useful conclusion. In this chapter, we will use graphs for initial data exploration.

NOTE: The proper choice of graph depends on the nature of the variable. *Different types of graphs for different types of data!*

Graphical Displays

Categorical Variables	Numerical Variables
<ul style="list-style-type: none">- Bar Chart- Pie Chart	<ul style="list-style-type: none">- Dot Plots- Stem Plots- Histograms- Time Plots- Box Plots- Scatterplots

Graphs for Categorical data

After categorical data has been sampled it should be summarized to provide the following information:

1. What values have been observed?

- Ex: Gender: Female or Male
- Ex: Car Color: red, white, blue, black, other
- Ex: Smoker: Yes or No

2. How often did every value occur?

The **distribution of a categorical variable** is given in form of a table providing with the following information:

- each possible category (and)
- frequency (or number) of individuals who fall into each category or
- relative frequency (or percentage) of individuals who fall into each category
- The *relative frequency* for a particular category is the percentage of the frequency that the category appears in the data set. It is calculated as

$$\text{relative frequency} = \frac{\text{frequency}}{\text{number of observations}}$$

Steps to Construct a Frequency Distribution Table

- 1) Define the levels of the variable (category)
- 2) Count the number of observations in the data set corresponding to each category (frequencies).
- 3) Summarize the results in a table (known as the *frequency distribution table*).
- 4) A relative frequency table is similar, but gives the percentages (instead of counts) for each category.

Example:

Construct a frequency and relative frequency distribution table of Favourite Ice-Cream Flavour for this Stat 151 Class:

Frequency Table

Category	Frequency
Chocolate	
Strawberry	
Vanilla	
Other	
Total	

Relative Frequency Table

Category	Relative Frequency
Chocolate	
Strawberry	
Vanilla	
Other	
Total	

Once the data is summarized in a frequency distribution table, the data can be displayed in a **bar chart** or **pie chart**. The bar chart will effectively show the frequencies or percent in the different categories, whereas the pie chart will show the relationship between the parts and the whole.

Bar chart

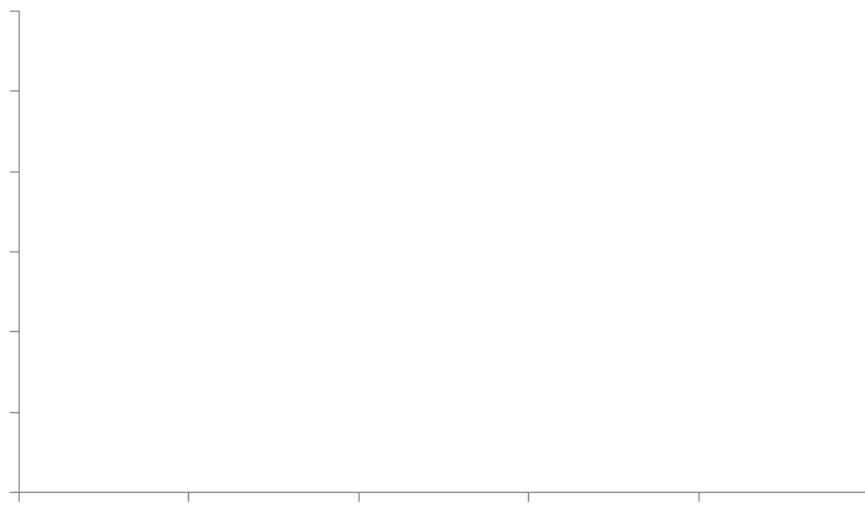
- a graph of the distribution of a categorical variable showing the counts for each category next to each other

Steps to make a bar chart:

- 1) Put every category (or levels of the categorical variable) evenly on the x -axis (can be marked with a tick).
- 2) Each category is represented by a bar of equal width, and the height of the bar is proportional to the corresponding frequency (relative frequency) of that category.
- 3) Label the y -axis (frequency or relative frequency).

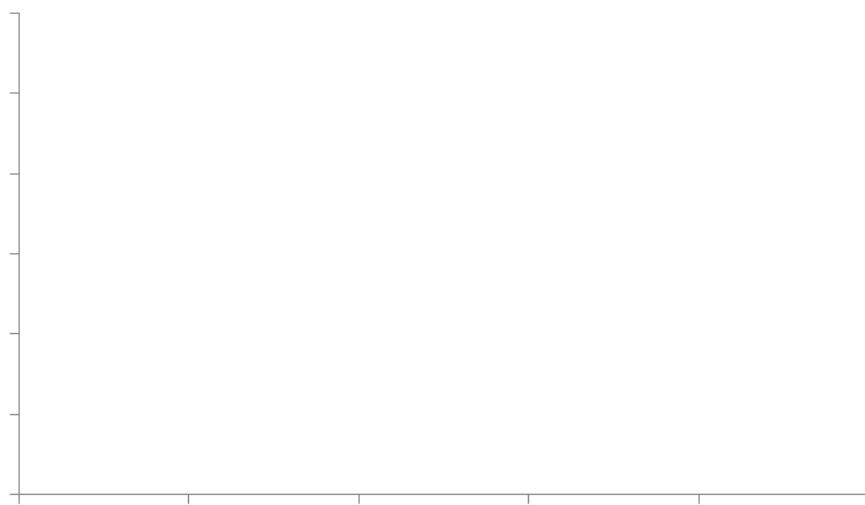
Example:

Construct a bar chart of the favourite ice-cream flavor for this Stat 151 class.



Example:

Construct a relative frequency bar chart of the favourite ice-cream flavor for this Stat 151 class.



NOTE: A relative frequency bar chart displays the relative proportion of counts for each category by replacing counts with percentages.

Pie charts

- provide an alternative kind of graph for categorical data.
- a circle is used to represent the sample.
- The size of the slice representing a particular category is proportional to the corresponding frequency (or relative frequency).

NOTE: Use a pie chart **only** when you want to emphasize each category's relation to the whole. It is useful when there are a relatively small number of classes involved.

Steps to create a pie chart:

- 1) Draw a circle
- 2) Calculate the slice size (angle)
$$\text{slice size} = \text{category relative frequency} \times 360^\circ$$

(fraction of the circle for the category)
- 3) use protractor to mark the angles

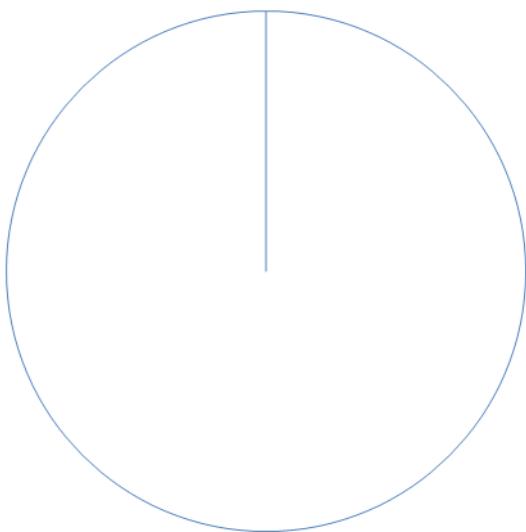
NOTE: In a pie chart, the proportions shown by each slice of the pie must add up to 100% and each individual must fall into only 1 category.

NOTE: Be sure to use enough individuals.

Example:

Using the data from previous example:

Category	Frequency	Relative Frequency (%)	Angle
Chocolate			
Strawberry			
Vanilla			
Other			



Example 5 (please read on your own):

On the M&M's webpage the following information on the distribution of colors in peanut M&M's is provided:

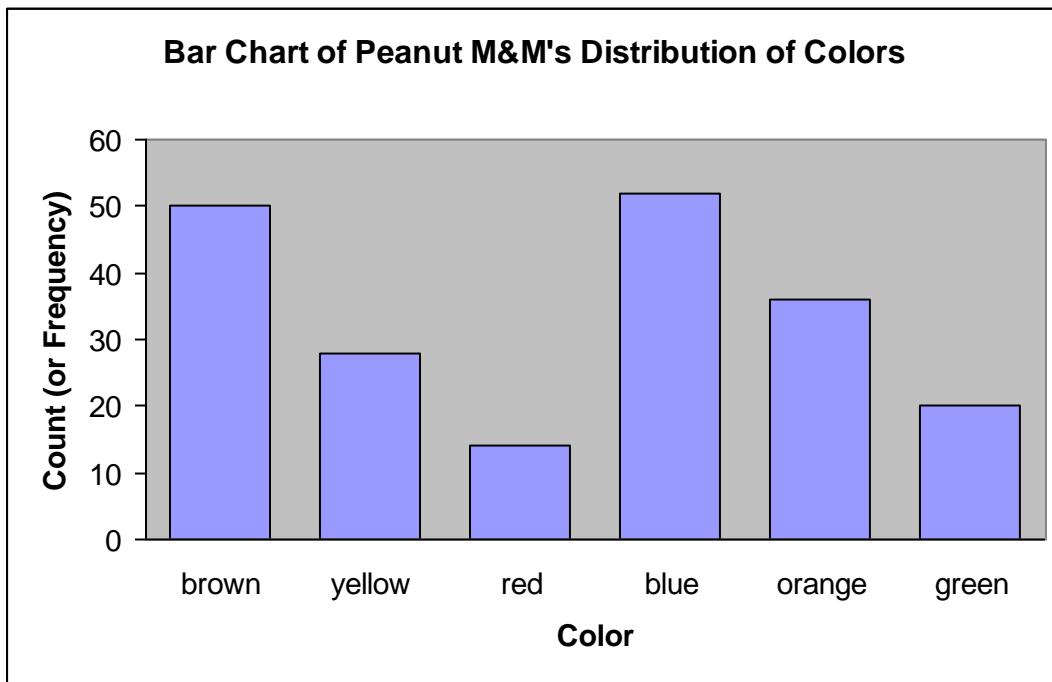
Color	brown	yellow	red	blue	orange	green
Percent	12%	15%	12%	23%	23%	15%

In order to check if this distribution is a "true" description of what is in a bag, someone bought a bag with 200 peanut M&M's and wants to describe the colors of the contents.

Color is a categorical variable, so a relative frequency table shall be obtained:

color	Count	Rel. Freq
brown	50	25%
yellow	28	14%
red	14	7%
blue	52	26%
orange	36	18%
green	20	10%
Total	200	100%

A bar chart would look like this:

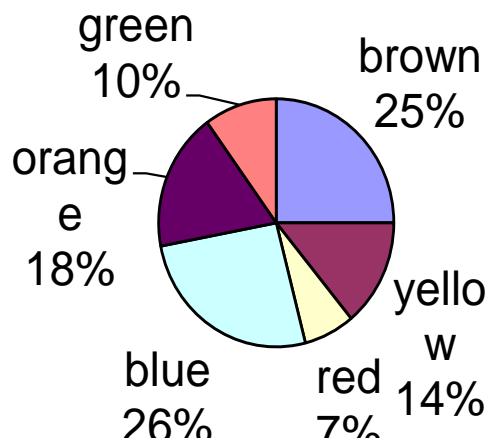


For the pie chart, the angles of the slices have to be determined:

color	Count	Rel. Freq	Angle
brown	50	25%	90.0°
yellow	28	14%	50.4°
red	14	7%	25.2°
blue	52	26%	93.6°
orange	36	18%	64.8°
green	20	10%	36.0°
Total	200	100%	360.0°

This results in the following pie chart:

Pie Chart of Peanut M&M's Distribution of Colors



2.2 Exploring the Relationship Between 2 Categorical Variables

Contingency Table

- allows us to explore the relationship between 2 categorical variables
- shows how individuals are distributed along each variable, contingent on the value of the other variable.
 - Ex: we can examine the gender and see whether a person likes chocolate ice-cream
- The margins of the table (both on the right and on the bottom) give totals and the frequency distributions for each of the variables.

- Each frequency distribution is called a **marginal distribution** of its respective variable.
- Each **cell** of the table gives the count for a combination of values of the two variables.
- A **conditional distribution** shows the distribution of one variable for just the individuals who satisfy some condition on another variable.
- The variables are considered **independent** when the distribution of one variable in a contingency table is the same for all categories of the other variable.

Heart Disease Example:

A study had been set up to study if smoking is a risk factor for heart disease. The result is given in the following table:

		Smoker		Total
		Yes	No	
Heart disease	Yes	23	15	38
	No	69	259*	328
Total		92	274	366

*This second cell in the nonsmoking column tells us that 259 nonsmokers had no heart disease.

The following is the marginal distribution of Heart Disease:

		Total	% of column
Heart disease	Yes	38	10.4%
	No	328	89.6%
Total		366	100%

The following is the marginal distribution of Smoker:

		Smoker		Total
		Yes	No	
Total		92	274	366
Row %		25.1%	74.9%	100%

The following is the conditional distribution of Heart Disease, conditional on smoking:

		Smoker	
		Yes	% of column
Heart disease	Yes	23	25%
	No	69	75%
Total		92	100%

The following is the conditional distribution of Heart Disease, conditional on nonsmoking:

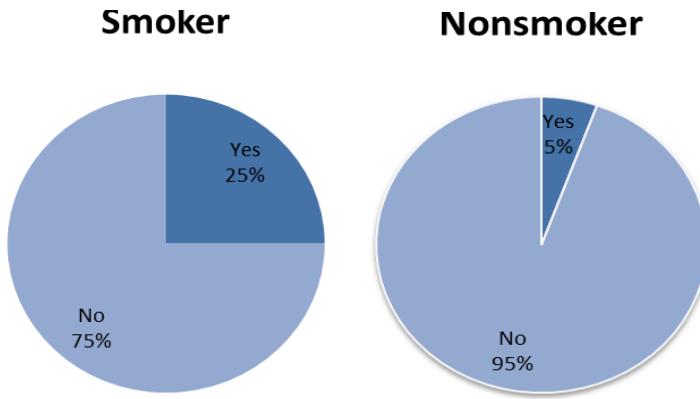
		Smoker	
		No	% of column
Heart disease	Yes	15	5.5%
	No	259	94.5%
Total		274	100%

NOTE: Do not confuse similar-sounding percentages:

- The percentage of people who have heart disease and smoked
- The percentage of smokers who have heart disease
- The percentage of people with heart disease who smoked

What does the conditional distributions tell us?

- there is a difference in having heart disease for those who smoked and those who don't.
- This is better shown with pie charts of the two distributions:



- This leads us to believe that heart disease and the status of smoking are associated, ie. *They are not independent.*

Independent variables

- Variables are said to be *independent* if the conditional distribution of one variable is the same for each category of another.
- In other words, there is *no association* between these variables.

Example of independent variables:

		Smoker		Total
		Yes	No	
Heart disease	Yes	30	15	45
	No	90	45	135
Total		120	60	180

The following is the conditional distribution of Heart Disease, conditional on the factor smoker:

		Smoker		Total
		Yes	No	
Heart disease	Yes	25%	25%	25%
	No	75%	75%	75%
Total		100%	100%	100%

- NOTE: We see that the distribution of having heart disease for the smokers is not different from that of the nonsmokers, so the two variables are independent.
- It is rare for 2 variables to be entirely independent.

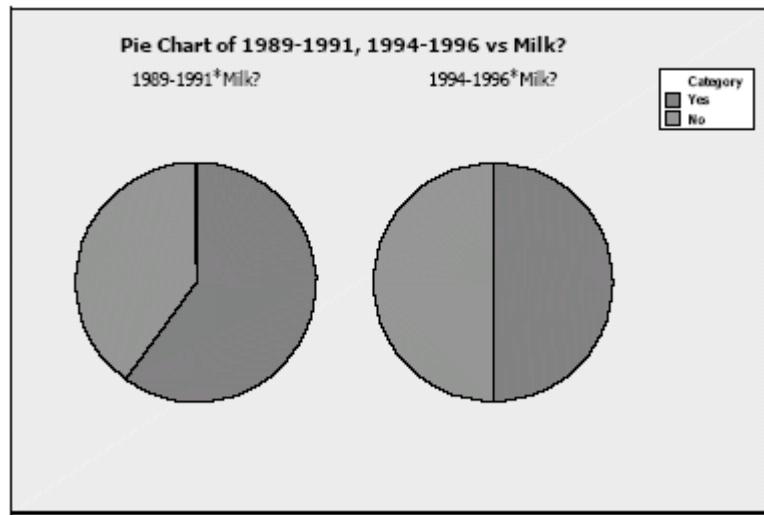
Example: Has the percentage of young girls drinking milk changed over time? The following table is consistent with the results from “Beverage Choices of Young Females: Changes and Impact on Nutrient Intakes” (by Shanthi A. Bowman)

		Nationwide Food Survey Years			
		1987-1988	1989-1991	1994-1996	Total
Drinks Fluid Milk	Yes	354	502	366	1222
	No	226	335	366	927
	Total	580	837	732	2149

1. Find the following:

- a. What percent of the young girls reported that they drink milk?
 - b. What percent of the young girls were in the 1989-1991 survey?
 - c. What percent of the young girls who reported that they drink milk were in the 1989-1991 survey?
 - d. What percent of the young girls in 1989-1991 reported that they drink milk?
2. What is the marginal distribution of milk consumption?
3. Do you think that milk consumption by young girls is independent of the nationwide survey year? Use statistics to justify your reasoning.

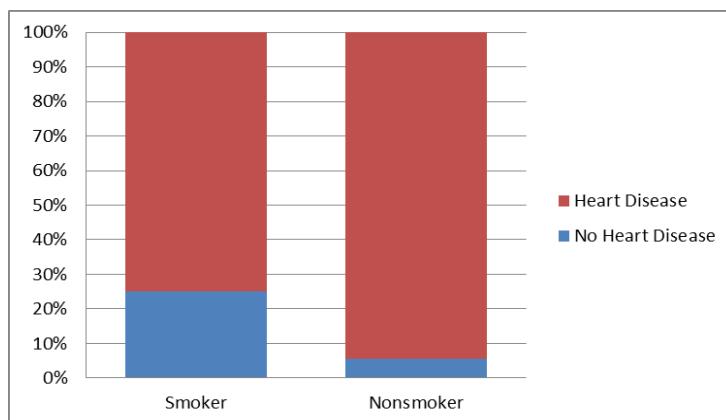
4. Consider the following pie charts for a subset of the data above:



Do the pie charts above indicate that milk consumption by young girls is independent of the nationwide survey year? Explain.

Segmented Bar Chart

- Displays the same info as a pie chart, but in the form of bars instead of circles.
- The following is a segmented bar chart for heart disease by smoking status:



Chapter 3: Displaying and Summarizing Quantitative Data

3.1 Displaying Quantitative Variables with Graphs

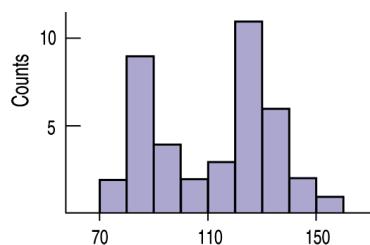
Graphs for Numerical Data

Numerical variables often take many values. We need to introduce other types of graphs to display the data for a quantitative variable in a fashion so that the distribution of the data becomes apparent.

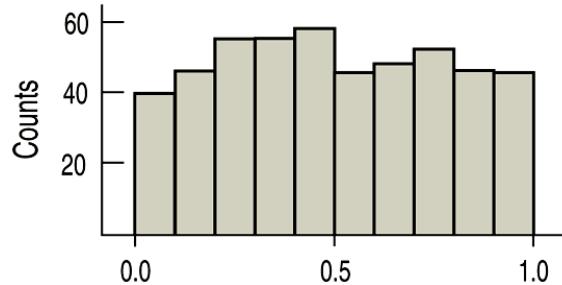
Describing a distribution of a plot:

1) shapes:

- a) nature of distribution (unimodal, bimodal, multimodal)
 - One characterization of general shape relates to the number of humps, or **modes**.
 - unimodal – a single peak
 - bimodal – two peaks; can occur when the data set consists of observation on two quite different kinds of individuals or objects



- multimodal – more than 2 peaks; rarely occurs
- uniform – no modes

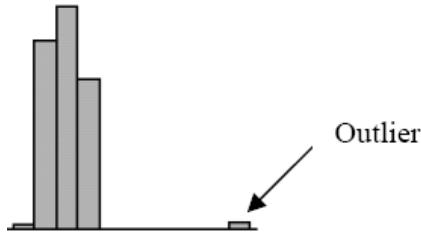


b) symmetrical or skewed to the right/left.

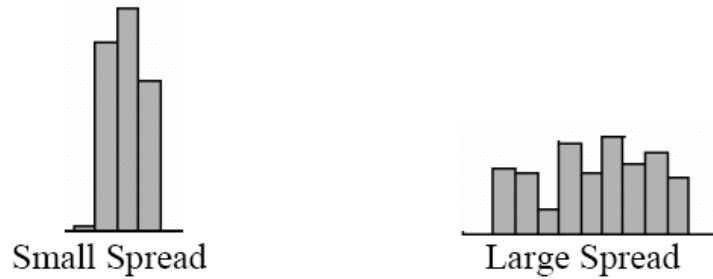
- **Symmetry:** if you can draw a vertical line so that the part to the left is a mirror image of the part to the right, then it is symmetric.
- Nonsymmetric graphs are **skewed**.
 - If the upper tail of the histogram stretches out farther than the lower tail, then is the histogram **positively skewed**, or **skewed to the right**.
 - If the lower tail longer than the upper tail the histogram is **negatively skewed**, or **skewed to the left**.

c) unusual values or deviations from the overall pattern.

- An important kind of deviation is an **outlier**, an individual value that falls outside the overall pattern.



- 2) center – the value that splits the data in half or a typical range of values at the center of the graph
 - mean, median, mode
- 3) spread – the range of values; concentration; are most of the values close to or far from the center?
 - Range, standard deviation, IQR



Dotplots

A **dot plot** is a plot that portrays the individual observations.

To construct a dot plot:

- 1) draw a horizontal (or vertical) line
- 2) label the line with the name of the variable, and mark regular values of the variable on it

3) for each observation, place a dot above (or next to) its value on the number line

NOTE1: The number of dots above a value on the number line represents the frequency of occurrence of that value.

NOTE2: The dot plots work well for small sets of data ($n \leq 50$).

Example 6a:

Construct a dotplot for the prices of 17 walking shoes (in \$): 90 70 70 70 75 70 65 68 60 74 70 95 75 68 85 40 65



Stem-and-Leaf Displays/Stemplot

Another way to portray the individual observations of quantitative data is a *stem-and-leaf display*, which works well for **small** sets of data ($n \leq 50$).

Each observed number is broken into two pieces called the *stem* and the *leaf*.

How to make a stemplot:

1. Order the data from smallest to largest.
2. Divide each data value into two parts:
 - The leading digits of the number are the **stems**.
 - The rest of the digits of the number are the **leaves**.
 - NOTE: use the stems to label the bins (the equal-width interval)
 - NOTE 2: use only one digit for each leaf – either round or truncate the data values to one decimal place after the stem.
3. List the stems in a column (with the smallest at the top), and place a vertical line to the right of this column.
4. For each measurement, record the leaf portion in the same row as its corresponding stem.
5. Provide a **key** to your stem and leaf coding.

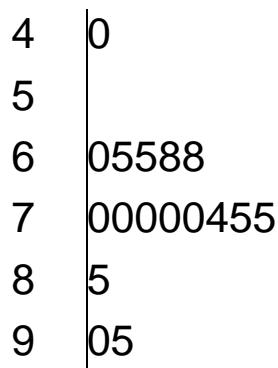
Example 6b: Draw a stemplot for the prices of walking shoes.

Prices of walking shoes in \$:

90 70 70 70 75 70 65 68 60 74 70 95 75 68 85 40 65

First, order them from smallest to largest:

40 60 65 65 68 68 70 70 70 70 70 74 75 75 85 90 95



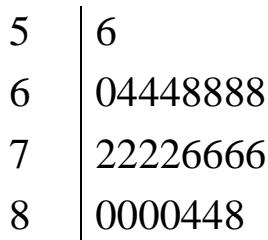
Prices of Walking Shoes

Key: 9|0 means \$90

Example 7a: Draw a stemplot for the prices of running shoes.

Prices of running shoes:

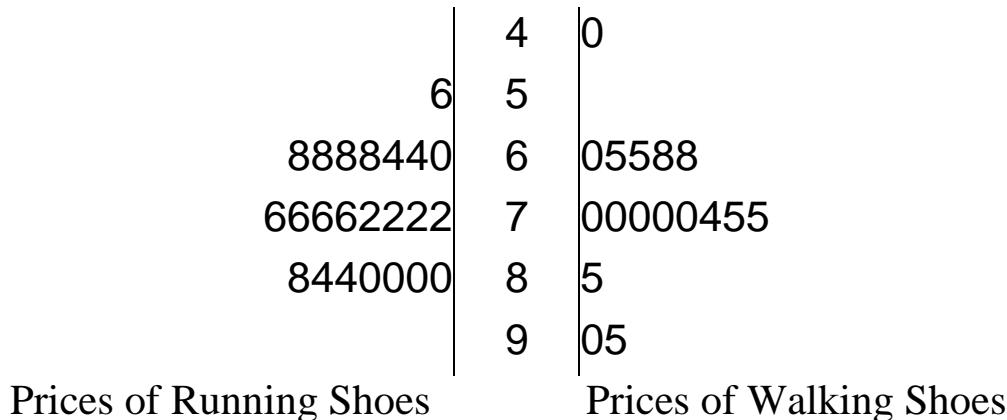
56, 60, 64, 64, 64, 68, 68, 68, 68, 72, 72, 72, 72, 76, 76, 76, 76, 80,
80, 80, 80, 84, 84, 88



Prices of Running Shoes

Key: 8|8 means \$88

You also can use stem-and-leaf plots for the comparison of the distribution of *two* groups (**back-to-back stemplot**)



Histogram

The most common graph for describing numerical data is the histogram. It helps to visualize the distribution of the underlying variable very well, especially for large data sets.

A **histogram** for a quantitative variable is a graph that uses bars to show "how often" (measured as frequency or relative frequency) measurements falls in a particular equal-width interval (or bin).

How to construct a histogram:

1. Decide which *intervals* of **equal** length to use for the histogram.

The intervals should have the same width and the boundaries are if possible whole numbers or tenth. The resulting intervals are called *bins*.

Informal Rule:

- 6 to 10 bins for smaller data sets
 - 10 to 25 bins for large ones
2. Create a frequency table for the class intervals using the *method of left inclusion*.
 - a) Example: Where should we include 28.0, which falls on the boundary between the classes [27.5, 28.0) and [28.0, 28.5)?
 3. Mark the boundaries of the class intervals on a horizontal axis.
 4. Use the frequency or the relative frequency on the vertical axis.
 5. Draw a bar for each class interval, with heights according to the frequency or relative frequency of the corresponding interval.

NOTE: data values are retained with the stem-and-leaf plot and dot plot but not with the histogram.

NOTE: bar charts of categorical variables had spaces between the bars to separate the counts of different categories. But in a histogram, the bins slice up all the values of the quantitative variable, so any spaces in a histogram are actual gaps in the data, indicating a region where there are no values.

Example: Construct a histogram for prices of walking shoes with bin width of 10.

Prices of walking shoes in \$:

40 60 65 65 68 68 70 70 70 70 74 75 75 85 90 95

class intervals	Frequency	Rel Freq	
[40; 50)			
[50; 60)			
[60; 70)			
[70; 80)			
[80; 90)			
[90; 100)			
Total			

We can even obtain more info from the relative frequency table and histogram:

1. What is the proportion of walking shoe prices that fall on or above \$70?
2. What is the percent of walking shoe prices that fall below \$70?

Example: Which of the following variables regarding student info requires a histogram?

- a) Name
- b) Student ID
- c) Age
- d) Birthplace
- e) Nationality

3.2 Describing Quantitative Variables with Numbers

Numerical Summaries

When you have large data sets, you may want to reduce them into a few important and meaningful numbers that preserves the relevant features of the data set so that you can draw useful conclusions.

Notations:

y = the variable for which we have sample data (ie. the variable of interest; we use capital letter for variables)

n = sample size = number of observations of the variable y

y_1 = the first sample observation of the variable y

y_2 = the second sample observation of the variable y

:

y_n = the n^{th} sample observation of the variable y

Example 1a:

We have a sample of $n = 4$ observations on $y = \text{battery lifetime (hrs)}$:

$$y_1 = 5.9$$

$$y_2 = 7.3$$

$$y_3 = 6.6$$

$$y_4 = 5.7$$

The sum of y_1, y_2, \dots, y_n can be denoted by $y_1 + y_2 + \dots + y_n$, but this becomes cumbersome when n is large.

The Greek letter Σ is traditionally used in mathematics to denote summation. In particular,

$$\sum_{i=1}^n y_i = y_1 + y_2 + \dots + y_n$$

Example 1b:

Using data in Ex 1a, find $\sum_{i=1}^4 y_i$.

Describing the center of a data set

The center of a data set can be described as a typical or representative value. It is a value that can be used as a benchmark for all other values.

The three most common measures for center are:

- a) The mean (center)
- b) The median
- c) The mode

a) The Mean

The **mean** of a set of numerical observations is the familiar arithmetic average.

Suppose you have a sample of n observations y_1, y_2, \dots, y_n . The *mean* of these values, \bar{y} , is

$$\bar{y} = \frac{\text{sum of all observations}}{\text{number of observations}} = \frac{y_1 + y_2 + \dots + y_n}{n} = \frac{\sum_{i=1}^n y_i}{n}$$

Example 1c:

The mean battery life of this sample is

Example 1d: Suppose you have a fifth observation for battery lifetime. What is the battery lifetime you will need in your fifth observation in order to have an average battery lifetime of at least 6.5 hours?

Calculating the Mean from a Frequency Distribution:

Let y_1, \dots, y_k be the distinct observations and f_1, \dots, f_k be the frequencies of these observations, respectively. The sample mean is:

$$\bar{y} = \frac{\sum_{i=1}^k y_i f_i}{\sum_{i=1}^k f_i}, \text{ where } \sum_{i=1}^k f_i = n$$

Example:

Suppose the class has 20 students, and their final grades are:

Grades	Frequency
50%	18
52% (You)	1
100%	1

Important note:

If a set of observations includes an outlier (an unusual value), the mean will be drawn into the direction of this outlier, as a result, it will not describe the true center of the distribution.

NOTE1: The mean is *not* resistant to outliers.

NOTE2: For skewed data sets, the mean may not represent the “center” of the data set very well.

For this reason, we are looking for an alternative value to measure the center of a distribution → **median**.

b) Median

The *median*, M , is the value that divides the ordered sample in two sets of the same size; one half of the data lies below M , and the other half above M .

Steps to find the median:

- 1) The median is determined by first ordering the n observations from smallest to largest.
- 2) Then

$$M = \text{median}$$

$$= \begin{cases} \text{the single middle value} & \text{if } n \text{ is odd} \\ \text{the average of the middle 2 values} & \text{if } n \text{ is even} \end{cases}$$

Example 4a:

Suppose you have the following ordered sample of size 5:

2 5 6 7 9

Example 4b:

Suppose you have the following ordered sample of size 6:

2 5 6 7 9 11

Example 5:

Consider the class of 20 students again, you will discover that the median grade in the class is _____.

c) Mode

The **mode** is the value that occurs with the highest frequency in a data set.

Example 6a: Using Previous Data of Prices of walking shoes in \$:

90 70 70 70 75 70 65 68 60 74 70 95 75 68 85 40 65

Find the mode.

Example 6b: Prices of running shoes in \$:

49 59 65 79 80 95 105 69 69 60 77 95 78 85 89

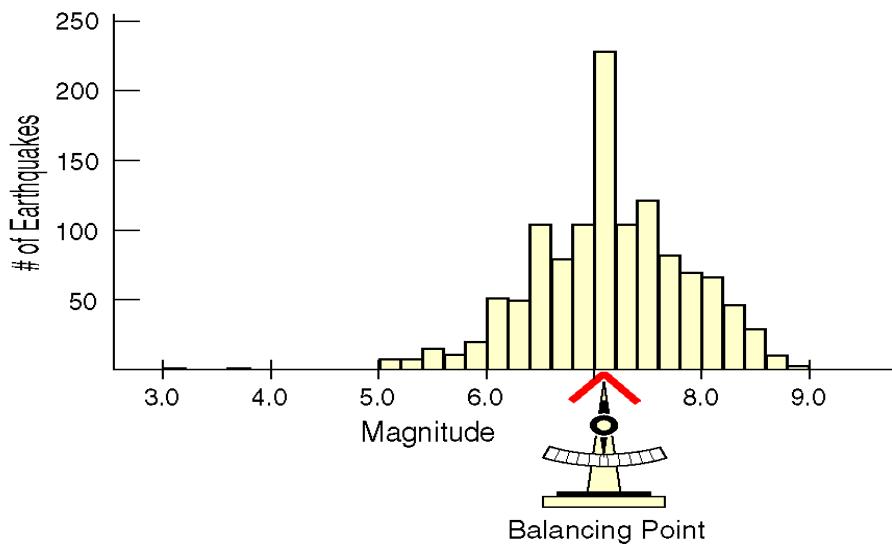
Find the mode.

Example 6c: Consider the battery lifetime example again. Find the mode.

NOTE: Mode is simple to locate, but it has a major shortcoming in that a data set may have none or may have more than 1 mode, whereas it will have only one mean and only one median.

Comparing mean and median.

The mean is the balance point of the distribution.



The median is the point where the distribution is cut into two parts of the same area.

In a symmetric distribution, $\text{mean} = \text{median} = \text{mode}$.

In a positively (right) skewed distribution, mean > median > mode.

In a negatively (left) skewed distribution, mean < median < mode.

Conclusion: When a distributions or data set is symmetric, the mean is a better representation to describe the center. When a distributions or data set is skewed, the median is better. Why?

The median is resistant as it only takes into account the rank of each observation not its magnitude.

Describing the variability (spread) of a distribution

It is not enough just to report a number that describes the center of a sample. The spread or variability in a sample is also an important characteristic of a sample.

The three most common measures for variability are:

- a) Range
- b) Variance and Standard deviation
- c) Interquartile Range (IQR)

a) Range

The **range** of a sample is the simplest numerical measure of variability that gives the difference between the largest (maximum) and the smallest (minimum) value in the sample.

$$\text{Range} = \text{max} - \text{min}$$

Example: Find the range for each of the sample in Ex 4.

4a) Range =

4b) Range =

Rule: Usually the greater the range the larger the variability.

However, variability depends on more than just the distance between the two most extreme values. It is a characteristic of the whole data set and every observation contributes to it.

Sample 1: * * * * * o * * * *

Sample 2: * ****o**** *

b) the variance and standard deviation

The value $y_i - \bar{y}$ is the deviation of the observation y_i from the mean \bar{y} . In a sample with n observations, we will get n deviations from the sample mean

$$(y_1 - \bar{y}), (y_2 - \bar{y}), \dots, (y_n - \bar{y})$$

NOTE:

- 1) A specific deviation is *positive* if the value is greater than \bar{y} and *negative* if it is less than \bar{y} .

2) The set of deviations describes the variability of the data set,

but it is always *true* that $\sum (y_i - \bar{y}) = 0$, so we need to introduce some easy calculation techniques to the deviations in order to characterize the variability in the data set $\rightarrow \textit{Squaring}$ every deviation before summing them up, ie.

$$(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2 = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2$$

Definition:

The *variance*, denoted by s^2 , is the sum of squared deviations from the mean divided by $n - 1$. That is,

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

NOTE: variance is problematic as a measure of spread as it is measure in squared units!

The *standard deviation*, s , is the square root of the variance:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^n y_i^2 - n\bar{y}^2}{n-1}}$$

NOTE:

- standard deviation is the most used measure of variability
- The value of the standard deviation tells how closely the values of a data set are clustered around the mean.
- It is measured in the same units as the original data

Example:

Calculate the variance and standard deviation of the 4 battery lives.

Recall: $\bar{y} = 6.375$

Properties of the standard deviation:

- The standard deviation s measures the spread about the mean and should be used *only* when the mean is chosen as the measure of center.
- $s = 0$ only when there is *no* spread. This happens only when all observations have the same value.
- Otherwise, $s > 0$. The standard deviation increases as the observations become more spread out.
- Like the mean, the standard deviation is not resistant to outliers, (a few outliers can make s very large), so we are in need of a value that measures the spread and is resistant to outliers.

c) Interquartile Range (IQR)

An alternative measure of variability is the *interquartile range*, which is the range of the middle half of the data. The IQR, like the median, is resistant to outliers. It is based on quartiles that divide the data into 4 equal sections.

Percentiles

The p^{th} *percentile* is the value so that $p\%$ of the measurements fall below the p^{th} percentile and $(100-p)\%$ fall above it.

NOTE: the set of measurements on the variable x must be arranged in order of magnitude.

NOTE: p is a number between 0 and 100.

NOTE: The median is the 50^{th} percentile.

Quartiles

- 1) The *lower quartile* Q_1 is the 25^{th} percentile (that is, it separates the bottom 25% of the measurements from the top 75%)
- 2) The *upper quartile* Q_3 is the 75^{th} percentile (that is, it separates the top 25% of the measurements from the bottom 75%)

Therefore, the middle 50% of the measurements fall between Q_1 and Q_3 .

How to calculate the quartiles?

- Order the measurements in order of magnitude.
- Q_1 is the median of those measurements that fall below the overall median.
- Q_3 is the median of those measurements that fall above the overall median.
- NOTE: the median of the entire sample is **not** included in both halves.

Example 4a:

Using previous examples, find Q_1 and Q_3 :

Suppose you have the following ordered sample of size 5:

2 5 6 7 9

Example 4b:

Suppose you have the following ordered sample of size 6:

2 5 6 7 9 11

Definition:

The **interquartile range (IQR)** for a set of measurements is the difference between the upper and the lower quartile:

$$\text{IQR} = Q_3 - Q_1.$$

- Ignore extreme data values and concentrate on the middle of the data
- When the data values are tightly clustered around the center of the distribution, the IQR will be small.
- When the data values are scattered far from the center, the IQR will be large.

Using previous examples, find the interquartile range.

Example 4a:

Recall: $Q_1 = 3.5$ and $Q_3 = 8$

Thus, IQR = $Q_3 - Q_1 =$

Example 4b:

Recall: $Q_1 = 5$ and $Q_3 = 9$

Thus, IQR = $Q_3 - Q_1 =$

The Five-Number-Summary and Boxplots

The **five-number-summary** of a set of measurements consists of the minimum, the first quartile, the median, the third quartile, and the maximum. These numbers give a good summary of a distribution of quantitative observations.

In symbols, the 5-number summary are:

Minimum	Q_1	M	Q_3	Maximum
---------	-------	-----	-------	---------

Example 4a:

You have a data set: 2 5 6 7 9

Recall: $M = 6$, $Q_1 = 3.5$ and $Q_3 = 8$

Therefore, the 5-number summary: 2 3.5 6 8 9

Example 4b:

You have a data set: 2 5 6 7 9 11

Recall: $M = 6.5$, $Q_1 = 5$ and $Q_3 = 9$

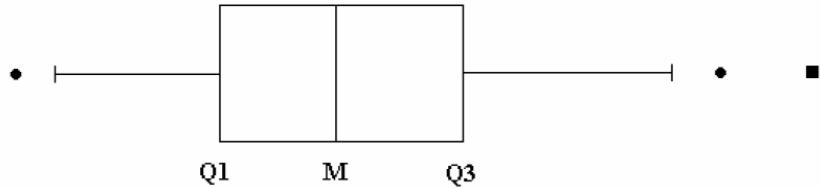
Therefore, the 5-number summary: 2 5 6.5 9 11

Conclusion: Choosing a summary

- Before choosing a summary, always draw a graph and learn about the distribution of the data.
- Use the mean and the standard deviation for reasonably symmetric distributions.
- Use the median and IQR (or the five number summary) for describing skewed distributions.
- If there are multiple modes, try to understand why. If you identify a reason for the separate modes, it may be good to split the data into two groups.
- If there are any clear outliers and you are reporting the mean and standard deviation, report them with the outliers present and with the outliers removed. The differences may be quite revealing.

Boxplot

The five number summary leads to another visual representation of a data set called a *boxplot*.



Definition:

The **boxplot** is a powerful graphical tool for summarizing data. It shows the center, the spread, and the symmetry or the skewness at the same time. Boxplots are particularly useful when comparing groups.

Construction of a boxplot

1. Draw a vertical measurement scale spanning the range of the data.
2. Draw a rectangular box, whose lower edge is at the lower quartile and whose upper edge is at the upper quartile.
3. Draw a line segment inside the box at the location of the median.
4. Determine the “fences” in order to check for outliers.

An observation is an **outlier** if it falls more than 1.5 IQR above the third quartile or below the first quartile.

In order to determine outliers, calculate an upper and a lower fence:

- upper fence = $Q_3 + 1.5 \times \text{IQR}$
 - every measurement **above** the upper fence is an outlier.

- lower fence = $Q_1 - 1.5 \times \text{IQR}$
 - every measurement **beneath** the lower fence is an outlier.

5. Add line segments (whiskers) from each end of the box to the most extreme data values found within the fences. If a data value falls outside one of the fences, we **do not** connect it with a whisker.

6. Add the outliers by displaying any data values beyond the fences with special symbols (such as circles).

- We often use a different symbol for “*far outliers*” that are farther than 3 IQRs from the quartiles.

Example 7:

Using Previous Data:

Prices of walking shoes in \$:

40 60 65 65 68 68 70 70 70 70 74 75 75 85 90 95

Find the outliers of this data set and make a boxplot.

Interpreting Boxplots:

Symmetric distribution: median line in center of box and whiskers of equal length

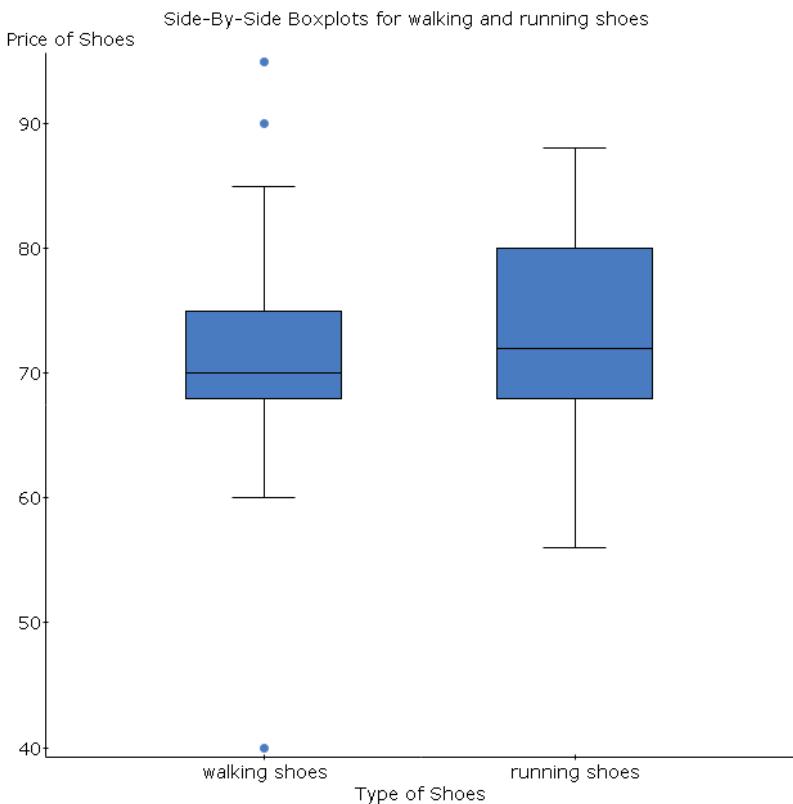
Skewed right: median line left of center and long right whisker

Skewed left: median line right of center and long left whisker

One can create *comparative boxplots* by drawing several boxes in one graph. This is a good tool for comparing variables in different categories.

Eg: Resting pulse and pulse after exercise boxplots in one graph.

Eg: Prices of walking shoes and prices of running shoes



Refer to the comparative boxplot, which of the following is TRUE?

- a) the walking shoes tend to be more expensive than the running shoes.
- b) The price distribution for either type of shoe is left-skewed.
- c) **More than 25% of running shoes cost less than \$70.**
- d) Majority of the running shoes cost less than \$70.
- e) Less than 25% of walking shoes cost more than \$70.

Time plots

For some data sets, we are interested in how the data behave over time. In these cases, we construct time plot of the data.

A *time plot* of a variable plots each observation against the time at which it was measured.

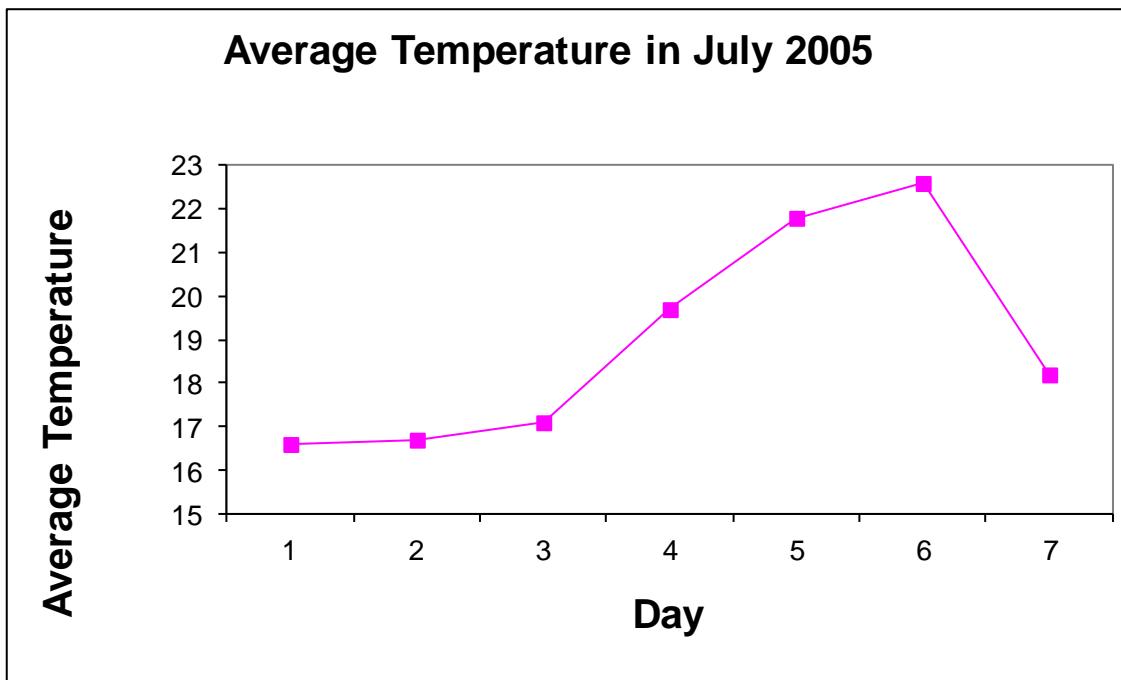
- Time is always used on the horizontal axis.

Example 9:

Average temperature in Edmonton in first week of July 2005 (data from Environment Canada)

Day	1	2	3	4	5	6	7
temperature	16.6	16.7	17.1	19.7	21.8	22.6	18.2

Since the variable (temperature) is measured at intervals over time (of one day), it's appropriate to apply the time plot.



Ch 5 The Standard Deviation as a Ruler and the Normal Model

5.2 Shifting and Scaling

Shifting data:

- Adding (or subtracting) a *constant* to every data value adds (or subtracts) the same constant to measures of position.
- Adding (or subtracting) a *constant* to each value will increase (or decrease) measures of position: center, percentiles, max or min by the same constant.
- Its shape and spread - range, IQR, standard deviation - remain unchanged.

Example: You have a data set: $y_1 = 1, y_2 = 2, y_3 = 3, y_4 = 4, y_5 = 5$. If you want to add a constant $c = 2$ to each observation in this data set, how does it affect the mean, median, Q_1 , Q_3 , max, min, range, standard deviation, IQR, and its shape?

Summary statistics:

Column	n	Mean	Variance	Std. Dev.	Std. Err.	Median	Range	Min	Max	Q1	Q3
y	5	3	2.5	1.5811388	0.70710677	3	4	1	5	2	4
y + 2	5	5	2.5	1.5811388	0.70710677	5	4	3	7	4	6

Summary:

If $y_{new} = y_{original} + c$ for each observation

- For measures of center or position:

- $Center_{new} = Center_{original} + c$
 - $Position_{new} = Position_{original} + c$
- For measures of spread and shape:
 - $Spread_{new} = Spread_{original}$
 - $Shape_{new} = Shape_{original}$

Rescaling Data:

- When we multiply (or divide) all the data values by any constant, all measures of position (such as the mean, median, and percentiles) and measures of spread (such as the range, the IQR, and the standard deviation) are multiplied (or divided) by that same constant.

Example: You have a data set: $y_1 = 1, y_2 = 2, y_3 = 3, y_4 = 4, y_5 = 5$. If you want to multiply each observation with a constant $d =$

2 in this data set, how does it affect the mean, median, Q_1 , Q_3 , max, min, range, standard deviation, and IQR?

Summary statistics:

Column	n	Mean	Variance	Std. Dev.	Std. Err.	Median	Range	Min	Max	Q1	Q3
y	5	3	2.5	1.5811388	0.70710677	3	4	1	5	2	4
y * 2	5	6	10	3.1622777	1.4142135	6	8	2	10	4	8

Summary:

If $y_{new} = d \times y_{original}$ for each observation

- For measures of center or position:
 - o $Center_{new} = d \times Center_{original}$
 - o $Position_{new} = d \times Position_{original}$
- For measures of spread and shape:
 - o $Spread_{new} = d \times Spread_{original}$
 - o $Shape_{new} = Shape_{original}$

Thus, if you rescale and shift data: $y_{new} = d \times y_{original} + c$ for each observation

- For measures of center or position:

- $Center_{new} = d \times Center_{original} + c$
 - $Position_{new} = d \times Position_{original} + c$

- For measures of spread and shape:

- $Spread_{new} = d \times Spread_{original}$
 - $Shape_{new} = Shape_{original}$

Example: Students taking an intro stats class reported the number of credit hours that they were taking that quarter.

Summary statistics are shown in the table.

Mean	Std Dev	Min	Q1	Median	Q3	Max
16.65	2.96	5	15	16	19	28

Suppose the college charges \$73 per credit hour plus a flat fee of \$35 per quarter. For example, a student taking 12 credit hours would pay $\$35 + 12(\$73) = \$911$ for that quarter.

What is the mean fee paid?

What is the standard deviation for the fees paid?

What is the median fee paid?

The Standard Deviation as a Ruler

The distance to the mean of a specific observation measured in standard deviations gives information about the location of this observation in relation to the other observations in the sample.

If you obtained a score in an achievement test you might want to know your standing in relation to other people who have taken the test. Are you below or above the mean, how far above or below in relation to the other people.

5.1 Standardizing with z-score

z-score or *standardized value* is a measure of relative standing.

If y is an observation from a sample with mean \bar{y} and standard deviation s , then the *standardized value* or *z-score* of y is

$$z = \frac{y - \bar{y}}{s}$$

- tells “how many standard deviations away from the mean does the measurement lie and in which direction?”
 - The standardization makes numbers from different contexts comparable.
 - Positive z-score →
 - Negative z-score →
 - z-score of 0 →

NOTE: standardized values have no units.

Benefits of Standardizing

- Standardized values have been converted from their original units to the standard statistical unit of *standard deviations from the mean*.
- Thus, we can compare values that are measured on different scales, with different units, or from different populations.
- Standardizing data into *z*-scores *shifts* the data by subtracting the mean and *rescales* the values by dividing by their standard deviation.
 - Standardizing into *z*-scores does not change the **shape** of the distribution.

- Standardizing into z -scores changes the **center** by making the mean 0.
- Standardizing into z -scores changes the **spread** by making the standard deviation 1.

Example:

In a class of 3 students, the average score on the final exam is 75% and standard deviation of 5%. Amy got 80% on the final exam. How many standard deviations better than the mean is that?

Mandy got 70%. How many standard deviations is Mandy's score deviate from the mean?

Judy got 75%. How many standard deviation is Judy's score deviate from the mean?

NOTE:

- 1) standardizing does not change the shape of the distribution of a variable
- 2) standardizing changes the center by make the mean 0
- 3) standardizing changes the spread by making the standard deviation 1.

Example:

Two graduate students:

- accounting major gets job offer for \$ 35000
- advertising major gets job offer for \$ 33000

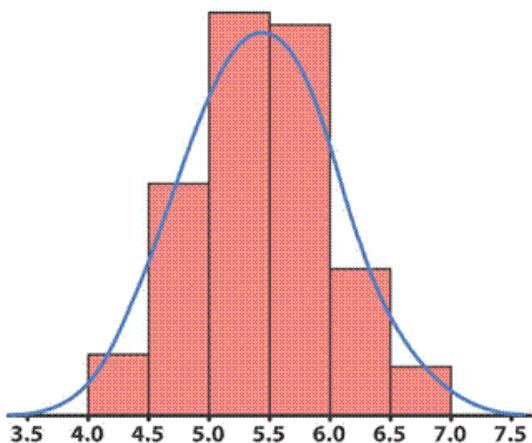
other students of:

- accounting: $\bar{y} = 34500$ and $s = 1500$
- advertising: $\bar{y} = 32500$ and $s = 1000$

Based on their own z-score, who do you think is happier about his/her job offer?

5.3 Density Curve and Normal Model

- Recall: Continuous random variables are described by histograms. For histograms, the measurement scale is divided in class intervals; and the area of the rectangle for each interval is proportional to the relative frequency of the data falling into each interval.
- A smooth curve may fit on the histogram and such curves are called *density curves*.



- There are many possible curves that might serve as useful models, but any density curve has to satisfy the following conditions:
 - A density curve is always on or above the horizontal axis.
 - The total area between the horizontal axis and under the density curve equals to 1

- c) The area under the curve above a certain interval is the proportion of all observations that fall in that range.

Ie. $P(a < x < b) = \text{area under the curve between } a \text{ and } b$

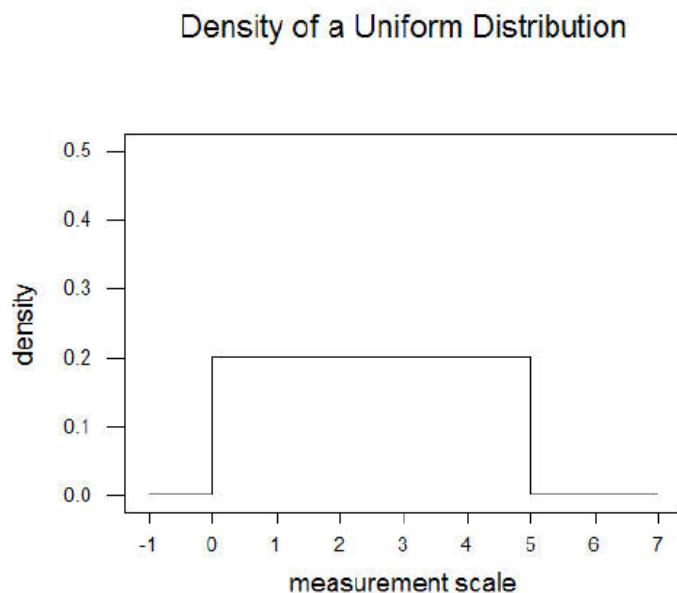
- d) There is no probability attached to any single value.

Ie. $P(x = a) = 0$ (**This is generally not true for discrete random variables.**)

Thus, $P(a < x < b) = P(a \leq x \leq b) = P(a < x \leq b) = P(a \leq x < b)$

Example:

The density of a uniform distribution in an interval $[0, 5]$ looks like this:



- a) Verify by geometry that the area under the curve is 1.

$$A = lw = 5 \times 0.2 = 1$$

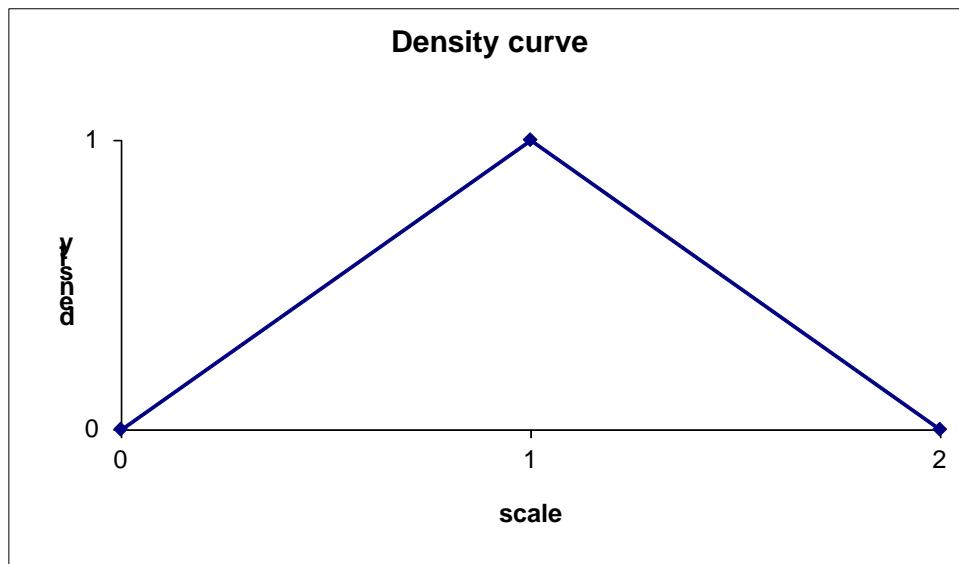
b) Using the above density curve, calculate the following probabilities:

a. $P(X \leq 3) = \text{area under the curve from } -\infty \text{ to } 3$
 $= 3 \cdot 0.2 = 0.6$

b. $P(1 \leq X \leq 2) = \text{area under the curve from } 1 \text{ to } 2$
 $= 1 \cdot 0.2 = 0.2$

c. $P(X > 3) = \text{area under the curve from } 3 \text{ to } \infty$
 $= 2 \cdot 0.2 = 0.4$
 $= 1 - 0.6 = 0.4$

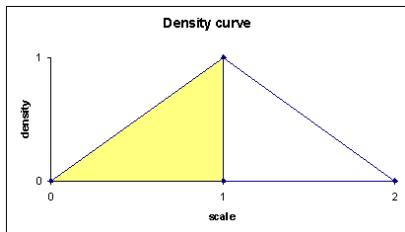
Example:



a) Verify by geometry that the area under the curve is 1.

$$A = \frac{1}{2} b h = \frac{1}{2} \times 2 \times 1 = 1$$

b) What is the probability that X is less than 1?



$$P(X < 1) = \frac{1}{2} \times 1 \times 1 = 0.5$$

c) What is the probability that X is greater than 1.5?

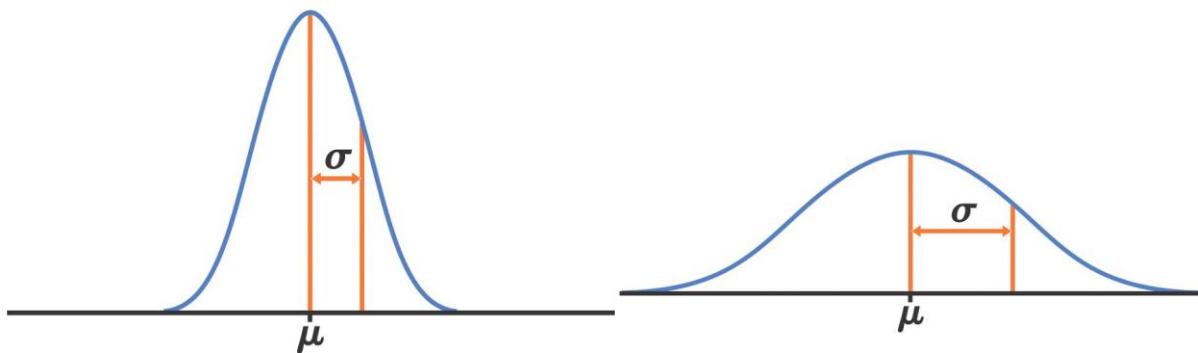
$$P(X > 1.5) = \frac{1}{2} \times 0.5 \times 0.5 = 0.125$$

Many numerical variables have bell shaped histograms. For example, heights, weights, lifetime of a light bulb, etc. The ***normal distribution*** provides a reasonable approximation for modeling this type of data. It is the most important and most widely used of all probability distributions.

Properties of normal distributions.

- These curves are symmetric, unimodal, and bell-shaped.
- For every combination of a mean μ and a standard deviation σ , there is a different curve.
 - μ is the center of the distribution (right at the highest point of the density distribution function)

- σ controls the spread of the distribution.
- Notation: $N(\mu, \sigma)$ represents a Normal model with a mean of μ and a standard deviation of σ .



Recall:

- 1) **Population parameter** is a numerical measure such as the mean, median, mode, range, variance, or standard deviation calculated for a population data; and is written with Greek letters. Eg. μ and σ .
- 2) **Sample statistic** is a summary measure calculated for a sample data set; it is written with Latin letters. Eg. \bar{y} , s

When we standardize Normal data, we still call the standardized value a ***z-score***, and we write

$$z = \frac{y - \mu}{\sigma}$$

- The *standardized value*, z , follows a standard normal distribution ($z \sim N(0,1)$).

Checking the Normality Assumption:

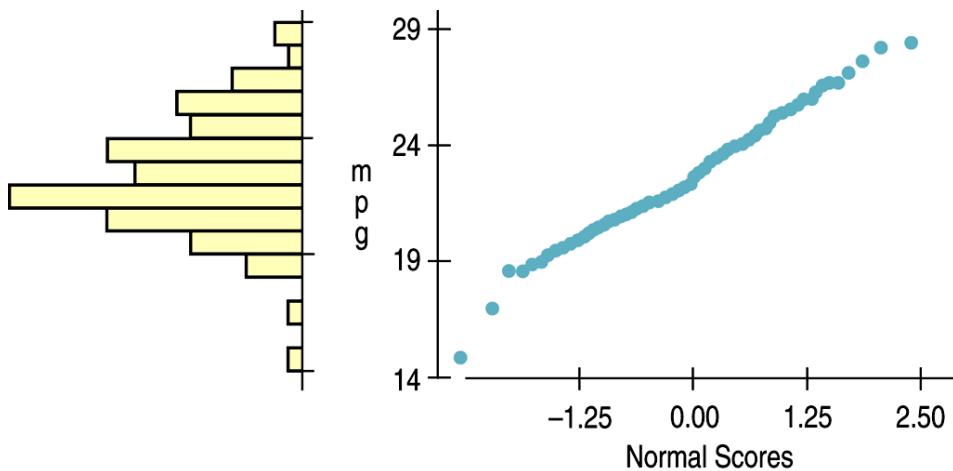
- When we use the Normal model, we are assuming the distribution is Normal.
- We cannot check this assumption in practice, so we check the following condition:
 - Nearly Normal Condition: The shape of the data's distribution is unimodal and symmetric.
 - This condition can be checked by making a histogram or a Normal probability plot.

5.5 Normal Probability Plot

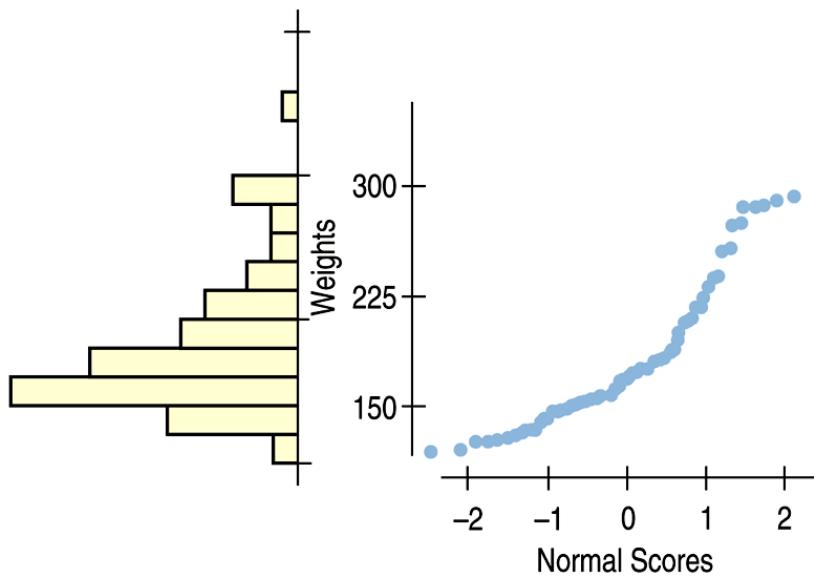
- A more specialized graphical display that can help to decide whether a normal model is appropriate
- If the distribution of the data is roughly normal, the normal probability plot approximates a diagonal straight line. Deviations from a straight line indicate that the distribution is not Normal

Example:

- Nearly normal data have a histogram and a normal probability plot that look somewhat like this example:



- A skewed distribution might have a histogram and normal probability plot like this:

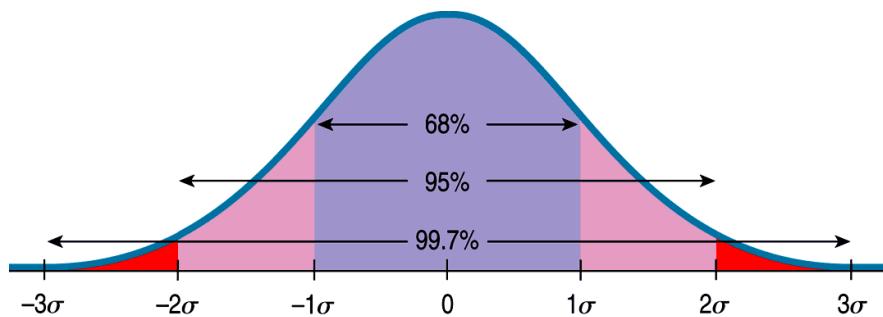


5.4 The Empirical Rule (68–95–99.7 Rule)

The Empirical Rule is derived from practical experience, which has shown that the normal curve provides a good model for a lot of different kind of variables. Thus, it applies to normal distributions and can be used for distributions that can be reasonably well described by a normal curve.

For all normal distributions, the **68–95–99.7 rule** holds:

- Approximately **68%** of the observations fall within 1σ of the mean μ .
- Approximately **95%** of the observations fall within 2σ of the mean μ .
- Approximately **99.7%** of the observations fall within 3σ of the mean μ .



Example:

The height of 112 children follows a normal distribution with $\bar{y} = 104.5$ and standard deviation $s = 16.3$.

Fill in the following chart:

k	$\bar{y} \pm ks$	Empirical
1		
2		
3		

Example:

The time to complete a standardized exam is approximately normal with a mean of 70 minutes and a standard deviation of 10 minutes. Using the 68-95-99.7 rule,

- what percentage of students will complete the exam in under an hour?
- what percentage of students will complete the exam between 60 minutes and 70 minutes?

- c) in what time interval would you expect the central 95% of students to be found?

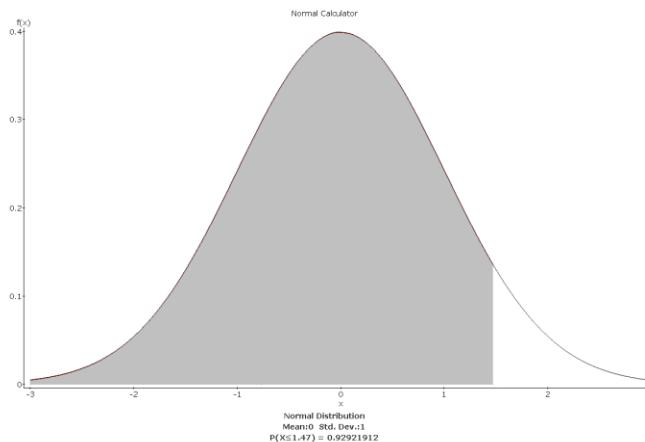
Suppose we want to find the probability within 1.5 standard deviations away from the mean, how to find such probability value?

Table z from the textbook (Appendix C) tabulates, for many different values of z^* , the area under the curve from $-\infty$ to z , which is called the **cumulative area** or **cumulative proportion**, for *standard normal* distributed variables.

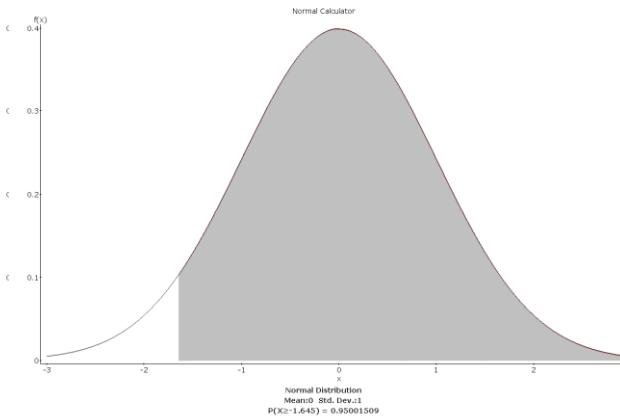
Example 2:

Using the table z, find the area under the standard normal curve,

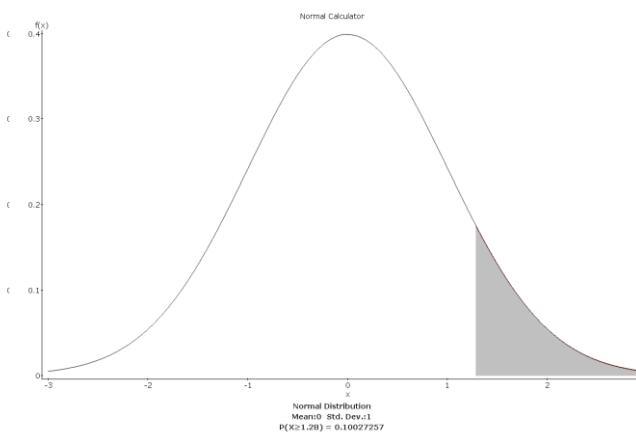
- to the left of $z = 1.47$:



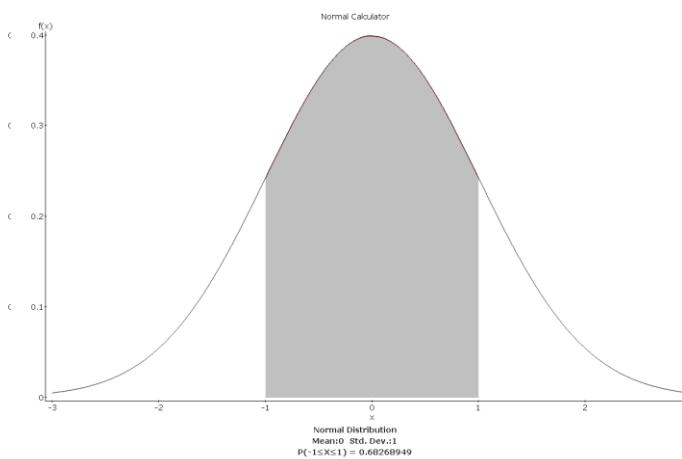
- to the right of -1.645:



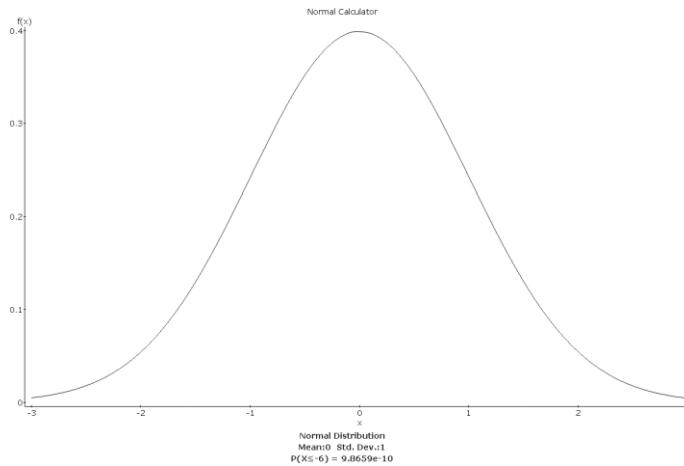
- to the right of 1.28:



- between -1 and 1:



- to the left of -6:



How to find the area for *any* normal distribution?

Lemma: If y is normal distributed with mean μ and standard deviation σ (ie. $y \sim N(\mu, \sigma)$), then the standardized variable

$$z = \frac{y - \mu}{\sigma}$$

is normal distributed with $\mu = 0$ and $\sigma = 1$ or ($z \sim N(0, 1)$).

The following example illustrates how the probability and the percentiles can be calculated by using the standardization process.

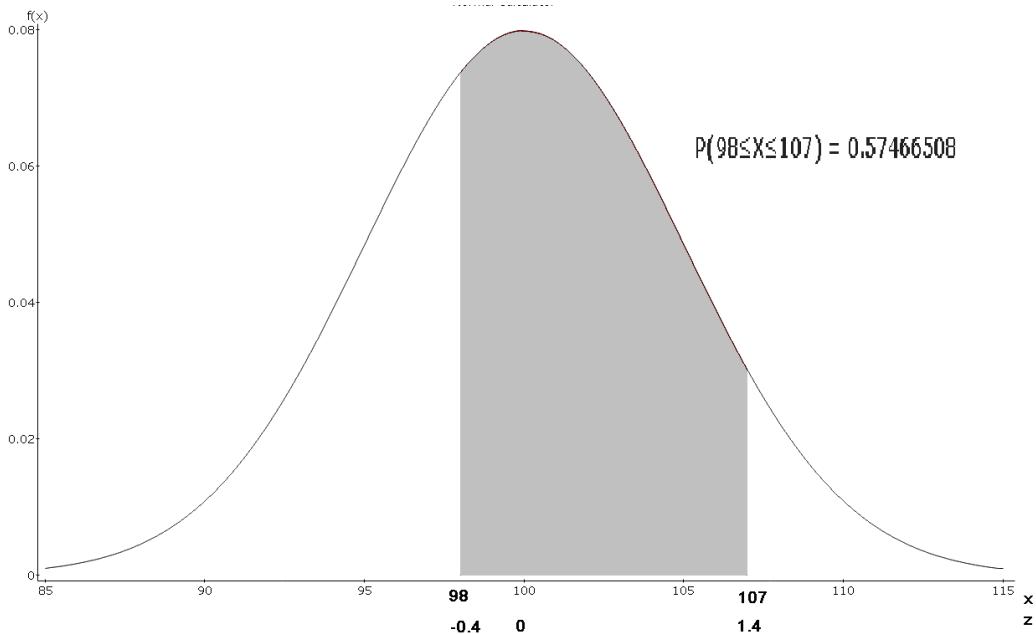
Example 3:

Let y be normal distributed with $\mu = 100$ and $\sigma = 5$, so

$y \sim N(100; 5)$. (NOTE: y is not standard normal)

Calculate the area under the curve between 98 and 107 for the distribution chosen above.

$$\begin{aligned} P(98 < y < 107) &= P\left(\frac{98-100}{5} < \frac{y-100}{5} < \frac{107-100}{5}\right) \\ &= P(-0.4 < z < 1.4) \\ &= P(z < 1.4) - P(z < -0.4) \quad \text{using } g \quad \text{Table } z \\ &= 0.9192 - 0.3446 = 0.5746 \end{aligned}$$



Summary to find normal proportions:

- 1) State the problem in terms of the observed variable y .
- 2) Standardize y to restate the problem in terms of a standard normal variable z . To visualize better, draw a picture to show the area under the curve.

$$P(y < a) = P\left(\frac{y - \mu}{\sigma} < \frac{a - \mu}{\sigma}\right) = P(z < a^*),$$

$$P(y > a) = P\left(\frac{y - \mu}{\sigma} > \frac{a - \mu}{\sigma}\right) = P(z > a^*),$$

$$P(a < y < b) = P\left(\frac{a - \mu}{\sigma} < \frac{y - \mu}{\sigma} < \frac{b - \mu}{\sigma}\right) = P(a^* < z < b^*),$$

$$\text{where } a^* = \frac{a - \mu}{\sigma}, b^* = \frac{b - \mu}{\sigma}$$

- 3) Find the required area using Table z.

Inverse normal calculations

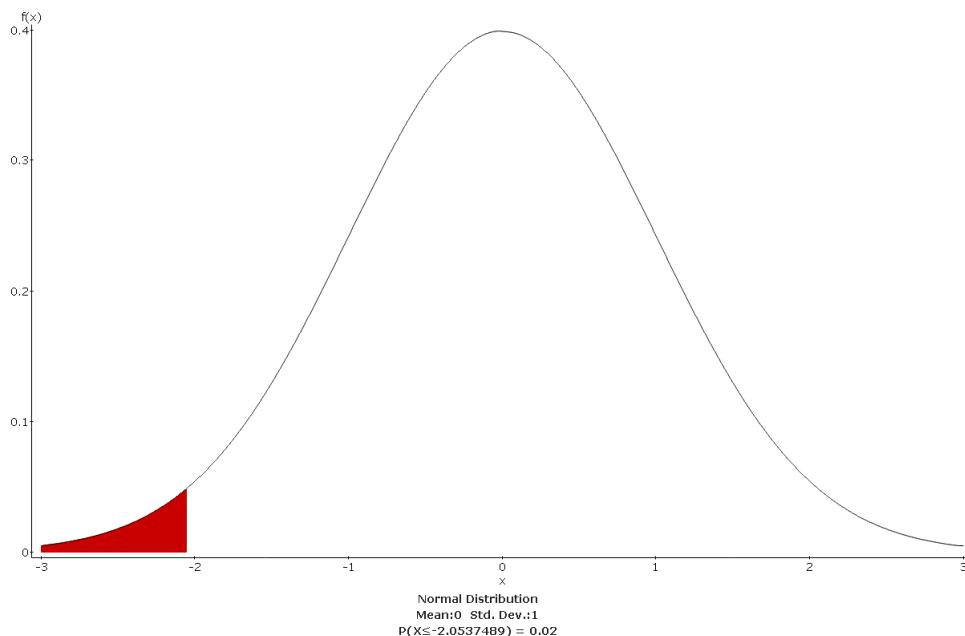
In other situations, we will not be given an interval for which we want to find the area above, but an area will be given and we are to find the value on the measurement scale so that the cumulative area up to this value matches the area given.

Example 4a:

Suppose we want to find the z-value that make up the smallest 2%.

In another words, we want to find z^* such that

$$P(z < z^*) = 0.02$$



Example 4b:

Suppose now we are interested in the largest 5%. So we are looking for z^* , with

$$P(z > z^*) = 0.05.$$

Example 4c:

Now suppose we are interested in the middle 95%. So we are looking for $-z^*$ and z^* , with

$$P(-z^* \leq z \leq z^*) = 0.95.$$

Example 5:

Let y be normal distributed with $\mu = 100$ and $\sigma = 5$. Find the value that makes up the smallest 30% for this distribution.

NOTE: This is NOT a standard normal distribution.

Summary for “backward” normal calculations:

- 1) State the problem in terms of the proportion.
- 2) Look in the body of Table z for the entry closest to the proportion, and find the corresponding z value.
- 3) Unstandardize to transform the z value back to the original y scale using $y = \mu + z\sigma$.

Example:

Assume that the length of a human pregnancy follows a normal distribution with mean 266 and standard deviation 16.

What is the probability that a human pregnancy lasts longer than 280 days?

How long do the 10% shortest pregnancies last?

Example: (please try this example on your own)

The weights of packages of ground beef are normal distributed with mean of 1 and standard deviation of 0.10.

- a) What is the probability that a randomly selected package weights between 0.8 and 0.85?

$$\begin{aligned} P(0.8 < y < 0.85) &= P(-2 < z < -1.5) \text{ standardize} \\ &= 0.0668 - 0.0228 = 0.440 \end{aligned}$$

- b) What is the weight of a package such that only 1% of all packages exceed this weight?

Find y^* such that $P(y > y^*) = 0.01$

This is equivalent to find z^* such that $P(z > z^*) = 0.01$

$$\Rightarrow P(z < z^*) = 1 - 0.01 = 0.99$$

Using Table Z, we get $z^* = 2.33$.

Unstandardize to get y^*

$$y^* = 1 + 0.1 * 2.33 = 1.233$$

Example: (please try this example on your own)

At a local high school track meet, the amount of time to run 100 meters follows a normal distribution with a mean of 25 seconds and a standard deviation of 3 seconds. According to this model, what time corresponds to the fastest 5% of this local high school students participating in this event?

- A) 17.23 seconds
- B) 20.07 seconds**
- C) 24.12 seconds
- D) 29.94 seconds
- E) 30.62 seconds

Solution:

Find y^* such that $P(y < y^*) = 0.05$

This is equivalent to find z^* such that $P(z < z^*) = 0.05$

Using Table Z, we get $z^* = -1.645$.

Unstandardize to get y^*

$$y^* = 25 - 1.645 \times 3 = 20.065$$

Ch 6 Scatterplots, Association, and Correlation

We will be investigating the relationship and association between two quantitative variables (bivariate data), such as height and weight, the concentration of an injected drug and heart rate, or the consumption level of some nutrient and weight gain.

Sometimes the purpose of a study is to show that one variable can explain the outcome of another variable.

Definition:

- **Response (or dependent) variable** (symbol: y) - measures an outcome of a study
- **Explanatory (or independent) variable** (symbol: x) explains or causes changes in the response variable.

Example 1: Distinguish the x and y variables

a) What is the effect of rainfall on crop yield?

- x :
- y :

b) What is the effect of the midterm score on the final grade?

- x :
- y :

Data:

- we measure x and y for each individual
- observations are recorded in the form (x, y)
- our sample of n bivariate observations is

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

6.1 Scatterplot

- is the best way to start observing the relationship and the ideal way to picture associations between two quantitative variables
- is a plot of pairs of observed values of two different quantitative variables. It helps to evaluate the **quality** of the relationship.
- The **x -axis** is the **horizontal** axis and **y -axis** is the **vertical** axis.
- Each observation is then plotted according to its value from the x variable and its value from the y variable.

Example:

Does the number of years invested in schooling pay off in the job market?

Thought: the better educated you are, the more money you will earn.

The data in the following table give the median annual income of full-time workers age 25 or older by the number of years of schooling completed.

$x = \text{Years of Schooling}$	$y = \text{Salary (dollars)}$
8	18,000
10	20,500
12	25,000
14	28,100
16	34,500
19	39,700

Create a scatterplot for x and y .



NOTE: If you want to make a scatterplot with more than 1 group, then use different symbols for each group.

NOTE: Axes need not to intersect at $(0, 0)$.

Examining a Scatterplot:

In any graph of data, look for the overall pattern and for striking deviations (ex. outliers) from this pattern. You can describe the overall pattern of a scatterplot by the form, direction, and strength of the relationship.

1) Direction (positive and negative associations)

- 2 variables are positively associated when x increases, y also

increases.



- 2 variables are negatively associated when x increases, y

decreases.



2) Form of relationship

- linear – where the points roughly follow a straight line
- curved relationship and clusters

3) Strength of the Relationship

- determined by how close the points in the scatterplot lie to a simple form such as a line

- the closer the observations appear to fit a line, the stronger the relationship.



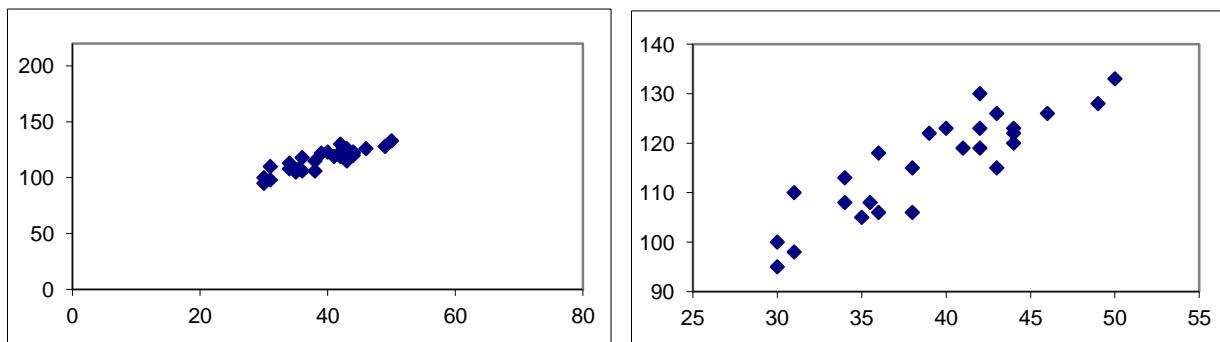
4) outliers or unusual observations

- look for any striking deviations from the overall pattern

Example:

Describe the pattern of the scatterplot above.

6.2 Correlation



If the scatterplot shows a reasonable linear relationship, calculate **correlation coefficient** to evaluate the **direction** and **strength** of the **linear** relationship between two numerical variables.

Correlation coefficient r :

- a numerical measurement of the strength of the linear relationship between the explanatory and response variables

$$r = \frac{\sum z_x z_y}{n-1} = \frac{\sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)}{n-1}.$$

- This is the sum of the products of the standardized values for each paired observation, all divided by $n - 1$.

Example:

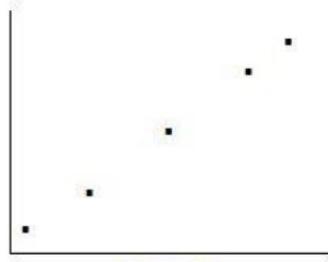
$x = \text{Years of Schooling}$	$y = \text{Salary (dollars)}$
8	18,000
10	20,500
12	25,000
14	28,100
16	34,500
19	39,700

Using Computer software, we obtained 0.9941 as the correlation coefficient between years of schooling and salary. What does this number imply?

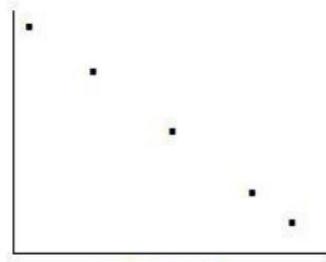
Facts about Pearson's correlation coefficient (r):

- 1) Correlation measures the strength of a **linear** relationship between two ***quantitative*** variables. Check a scatterplot first.
 - a. Correlation requires both variables to be numerical;
Cannot be applied to categorical data
 - b. does **NOT** apply to nonlinear relations
 - c. outliers can distort the correlation dramatically (an outlier can make a weak association look strong or a strong one look weak)
- 2) Correlation makes **no** distinction between explanatory and response variables, ie. The correlation of x with y is the same as the correlation of y with x .
- 3) Correlation has **no** units
- 4) Correlation is a number between -1 and 1
- 5) The **absolute value** of the coefficient measures how closely the variables are related.
 - The closer it is to 1 , the closer the relationship.
 - ❖ $|r| > 0.8 \rightarrow$ a strong correlation between the variables.
 - ❖ $r \approx 0 \rightarrow$ a weak linear association

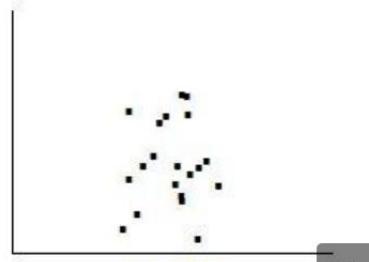
- 6) Like the mean and standard deviation, the correlation is strongly affected by outliers.
- 7) Correlation is not affected by changes in the center or scale of either variable.
- Correlation depends only on the z -scores, and they are unaffected by changes in center or scale.
- 8) The sign of the correlation coefficient tells you of the trend in the relationship.
- ❖ $r > 0$ indicates a positive relation between the variables
 - ❖ $r < 0$ indicates a negative relation between the variables



$r = 1$
Perfect positive correlation

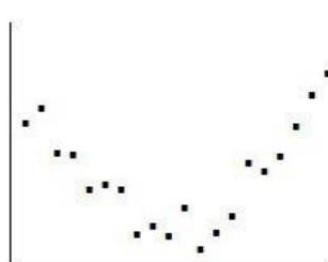


$r = -1$
Perfect negative correlation

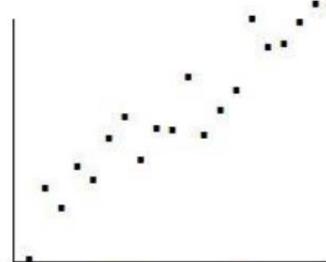


$r = 0$
Variables not related

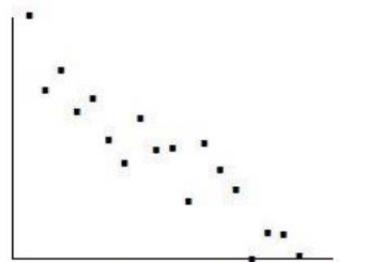
15



$r = 0$
Curved relationship



$0 < r < 1$
Positive correlation



$-1 < r < 0$
Negative correlation

6.3 Warning: Correlation \neq Causation

Lurking Variable

- It is a variable that is not among the explanatory or response variables in a study and yet may influence the interpretation of relationships among those variables.
- It can falsely suggest a strong relationship between x and y when there is none, or it can hide a relationship that is really there.

Example: education and salary.

y = salary; x = number of years of schooling

lurking variable: number of years of experience, company size, etc...

Correlation does NOT imply causation

- no matter how strong the correlation, no matter how large the r value, no matter how straight the line, there is ***no way to conclude from a regression alone that one variable causes the other.***

- the lurking variables warn us that an association between x and y doesn't prove that changes in x cause changes in y .
- With observational data, as opposed to data from a designed experiment, there is no way to be sure that a **lurking variable** is not the cause of any apparent association.

Example: damage caused to a forest by fire and the number of firefighters

A scatterplot of the damage caused to a forest by fire would show a strong positive correlation with the number of firefighters at the scene.

- 1) Surely the damage doesn't cause firefighters.
- 2) firefighters do seem to cause some of the damage through spraying water all around and chopping holes.

Does that mean we shouldn't call the fire department?

How to solve the problem?

The best way to get good evidence that x causes y is to do an experiment in which we change x and keep lurking variables under control.

Ch 7 Linear Regression & Ch 8 Regression Wisdom

7.1 – 7.2 Least Squares: The Line of “Best Fit” and The Linear Model

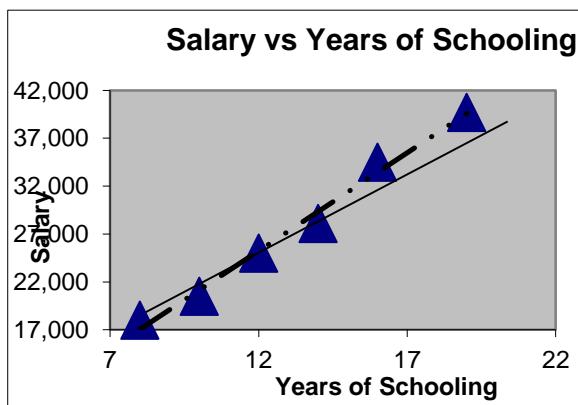
Idea: To fit a straight line through the data so that we can predict values of the response at specified values of x .

When we have one dependent variable and one independent variable **and** the relationship between two variables follows a linear pattern, it is possible to describe the relationship by a straight line and by an equation of the form:

$$y = b_0 + b_1 x$$

where b_0 is called the **y-intercept** and b_1 the **slope** of the equation.

The b 's are called the **coefficients of the linear model**.



The slope is the amount by which y increases when x increases by 1 unit.

How to find the line that best describes the linear relationship?

Estimate: $\hat{y} = b_0 + b_1x$

- \hat{y} gives an estimate (predicted response) for y for a given value of x
- $\hat{y} = b_0 + b_1x$ is called the ***line of best fit*** or the ***least squares regression line***.

Note 1: $\hat{y} \neq y$. The vertical distance from a data point (x, y) to the line is called the **error of prediction** or deviation or residuals.

Residual or Deviation of the i^{th} data point (x_i, y_i) is:

$$\text{residual} = \text{observed} - \text{predicted} = y_i - \hat{y}_i$$

- A negative residual means the predicted value is too big (an overestimate)
- A positive residual means the predicted value is too small (an underestimate)

Note 2: Sum of the residuals is always 0. Thus, we can't assess how well the line fits by adding up all the residuals.

Note 3: Similar to what we did with deviations, we square the residuals and add the squares.

Note 4: the smaller the sum, the better the fit.

Conclusion: The best fitted line is the one that minimizes the sum of the squared residuals between the data points and the line itself.

$$\min_{a,b} \sum_i^n (\text{residuals})^2 = \min_{a,b} \sum_i^n (y_i - \hat{y}_i)^2$$

- Minimize the sum by choosing the appropriate parameters b_0 and b_1 .
- The resulting line is called the *least square line* or *best fitted line*: $\hat{y} = b_0 + b_1 x$
- Thus, the sum of squared residuals of the least square line is smaller than that for any other straight-line model.

After the problem is stated, it can be solved mathematically for the *best* parameters b_1 and b_0 :

Slope: $b_1 = r \left(\frac{s_y}{s_x} \right)$

- Slope is always in units of y per unit of x
- Moving one standard deviation away from the mean in x moves us r standard deviations away from the mean in y

Intercept: $b_0 = \bar{y} - b_1 \bar{x}$

- Intercept is always in units of y

Example:

The following best fitted line predicts the Stat 151 final exam mark by the number of studying hours during the term:

$$\hat{final} = 18 + 0.7 \text{hours}$$

- The **slope**, 0.7, says that a Stat 151 student can be expected, on average, to score 0.7% more when he studies an additional hour during the term.
- Algebraically, the intercept is the value the line takes when x is zero. i.e. A student who did not study would have, on average, about 18% on the final exam. Normally, the intercept has some meaning if the x -values include zero values.

Continue Example (from Ch6):

Since the salary and the years of schooling show such a strong linear relationship and the salary can be viewed as a variable depending on the years of schooling, find the best fitted line with the salary as the response variable and the years of schooling as the predictor variable with the following summary statistics table:

Summary statistics:

Column	n	Mean	Variance	Std. Dev.	Median	Min	Max	Sum
x	6	13.166667	16.166666	4.020779	13	8	19	79
y	6	27633.334	6.8718664E7	8289.672	26550	18000	39700	165800

Simple linear regression results:

Dependent Variable: y

Independent Variable: x

$$y = 648.4536 + 2049.4846 x$$

Sample size: 6

R (correlation coefficient) = 0.9941

R-sq = 0.9881777

Estimate of error standard deviation: 1007.72784

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-Value
Intercept	648.4536	1532.058	$\neq 0$	4	0.42325658	0.6939
Slope	2049.4846	112.08514	$\neq 0$	4	18.28507	<0.0001

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	3.39531264E8	3.39531264E8	334.34378	<0.0001
Error	4	4062061.8	1015515.44		
Total	5	3.43593344E8			

Using this regression model, what is the estimated average salary after 18 years of schooling?

NOTE: *Don't use the regression line for values outside the range of the observed values (extrapolations). This is a model that **only** has been proved valid for the given range.*

Properties of the regression or least squares line

1. The least squares line passes through the balance point (\bar{x}, \bar{y}) of the data set.
2. The regression line of y on x should not be used to predict x , since it is not the line that minimizes the sum of squared x deviations.
3. The line shows that a unit increase in x corresponds to b_1 units of change in our predicted y value.
 - $b_1 < 0$, then y decreases as x increases.
 - $b_1 > 0$, then y increases as x increases.

Ch 11 From Randomness to Probability

11.1 Random Phenomena

Many things in our world depend on randomness:

- What is the chance to observe a Head while flipping a coin?
- What is the chance to roll a 6 with a "fair" die?
- What is the probability that the bus will be on time today?
- What is the probability that you lose your investment?

Even those events occur randomly, there is an underlying pattern in the occurrence of these events. This is the basis of Probability Theory.

Definition:

1. A phenomenon is **random** if we know what outcomes could happen, but we don't know which particular outcome did or will happen.
 - Individual outcomes are unpredictable
 - With a large number of observations, predictable patterns occur

Examples for random phenomenon are:

- Count red blood cells in a blood sample

- Roll a die
 - Toss a coin
 - Toss a coin twice
2. In general, each occasion upon which we observe a random phenomenon is called a **trial**.
 3. At each trial, we note the value of the random phenomenon, and call it an **outcome**.
 4. When we combine outcomes, the resulting combination is an **event**.
 - An **event** is a subset of the sample space (or is the collection of one or more of the outcomes of an experiment)
 - Events are usually denoted by letters from the beginning of the alphabet, such as A and B , or by a letter or string of letters that describes the event.
 5. The collection of *all possible outcomes* is called the **sample space**.

Con't Example:

Give the sample spaces of above random phenomenon:

Examples for events:

Suppose you roll an unbiased die with 6 faces and observe the number on the upper face.

Let A be the event of rolling 1 or 2

Let B be the event of rolling 3 or 4

Let C be the event of rolling an odd number

A **Venn diagram** is a picture that depicts all the possible outcomes for an experiment.

The outer box represents the sample space, which contains all of the outcomes, and the appropriate events are circled and labeled.

Example:

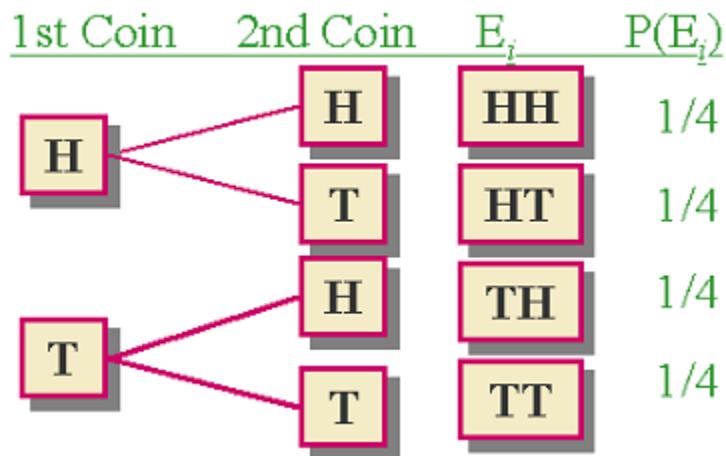
Draw a Venn Diagram for the dice example:

Example: the 2 coins toss.

Let A be the event that occurs if at least a head is tossed.

Example:

If we toss 2 coins, what is the sample space? Draw a tree diagram to illustrate it.



This is a **tree diagram** (each outcome is represented by a branch of the tree).

- An ideal way of visualizing sample spaces with a small number of outcomes
- As the number of trials or the number of possible outcomes on each trial increase, the tree diagram becomes impractical

Example:

In a pop quiz with 3 multiple-choice questions, each question has 5 options, and the student's answer is either correct (C) or incorrect (I). Determine the total number of possible outcomes.

What is the sample space for the correctness of a student's answers on this pop quiz. Draw a tree diagram to illustrate it.

Let A be the event a student answers all 3 questions correctly

Let B be the event a student passes (at least 2 correct)

6. The **probability** of any outcome (or event) of a random phenomenon is the proportion of times the outcome would occur in a very long series of repetitions. (*Probability is a measure for the likelihood or chance of a future event*)
 - This definition is based on the **Law of Large Numbers** (LLN): the long-run *relative frequency* of repeated independent events gets closer and closer to a single value.
 - The LLN says nothing about short-run behavior.
 - Relative frequencies even out *only in the long run*, and this long run is *really* long (*infinitely* long, in fact).

11.2 Modeling Probability

Equally likely

- It's equally likely to get any one of six outcomes from the roll of a fair die.

Outcome	1	2	3	4	5	6
Probability	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

- It's equally likely to get heads or tails from the toss of a fair coin.

Outcome	Head	Tail
Probability	$\frac{1}{2}$	$\frac{1}{2}$

- However, keep in mind that events are *not* always equally likely.
 - A skilled basketball player has a better than 50-50 chance of making a free throw.

Probability of an Event

In general, the probability of an event A is:

$P(A) = \text{sum of the probabilities of the individual outcomes contained in } A.$

With equally likely outcomes, the probability of an event A is:

$$P(A) = \frac{\text{Number of outcomes in } A}{\# \text{ of possible outcomes}}$$

Example: Find the probability of A .

- 1) Let A be the event of obtaining a ‘2’ in one roll of an unbiased die.
- 2) Let A be the event of obtaining a ‘H’ in one toss of an unbiased coin.
- 3) Let A be the event of obtaining a ‘HH’ in two tosses of an unbiased coin.
- 4) Let A be the event of obtaining an even number in one roll of an unbiased die.
- 5) Let A be the event of obtaining at least one tail in two tosses of an unbiased coin.

11.3 Formal Probability Rules

The probabilities must follow the following rules:

- 1) 2 requirements for a probability:
 - The probability of each simple event (individual outcome) is between 0 and 1. ie. For any event A , $0 \leq P(A) \leq 1$
 - $P(A) = 0$, if the event A never occurs.
 - $P(A) = 1$, if the event always occurs.
- 2) Total Probability Rule: the probability of the set of all possible outcomes of a trial must be 1. ie. then $P(S) = 1$, where S is the whole sample space or the set of all possible outcomes
 - Roll a 6-sided die: $P(\{1,2,3,4,5,6\}) =$
 - $P(\text{roll with a regular die a number} < 7) =$
 - For the toss of 1 coin: $P(S) = P(\{H, T\}) =$
 - For the toss of 2 coins: $P(S) = P(\{HH, HT, TH, TT\}) =$

Example: Blood Types

All human blood can be typed as one of O, A, B, or AB. Here is the blood distribution for everyone in a small town:

Blood Type	O	A	B	AB
Pbty	0.49	0.27	0.2	?

a) What is the probability of type AB blood? Why?

b) Maria has type B blood. She can safely receive blood transfusions from people with blood types B and O. What is the chance that a randomly selected person from this town can donate blood to Maria?

3. Complement Rule:

- a. Complement of an event A (denoted by A^c)
- a. Consists of all outcomes in the sample space that are *not* in A .
- b. The probabilities of A and A^c add to 1
(ie. $P(A^c) + P(A) = 1$)
- c. $P(A^c) = 1 - P(A)$

Remark: $P(A \cap A^c) = 0$

$$P(A \cup A^c) = 1$$

Example 2 cont:

Using the pop quiz example, draw a Venn diagram to illustrate B and B^c .

Recall: B is the event that a student passes (at least 2 correct);

$$B = \{CCI, CIC, ICC, CCC\}$$

$$B^c = \{CII, IIC, ICI, III\}$$

NOTE: B is the event that a student fails (less than 2 correct);

Example: Use the complement rule to find the probability of having at least one tail in two tosses of an unbiased coin?

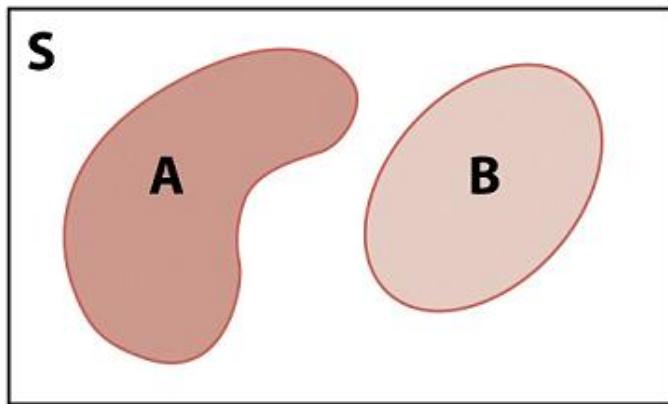
Recall: $P(A) = P(\{TH, HT, TT\}) = 3/4$

OR $P(A) = 1 - P(A^c) = 1 - P(\text{no tails in the two tosses})$

$$= 1 - P(\{HH\}) = 1 - 1/4 = 3/4$$

Definitions:

- If A and B are **disjoint events (or mutually exclusive events)**, then they have no outcomes in common (ie. if when one event occurs, the other cannot occur, and vice versa.)



Example Cont: Quiz

Recall:

A: student answers all 3 questions correctly

B: student passes (at least 2 correct)

Now, let

- Event C: Student answers exactly 1 question correctly
- Event D: Student answer exactly 2 questions correctly

Which of the events (A, B, C, D) are disjoint?

In another words:

$$A = \{\text{CCC}\}$$

$$B = \{\text{CCI}, \text{CIC}, \text{ICC}, \text{CCC}\}$$

$$C = \{\text{CII}, \text{ICI}, \text{IIC}\}$$

$$D = \{\text{CCI}, \text{CIC}, \text{ICC}\}$$

Notations:

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

$$A \cup B = A \text{ or } B$$

Let A and B be two events:

- The **intersection** of events A and B , denoted by $A \cap B$, consists of outcomes that are in both A and B .
- The **union** of events A and B , denoted by $A \cup B$, is the event that either A or B or both occur.

General Addition Rule: Probability of the Union of 2 Events

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

Addition Rule for Disjoint Events:

$$P(A \cup B) = P(A) + P(B)$$

Remark 1: The subtraction of $P(A \cap B)$ is necessary because this area is counted twice by the addition of $P(A)$ and $P(B)$, once in $P(A)$ and once in $P(B)$.

Remark 2: The probability of the intersection of disjoint events is

$$P(A \cap B) = 0$$

Example: A fair dice

Let A be the event to roll an odd number.

Let B be the event to roll a number greater than 2.

What is $P(A)$, $P(B)$, $P(A^c)$, $P(B^c)$, $P(A \cap B)$, $P(A \cup B)$?

$$P(A) = P(\{1,3,5\}) =$$

$$P(A^c) = 1 - P(A) = \quad \text{OR: } P(A^c) = P(\{2,4,6\}) =$$

$$P(B) = P(\{3,4,5,6\}) =$$

$$P(B^c) = 1 - P(B) = \quad \text{OR: } P(B^c) = P(\{1,2\}) =$$

$$P(A \cap B) = P(\{1,3,5\} \cap \{3,4,5,6\}) =$$

$$P(A \cup B) = P(\{1,3,5\} \cup \{3,4,5,6\}) =$$

$$\text{OR } P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Example: Color

Suppose that 55% like green, 25% like red, and 45% like yellow.

Also, suppose 15% like both green and red, 5% like all 3 colors, 25% like both green and yellow, and 5% only like red. How many like green or red or both?

How to find $P(A \text{ and } B)$?

Sometimes $P(A \cap B)$ is not given in a question, and the rule for calculating $P(A \cap B)$ depends on the idea of **independent and dependent events**.

12.2 Independence

Multiplication Rule: Probability of the Intersection of Independent Events

$$P(A \text{ and } B) = P(A) \times P(B)$$

Independence of Two Events

Two events are considered independent, when the occurrence of one of the events has no impact on the probability for the second event to occur.

- NOTE: it doesn't mean they are disjoint!!
 - Disjoint events *cannot* be independent! Well, why not?
 - Since we know that disjoint events have no outcomes in common, knowing that one occurred means the other didn't.
 - Thus, the probability of the second occurring changed based on our knowledge that the first occurred.
 - It follows, then, that the two events are *not* independent.

- A common error is to treat disjoint events as if they were independent, and apply the Multiplication Rule for independent events—don't make that mistake.

Example:

Toss a fair coin twice. What is the probability to toss two heads?

Recall: $P(\{HH\}) = \frac{1}{4}$

OR Define:

A: head on first toss

B: head on second toss

and find $P(A \text{ and } B) = P(A) * P(B) = \frac{1}{2} * \frac{1}{2} = \frac{1}{4}$

This is independent event.

Example:

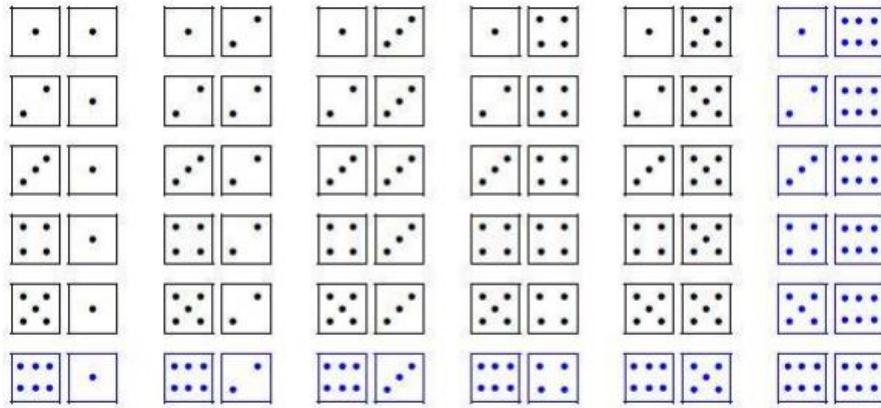
What is the probability to get the outcome (HTTH) with a biased coin which has a 0.1 chance to toss a head.

$$P(\{HTTH\}) = P(H)P(T)P(T)P(H)$$

Example: Two and Three Dice Rolling Game.

a) When you roll two dice, what is the probability to roll two 6's?

There are 36 possible outcomes:



b) What is the chance that none of the dice are divisible by 3 with 2 fair dice?

Use $P(A \text{ and } B) = P(A)P(B)$

Let A = no numbers are divisible by 3 with the first dice

Let B = no numbers are divisible by 3 with the second dice

c) What is the chance of rolling at least one  with 2 fair dice?

Method 1:

Method 2: Using Probability Rule 5, we have:

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

Let A to be the event that the first die is , and

let B to be the event that the second die is .

Method 3: Using the complement rule, we have:

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

Let A to be the event to have at least one 

d) What is the chance of rolling at least one  with 3 fair dice?

e) What is the chance of getting not all  with 3 fair dice?

Example: Pop Quiz with 3 Multiple-choice questions

- Each question has 5 options
- A student is totally unprepared and randomly guesses the answer to each question
- The probability of selecting the correct answer by guessing = 0.20
- Responses on each question are independent

What is the probability that a student answers

1. all questions correctly?

$$P(\{\text{CCC}\}) =$$

2. one question correctly?

$$P(\{\text{IIC}, \text{ICI}, \text{CII}\}) =$$

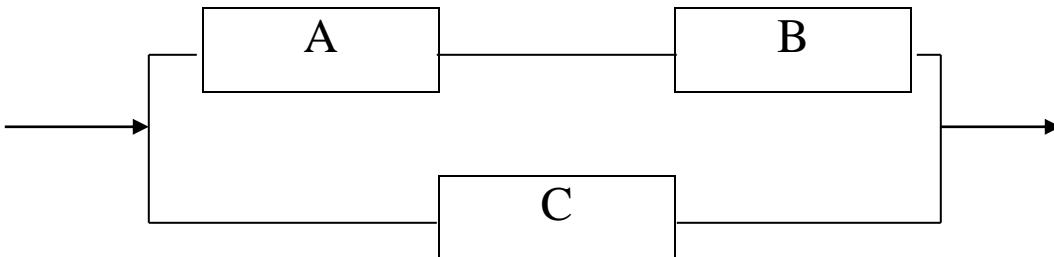
3. *at least* 1 questions correctly?

$$P(\{\text{CCC}, \text{CCI}, \text{CIC}, \text{ICC}, \text{IIC}, \text{ICI}, \text{CII}\}) =$$

4. *at least* 2 questions correctly?

$$P(\{\text{CCC}, \text{CCI}, \text{CIC}, \text{ICC}\}) =$$

Example: System reliability



The signal will pass through if “A and B” or “C” is working.

If these 3 independent components are 95% reliable, what is the reliability of the system?

Example:

A satellite has two power systems, a main and an independent backup system. Suppose the probability of failure in the first ten years for the main system is 0.05 and for the backup system 0.08.

- What is the probability that both systems fail in the first 10 years and the satellite will be lost?

Let F_M be the event that the main system fails in the first 10 years, F_B the event that the backup system fails in the first 10 years.

		Backup	Total
		F_B	F_B^c
Main	F_M		
	F_M^c		

- b) What is the probability that at least one of the systems is still functional after the first 10 years?

Let O be the event that at least one of the systems is still operational after 10 years

- c) What is the probability that both systems are still functional after the first 10 year?

NOTE: F_M^c is the event that the main system is functional, and F_B^c is the event that the backup system is still functional

Example: (Please try on your own)

In an intro stats class, 57% of students eat breakfast in the morning and 80% of students floss their teeth. 45.6% percent of students eat breakfast and floss their teeth.

- a) What is the probability that a student from this class eats breakfast and does not floss their teeth?

Let B = student eats breakfast and F = student flosses his teeth

Check for independence:

$$P(B \text{ and } F) = P(B) * P(F) = .57 * .8 = .456 \rightarrow \text{independent}$$

Thus,

$$P(B \text{ and } F^c) = P(B) * P(F^c) = .57 * (1 - .8) = .114$$

- b) What is the probability that a student from this class does not eat breakfast, or eat breakfast and does not floss their teeth?

$$\begin{aligned} P(B^c \text{ or } (B \text{ and } F^c)) &= P(B^c) + P(B \text{ and } F^c) \\ &= (1 - .57) + .114 = .43 + .114 = .544 \end{aligned}$$

		Breakfast		Total
		B	B^c	
Floss	F	$P(B \cap F) = 0.456$	$0.8 - 0.456 = 0.344$	$P(F) = 0.8$
	F^c	$0.57 - 0.456 = 0.114$	$0.2 - 0.114 = 0.086$	$1 - 0.8 = 0.2$
	Total	$P(B) = 0.57$	$1 - 0.57 = 0.43$	1

Example:

Insurance company records indicate that 12% of all teenage drivers have been ticketed for speeding and 9% for going through a red light. If 4% have been ticketed for both, what is the probability that a teenage driver

- a) has been issued a ticket for speeding but not for running a red light?

Let R = running a red light and S = get ticket for speeding

$$P(R \text{ and } S) = 0.04$$

$$P(S) = 0.12$$

$$P(R) = 0.09$$

		Speeding	Total
		S	S^c
Running a red light	R	0.04	0.09
	R^c		
	Total	0.12	1

b) has been issued a ticket for speeding or for running a red light but not both?

c) has not been issued a ticket for speeding nor for running a red light?

Often Events are Not Independent:

Pop Quiz Example cont:

Assume the instructor only gives 2 questions in the pop quiz and he finds the proportions for the actual responses of her students:

Outcome	II	IC	CI	CC
Pbty	0.26	0.11	0.05	0.58

Let A: {first question correct}

B: {2nd question correct}

Find P(A), P(B) and P(A and B). Are A and B independent events?

$$P(A) = P(\{CI, CC\}) =$$

$$P(B) = P(\{IC, CC\}) =$$

$$P(A \cap B) = P(\{CC\}) =$$

If A and B were independent, then

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

NOTE: Responses to different questions on a quiz are typically not independent. Most students do not guess randomly. Students who get the first question correct may have studied more than students who do not get the 1st question correct, and thus they may also be more likely to get the 2nd question correct.

NOTE: Don't assume that events are independent unless you have given this assumption careful thought and it seems plausible.

To understand the concept of independent and dependent events even further, we will introduce the idea of conditional probability.

Definition:

If A and B are events with $P(B) > 0$, the conditional probability of A given B is defined by

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

The interpretation of the conditional probability $P(A|B)$ is as follows:

Given that you know already event B occurred, what is the probability that A occurs.

Example:

Toss a fair coin twice. Define

A: head on second toss

B: head on first toss

Find $P(A|B)$ and $P(A|B^c)$.

Example:

Suppose I roll a fair die and ask, “What is the chance to get a ?”

The answer is $P(\text{---}) = 1/6$

Now suppose I give you a hint: the number on the die is even.

Now what is the chance that it is a ?

$P(\text{---}|\text{even number}) = 1/3$

Example: Consider the text example on the sample survey of all Canadians visiting Stonehenge on Canada Day. The distribution of official language commonly used by region is:

	English	French	Total
Atlantic	15	9	24
Quebec	5	50	55
Ontario	89	4	93
West	70	6	76
Total	179	69	248

Based on the contingency table, find:

- $P(\text{French Speakers}) = 69/248 = 0.2782$
- $P(\text{French and Atlantic}) = 9/248 = 0.0363$
- $P(\text{Quebecer}) = 55/248 = 0.2218$
- The probability that a selected person prefers French *given that we have selected a Quebecer*

$$P(\text{French} \mid \text{Quebecer}) = 50/55 = 0.909$$

The General Multiplication Rule:

Rearranging the equation in the definition for conditional probability, we get the General Multiplication Rule:

- 1) $P(A \cap B) = P(A) P(B|A))$, or
- 2) $P(A \cap B) = P(B) P(A|B))$

NOTE: $P(A/B) \neq P(A)$; $P(B/A) \neq P(B)$ for dependent events

Summary for Independent events:

We say events A and B are **independent** if any one of the following conditions is satisfied:

- 1) $P(A|B) = P(A);$
- 2) $P(B|A) = P(B); OR$
- 3) $P(A \cap B) = P(A) P(B)$

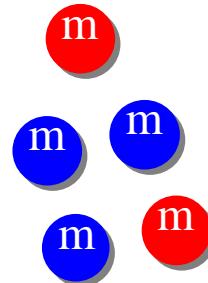
Example:

A bowl contains five M&Ms®, two red and three blue. Randomly select two candies without replacement, and define

A: second candy is red.

B: first candy is blue.

Find $P(A|B)$ and $P(A|B^c)$.

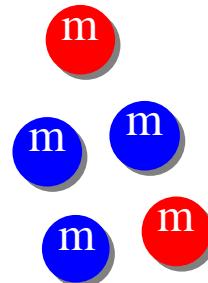


NOTE:

A bowl contains five M&Ms®, two red and three blue. Randomly select two candies with replacement, and define

A: second candy is red.

B: first candy is blue.



Find $P(A|B)$ and $P(A|B^c)$.

Example:

Five juniors and four seniors have applied for two open student council positions. School administrators have decided to pick the two new members randomly. What is the probability that one junior and one senior are chosen for the two positions?

Let J = the person chosen is junior; S = the person chosen is senior

Once you've decided whether or not two events are independent or dependent, you can use the following rule to calculate their intersection.

Multiplication Rule for Finding $P(A$ and $B)$

- For events A and B , the probability that A and B both occur:
 - $P(A \text{ and } B) = P(A|B) \times P(B)$; and
 - $P(A \text{ and } B) = P(B|A) \times P(A)$
- For **independent** events A and B , the probability that A and B both occur is
 - $P(A \cap B) = P(A) \times P(B)$

Example: Drawing Straws

Six soldiers have to decide between themselves which one goes on a suicide mission. They decide to draw straws: there are 5 long straws and 1 short one, and they take turns picking one. The guy with the short straw loses.

Is it better to pick first or second?

Example:

In a criminal trial, a person is being suspected as a murderer. In general, the probability that the jury convicts an innocent person is 0.04, and the probability that the jury sets a murderer free is 0.10. If a city has a criminal rate of 0.25, what is the probability that the jury:

- a) finds the person innocent when in fact he is innocent?
- b) finds the person guilty when in fact he is the murderer?
- c) makes the right decision?
- d) Finds the person innocent and in fact he is innocent given the jury makes the right decision?

Conditional Distribution:

		Decision		Total
		I_D	I_{D^c}	
Truth	I_T		$P(I_{D^c} I_T) = 0.04$	1
	I_T^c	$P(I_D I_T^c) = 0.1$		1
	Total			

		Decision		Total
		I_D	I_{D^c}	
Truth	I_T			
	I_T^c			$P(I_T^c)$ $= 0.25$
	Total			1

Example: (Please try it on your own) How Likely Are You to Win the Lotto? In a special city's Lottery, 3 numbers are randomly sampled without replacement from the integers 1 to 20.

a) If you buy a lotto ticket, how likely are you going to win? In another words, what is the probability that it is the winning ticket?

Let A = you have the same first number

B = you have the same second number

C = you have the same third number

$$P(\text{have all 3 numbers}) = P(A \text{ and } B \text{ and } C)$$

$$= P(A) \times P(B|A) \times P(C|A \text{ and } B)$$

$$= 3/20 \times 2/19 \times 1/18 = 0.000877$$

(NOTE: A, B and C are not independent events. You need A, before you can have B, and you need both A and B before you can have C.)

b) what is the probability that the ticket will match at least one of the 3 numbers?

Let A = at least one of 3 numbers

$$P(A) = 1 - P(A^c) = P(\text{have none of 3 numbers})$$

$$= 1 - P(\text{no 1}^{\text{st}} \& \text{ no 2}^{\text{nd}} \& \text{ no 3}^{\text{rd}})$$

$$= 1 - P(\text{no 1}^{\text{st}}) \times P(\text{no 2}^{\text{nd}} | \text{no 1}^{\text{st}}) \times P(\text{no 3}^{\text{rd}} | \text{no 1}^{\text{st}} \text{ nor } 2^{\text{nd}})$$

$$= 1 - 17/20 \times 16/19 \times 15/18 = 0.000877$$

Example: (Please try it on your own)

The probability that a student is a male is 0.5. The probability that a chosen male student has brown hair is 1/3. The probability that a chosen female student has brown hair is 0.5. Find the probability that a random chosen student

- a) is a male and does not have brown hair

Let $M = \text{male}$; $B = \text{brown hair}$

$$P(M \cap B^c) = P(M) * P(B^c|M) = 0.5 * (1 - 1/3) = 2/6 = 1/3$$

- b) is a female and has brown hair

$$P(M^c \cap B) = P(M^c) * P(B|M^c) = (1 - 0.5) * 0.5 = 0.25$$

- c) has brown hair?

$$\begin{aligned} P(B) &= P(M \cap B) + P(M^c \cap B) \\ &= P(M) * P(B|M) + P(M^c) * P(B|M^c) \\ &= 0.5 * 1/3 + (1 - 0.5) * 0.5 = 5/12 \end{aligned}$$

- d) is a female and/or has brown hair?

$$\begin{aligned} P(M^c \cup B) &= P(M^c) + P(B) - P(M^c \cap B) \\ &= 1/2 + 5/12 - 0.25 = 2/3 \end{aligned}$$

e) Now, assume that it is known that a student with brown hair is selected. What is the probability that the student selected is female?

$$P(M^c | B) = P(M^c \cap B)/P(B) = (0.25)/(5/12) = 3/5$$

Condition: Gender

		Brown		Total
		B	B^c	
Gender	M	1/3	$1 - 1/3 = 2/3$	1
	M^c	0.5	$1 - 0.5 = 0.5$	1

		Brown		Total
		B	B^c	
Gender	M	$1/3 \times 0.5 = 1/6$	$0.5 - 1/6 = 1/3$	0.5
	M^c	$0.5 \times 0.5 = 0.25$	$0.5 - 0.25 = 0.25$	$1 - 0.5 = 0.5$
		$1/6 + 0.25 = 5/12$	$1/3 + 0.25 = 7/12$	1

Example: (Please try it on your own)

A jar contains a large number of balls, each of which is either large or small and either blue or red. In this jar, 35% of the balls are large, 75% of the balls are blue, and 15% are both small and red. If one ball is randomly sampled from this container, what is the probability that it is a large blue ball?

- A) 0.500
- B) 0.350
- C) 0.100
- D) **0.250**
- E) 0.125

		Size		Total
		Large	Small	
Color	B	$0.35 - 0.1 = 0.25$	$0.65 - 0.15 = 0.5$	0.75
	R	$0.25 - 0.15 = 0.1$	0.15	$1 - 0.75 = 0.25$
		0.35	$= 1 - 0.35 = 0.65$	1

Example: (Please try it on your own)

There is a test for a rare disease. The test gives a positive result with probability 0.997 when the antibody is present. When the blood has no antibody, the test gives a positive result with probability 0.015. Suppose that only 1% of population carries the antibody. Given a positive test result, how likely is it that a person has the antibody?

Let $+$ = positive test result

A = person has the antibody

We know: $P(A) = 0.01$, $P(+|A) = 0.997$, $P(+|A^c) = 0.015$

We need to find: $P(A|+) = P(A \cap +)/P(+)$, where

$P(A \cap +) = P(+|A)P(A) = 0.997 \times 0.01 = 0.00997$, and

$$\begin{aligned} P(+) &= P(A \cap +) + P(A^c \cap +) = 0.00997 + P(+|A^c)P(A^c) \\ &= 0.00997 + 0.015 \times 0.99 = 0.02482 \end{aligned}$$

Thus, $P(A|+) = P(A \cap +)/P(+) = 0.00997/0.02482 = 0.4017$

Condition: Antibody

		Test Result		Total
		+	-	
Antibody	A	0.997	$1-0.997 = 0.003$	1
	A^c	0.015	$1-0.015 = 0.985$	1

		Test Result		Total
		+	-	
Antibody	A	0.997×0.01 $= 0.00997$	$0.01 - 0.00997$ $= 0.00003$	0.01
	A^c	0.015×0.99 $= 0.01485$	$0.99 - 0.01485$ $= 0.97515$	$1 - 0.01$ $= 0.99$
		$0.00997 + 0.01485$ $= 0.02482$	$0.00003 + 0.97515$ $= 0.97518$	1

Ch 13 Random Variables

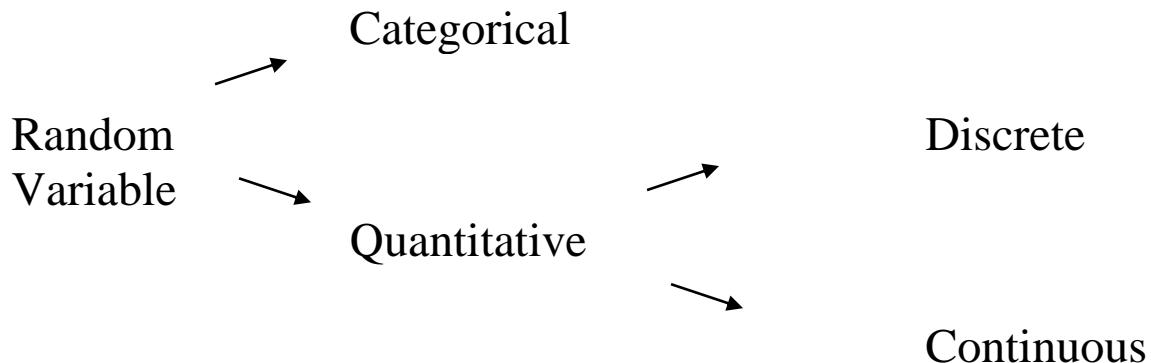
A variable X is a ***random variable*** (rv) if its value depends on the outcome of a random event.

- we use a capital letter, like X , to denote a random variable
- A particular value of a random variable will be denoted with a lower case letter, in this case x .

Example:

- X = number of observed "Tail" while tossing a coin 10 times
- X = survival time after specific treatment of a randomly selected patient
- X = SAT score for a randomly selected college applicant

Similar as for variables in sample data, rvs can be categorical or quantitative, and if they are quantitative, they can be either discrete or continuous.



Similar to data description, the models for rvs depend entirely on the type the rv. The models for continuous rvs will be different than those for discrete rvs.

Discrete random variables can take one of a finite number of distinct outcomes.

Example:

- the number of stores in a shopping mall
- the number of cars owned by a family
- the number of luggages each traveler carries in the airport

Continuous random variables can take any numeric value within an interval of values.

Example:

- Cost of books this term
- Height of football players

Definition:

- A **probability model** for a random variable consists of:
 - o The collection of all possible values of a random variable, and

- the probabilities that the values occur.

Value of X	x_1	x_2	x_3	...	x_n
Probability	$P(x_1)$	$P(x_2)$	$P(x_3)$...	$P(x_n)$

Properties of discrete probability distributions:

- $0 \leq P(x_i) \leq 1$
- $\sum P(x_i) = 1$

Example:

Toss two unbiased coins and let X equal the number of heads observed. Construct a probability distribution for X .

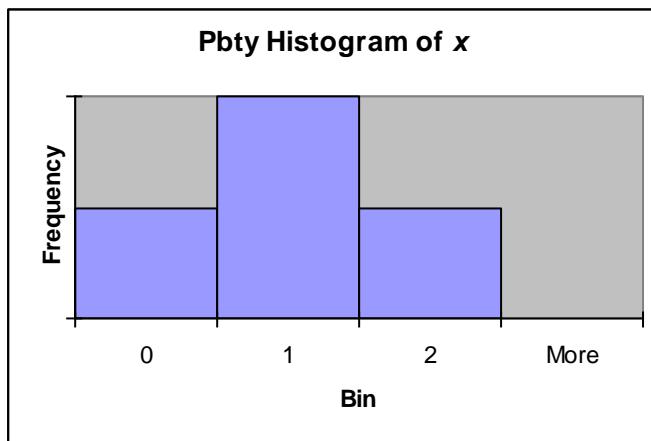
The simple events of this experiment are:

coin1	coin 2	X	P(X)
-------	--------	---	----------

So that we get the following distribution for $X = \text{number of heads observed}$:

<u>X</u>	<u>$P(X)$</u>
0	1/4
1	1/2
2	1/4

With the help of this distribution, find $P(X \leq 1)$.



Example

Rosana is planning whether she should continue to use myStatLab in her intro statistics class next Semester. She asked her current statistics students whether they like MyStatLab, and she found that _____ % of students like MyStatLab. Two students are randomly selected from this class. Let X denote the number of

students in this sample who likes MyStatLab. Develop the probability distribution of X .

Solution:

$Y = \text{the student selected likes MyStatLab}$

$N = \text{the student selected dislikes MyStatLab}$

So that we get the following distribution for X :

<u>X</u>	<u>$P(X)$</u>
0	
1	
<u>2</u>	

Expected Value: Center

The **expected value** $E(X)$ or **population mean** μ (mu) of a rv X is the value that you would expect to observe on average if the experiment is repeated over and over again. It is the center of the distribution.

Definition:

Let X be a discrete rv with probability distribution $P(X)$. The *population mean* or *expected value* of X is given as

$$\mu_X = \mu = E(X) = \sum x_i P(x_i).$$

In other words, the expected value of a (discrete) random variable can be found by summing the products of each possible value and the probability that it occurs.

NOTE:

- Be sure that every possible outcome is included in the sum
- Verify that you have a valid probability model to start with.

Example 1:

Find the expected value of the distribution of X = the number of heads observed tossing two coins.

$$\mu = E(X) =$$

Example 2:

Consider the *MyStatLab* example again. Let X be the number of students who like *MyStatLab* in a sample of two students. Find the expected value of the distribution of X .

$$\mu = E(X) =$$

Example 3:

A wheel comes up green 50% of the time and red 50% of the time. If it comes up green, you win \$100, if it come up red you win nothing. Intuitively, how much do you expect to win on one spin, on average?

Example 4:

BatCo, a company that sells batteries, claims that 99.5% of their batteries are in working order. How many batteries would you expect to buy, on average, to find one that does not work?

Example 5:

A friend of yours plans to toss a fair coin 200 times. You watch the first 40 tosses, noticing that she got only 16 heads. But then you get bored and leave. If the coin is fair, how many heads do you expect her to have when she has finished the 200 tosses?

Example: (Please try it on your own)

Suppose that an investor must decide between two strategies.

- 1) There is a 20% chance to make a profit of \$1000, and 80% chance to lose \$100.
- 2) There is a 15% chance to make a profit of \$12000, and an 85% chance to lose \$2000. Which strategy is better?

Calculate the expected profit for each strategy:

$$E(\text{profit with strategy 1}) = (-100)(0.8) + (1000)(0.2) = 120$$

$$E(\text{profit with strategy 2}) = (-2000)(0.85) + (12000)(0.15) = 100$$

Therefore strategy 1 would be better in the long run.

Standard Deviation (Spread)

Let X be a discrete rv with probability distribution $P(X)$. The *population variance* σ^2 of X is

$$\sigma_X^2 = \sigma^2 = \text{Var}(X) = \sum (x_i - \mu)^2 P(x_i)$$

The *population standard deviation* σ (sigma) of a rv X is equal to the square root of its variance.

$$\sigma = \sqrt{\sigma^2} = \sqrt{Var(X)} .$$

Example 1 (con't):

Find the population variance and standard deviation of $X =$ number of heads observed tossing two coins.

$$\sigma^2 = Var(X)$$

$$\sigma = SD(X)$$

Example 2:

Consider the *MyStatLab* example again. Let X be the number of students who like *MyStatLab* in a sample of two students. Find the population variance and standard deviation of X .

$$\sigma^2 = Var(X)$$

$$\sigma = SD(X)$$

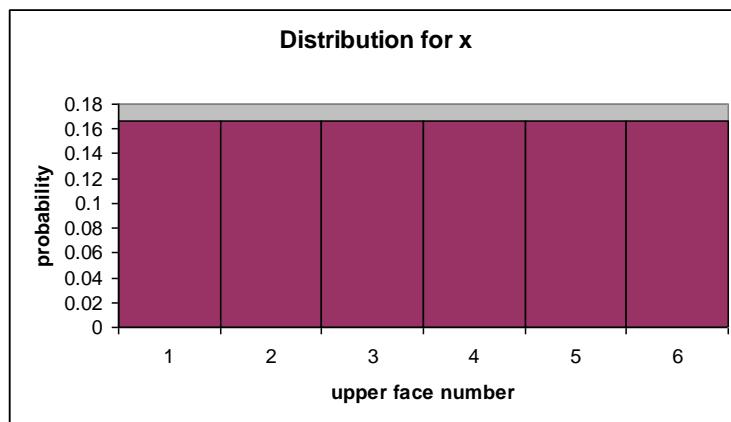
Example (Please try it on your own):

Consider tossing an unbiased dice and recording the number on upper face X . Find the expected value, variance and standard deviation of the distribution of X .

$$\mu = E(X) = \sum x_i P(x_i)$$
$$= 1(1/6) + 2(1/6) + 3(1/6) + 4(1/6) + 5(1/6) + 6(1/6) = 3.5$$

$$\sigma^2 = (1 - 3.5)^2 1/6 + (2 - 3.5)^2 1/6 + (3 - 3.5)^2 1/6 +$$
$$(4 - 3.5)^2 1/6 + (5 - 3.5)^2 1/6 + (6 - 3.5)^2 1/6$$
$$= 2.91666$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{2.91666} = 1.7078$$



More About Means and Variances

- Adding or subtracting a constant from data shifts the mean but doesn't change the variance or standard deviation:

$$E(X \pm c) = E(X) \pm c \quad \text{Var}(X \pm c) = \text{Var}(X)$$

Example: The average midterm mark for this Statistics Class is _____ with a standard deviation of _____.

Consider everyone in this class receives an extra 5%. What will be the mean and standard deviation of the average midterm mark after the increase in mark?

Let X = original mark and S = new mark, then $S = X + 5\%$

- Multiplying each value of a random variable by a constant multiplies the mean by that constant and the variance by the *square* of the constant:

$$E(aX) = aE(X) \quad Var(aX) = a^2Var(X)$$

Example: The average midterm mark for this Statistics Class is _____ with a standard deviation of _____.

Consider everyone in this class receives a 5% increase in marks.

For example, if a student got 60% on his exam, his new mark would be $60\% \times 1.05 = 63\%$. What will be the mean and standard deviation of the average midterm mark after the increase in mark?

Let X = original mark and S = new mark, then $S = 1.05 X$

Example: (Please try it on your own)

The average monthly income for ABC company's employees is \$5830 with a standard deviation of \$8620. Consider everyone in ABC company receiving a \$5000 increase in salary. What will be the mean and standard deviation of the monthly income after the increase in salary?

Let X = original salary, S = new salary, then $S = X + \$5000$

$$E(S) = E(X + 5000) = E(X) + 5000 = 5830 + 5000 = 10830$$

$$\text{Var}(S) = \text{Var}(X + 5000) = \text{Var}(X) = \$8620^2$$

$$\text{SD}(S) = \$8620 \quad \textit{No change in SD.}$$

Example: (Please try it on your own)

The average monthly income for ABC company's employees is \$5830 with a standard deviation of \$8620. Consider everyone in ABC company receiving a 10% increase in salary. What will be the mean and standard deviation of the monthly income after the increase in salary?

Let X = original salary, S = new salary, then $S = 1.1 X$

$$E(S) = E(1.1X) = 1.1E(X) = 1.1(5830) = 6413$$

$$\text{Var}(S) = \text{Var}(1.1X) = 1.1^2 \text{Var}(X) = 1.1^2 * \$8620^2$$

$$\text{SD}(S) = \$9482$$

Two Random Variables

- The mean of the sum (difference) of two random variables is the sum (difference) of the means.

$$E(X \pm Y) = E(X) \pm E(Y)$$

Thus, $E(aX \pm bY) = E(aX) \pm E(bY) = aE(X) \pm bE(Y)$

- If the random variables are *independent*, the variance of their sum *or* difference is always the sum of the variances.

$$\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y)$$

Thus, $\text{Var}(aX \pm bY) = \text{Var}(aX) + \text{Var}(bY)$

$$= a^2\text{Var}(X) + b^2\text{Var}(Y)$$

- NOTE:
 - Variances of independent random variables add.
Standard deviations don't.
 - The mean of the sum or difference of two random variables, discrete or continuous, is just the sum or difference of their means.
 - And, for *independent random variables*, the variance of their sum or difference is always the *sum* of their variances.

Example:

Given X is the study time for Andy each day and Y is the study time for Benny each day. Assume X and Y are independent random variables, and their means and standard deviations as shown:

	Mean	SD
X	5	3
Y	10	4

Find the mean and standard deviation of:

- a) Andy's study time if he doubled his study time each day, ie.
 $2X$
- b) Benny uses to study each day if he decided to increase his hours of studying by 6 hours each day, ie. $Y + 6$
- c) The difference between Andy and Benny's study time each day, ie. $X - Y$
- d) The total study time for Andy over 2 days, ie. $X_1 + X_2$

Example:

Some marathons allow two runners to “split” the marathon by each running a half marathon. Alice and Sharon plan to “split” a marathon. Alice’s half marathon times average 92 minutes with a standard deviation of 4 minutes, and Sharon’s half-marathon times average 96 minutes with a standard deviation of 2 minutes.

Assume that the women’s half marathon times are independent.

Let T = total time to complete a full marathon

X_A = time for Alice to complete half marathon

X_S = time for Sharon to complete half marathon

What is the expected time for Alice and Sharon to complete a full marathon?

What is the standard deviation of the total time?

If T is normally distributed, what is the probability that it will take Alice and Sharon more than 197 minutes to complete the full marathon?

Example:

Assume the heights of high school basketball players are normally distributed. For boys the mean is 74 inches with a standard deviation of 4.5 inches, while girl players have a mean height of 70 inches and standard deviation 3 inches. At a mixed 2-on-2 tournament teams are formed by randomly pairing boys with girls as teammates.

Let D = height difference between boy and girl = $X_b - X_g$

On average, how much taller do you expect the boy to be?

What will the variance be?

Example:

A coke filling machine is supposed to fill cans of coke with 12 fluid ounces. Suppose each fill is independent and that the fills are normally distributed with a mean of 12.1 oz and a standard deviation of .2 oz.

Let X stands for the amount of cola in a randomly selected bottle, and X_1, \dots, X_6 be the amounts in the six bottles and define:

- the total: Total = $X_1 + \dots + X_6$

- the average: $\bar{X} = (X_1 + \dots + X_6)/6$.

- a) What is the expected value of the TOTAL contents in a 6-pack of coke?

$$\begin{aligned} E(\text{Total}) &= E(X_1 + \dots + X_6) = E(X_1) + \dots + E(X_6) \\ &= 12.1 + \dots + 12.1 = 6 \times 12.1 = 72.6 \end{aligned}$$

- b) What is the standard deviation of the TOTAL contents in a 6-pack of coke?

$$\begin{aligned} \text{Var}(\text{Total}) &= \text{Var}(X_1 + \dots + X_6) = \text{Var}(X_1) + \dots + \text{Var}(X_6) \\ &= 0.2^2 + \dots + 0.2^2 \\ &= 6 \times 0.2^2 \end{aligned}$$

$$\text{SD}(\text{Total}) = 0.4899$$

- c) Find the expected value of the average contents in a 6-pack of coke.

$$E(\bar{X}) = E[(X_1 + \dots + X_6)/6] = (E(X_1) + \dots + E(X_6))/6 = 12.1.$$

- d) Find the standard deviation of the average contents in a 6-pack of coke.

$$\begin{aligned} \text{Var}(\bar{X}) &= \text{Var}[(X_1 + \dots + X_6)/6] \\ &= \frac{1}{6^2} (\text{Var}(X_1) + \dots + \text{Var}(X_6)) \\ &= (6 \times 0.2^2)/6^2 = 0.2^2/6 \end{aligned}$$

$$SD(\bar{X}) = \sqrt{\text{Var}(\bar{X})} = \frac{0.2}{\sqrt{6}}$$

Example:

Bolts are produced by a certain process have a width that is normally distributed with mean 1cm and standard deviation 0.01 cm. The nuts have mean width of 1.01cm and standard deviation 0.01 cm. What is the probability that a randomly chosen pair fit together?

Let X be the width of the nut, and Y the width of the bolt. The chance that they fit is $P(X \geq Y) = P(X - Y \geq 0)$

The random variable $X - Y$ has normal distribution with mean

$$E(X - Y) = 1.01 - 1.00 = 0.01 \text{ and}$$

$$\text{SD}(X - Y) = \sqrt{0.01^2 + 0.01^2} = \sqrt{2}(0.01).$$

Thus, $P(X - Y \geq 0) = P(Z \geq (0 - 0.01)/\sqrt{2}(0.01)) = P(Z \geq -1/\sqrt{2}) = 0.76025$

NOTE: Combining Random Variables (The Bad News)

- It would be nice if we could go directly from models of each random variable to a model for their sum.
- But, the probability model for the sum of two random variables is *not* necessarily the same as the model we started with *even when the variables are independent*.

- Thus, even though expected values may add, the probability model itself is different.

Combining Random Variables (The Good News)

- Nearly everything we've said about how discrete random variables behave is true of continuous random variables, as well.
- When two independent continuous random variables have Normal models, so does their sum or difference.
- This fact will let us apply our knowledge of Normal probabilities to questions about the sum or difference of independent random variables.

14 Sampling Distribution Models

Definition:

1) **Population parameter** is a numerical measure such as the mean, median, mode, range, variance, or standard deviation calculated for a population data; and is written with Greek letters. Eg. μ and σ .

- Usually unknown and constant

2) **Sample statistic** is a summary measure calculated for a sample data set; it is written with Latin letters. Eg. \bar{y} , s

- Regarded as random before sample is selected
- Observed after sample is selected

3) The value of the statistic varies from sample to sample, this is called **sampling variability**.

4) The distribution of all the values of a statistic is called its **sampling distribution**.

Population and Sample Proportions

Suppose we are just interested in one characteristic occurred in the population of interest. For convenience, we will call the outcome we are looking for “Success” (S). The **population proportion**, p ,

is obtained by taking the ratio of the number of successes in a population to the total number of elements in the population.

Example for population proportion:

- check for N students, how many are "nonresidents".
- check how many out of N patients survived at least five years, after a specific cancer treatment.

Recall: N is population size.

Looking at a SRS of the size n from a large population, the probability p can be estimated by calculating the sample proportion (relative frequency) of Successes, \hat{p} . That is,

$$\hat{p} = \frac{\text{number of Successes(S) in the sample}}{\text{sample size}} .$$

Example for sample proportion:

- flip n coins and observe if “Tail” was tossed.
- look at n random persons and survey how many have an IQ above 120.
- look at n random students and survey how many have more than two siblings.

Example:

Suppose a total of 10,000 patients in a hospital and 7,000 of them like to play basketball. A sample of 200 patients is selected from this hospital, and 128 of them like to play basketball. Find the proportion of patients who like to play basketball in the population and in the sample.

Find the sampling error for this case while assuming that the sample is random and no nonsampling error has been made.

Sampling error =

14.1 The Sampling Distribution of a Sample Proportion (\hat{p})

- Consider two different samples from a population, which you want to use for estimating the proportion of people with more than 2 siblings in the population. Use the statistic "proportion" for both samples. Are the outcomes the same?
- Most likely not! This is known as *sampling variability*.
- *imagine* we draw many samples and look at the sample proportions for these samples.

- The histogram we'd get if we could see *all the proportions from all possible samples* is called the **sampling distribution** of the proportions.
- What would the histogram of all the sample proportions look like?
- We would expect the histogram of the sample proportions to center at the true proportion, p , in the population.
- RULE 1: $\mu_{\hat{p}}$ is the **mean** of the sampling distribution of \hat{p} equals p :

$$\mu_{\hat{p}} = p$$

Notation: $\mu_{\hat{p}}$ is also denoted as $\mu(\hat{p})$

- RULE 2: The **standard deviation** of the sampling distribution of \hat{p} , $\sigma_{\hat{p}}$, is:

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

Notation: $\sigma_{\hat{p}}$ is also denoted as $SD(\hat{p})$.

NOTE: the standard deviation of a sampling distribution is called a **standard error**. Thus, $SD(\hat{p})$ is called a standard error.

- What about the shape of the sampling distribution of the proportions?
- It turns out that the histogram is unimodal, symmetric, and centered at p .
- More specifically, the sampling proportion \hat{p} is approximately normal distributed for large n . (Central Limit Theorem)

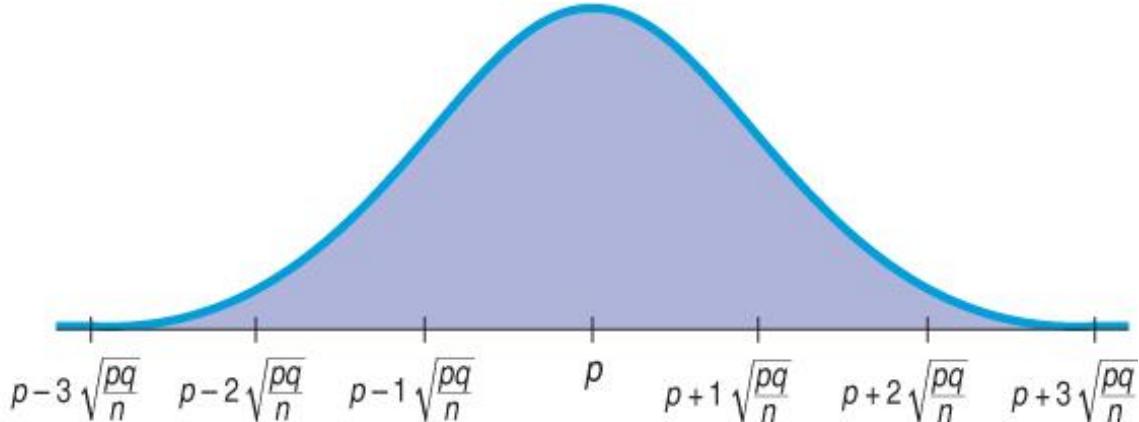
RULE 3: A rule of thumb states that the sample size is considered to be sufficiently large if:

$$np \geq 10 \text{ and } n(1 - p) \geq 10$$

NOTE: When n is large and p is not too close to 0 or 1, the sampling distribution of \hat{p} is approximately normal. The further p is away from 0.5, the larger n must be for accurate normal approximation of \hat{p} .

- A **sampling distribution model** for how a sample proportion varies from sample to sample allows us to quantify that variation and how likely it is that we'd observe a sample proportion in any particular interval.
- So, the distribution of the sample proportions is modeled with

a probability model that is $N\left(p, \sqrt{\frac{pq}{n}}\right)$



- Because we have a Normal model, for example, we know that 95% of Normally distributed values are within two standard deviations of the mean. So we should not be surprised if 95% of various polls gave results that were near the mean but varied above and below that by no more than two standard deviations.

Example: Which Brand of Pizza Do You Prefer?

- Two Choices: A or D.
- Assume that half of the population prefers Brand A and half prefers Brand D.
- Take a random sample of $n = 3$ tasters.

Find the sampling distribution for the sample proportion. Find the mean and standard deviation.

Sample	No. Prefer Pizza A	Proportion
(A,A,A)	3	1
(A,A,D)	2	2/3
(A,D,A)	2	2/3
(D,A,A)	2	2/3
(A,D,D)	1	1/3
(D,A,D)	1	1/3
(D,D,A)	1	1/3
(D,D,D)	0	0

Sample Proportion	Probability
0	1/8
1/3	3/8
2/3	3/8
1	1/8

14.2 Assumptions and Conditions

- Most models are useful only when specific assumptions are true.
- There are two assumptions in the case of the model for the distribution of sample proportions:
 1. **The Independence Assumption:** The sampled values must be independent of each other.
 2. **The Sample Size Assumption:** The sample size, n , must be large enough.
- Assumptions are hard—often impossible—to check. That's why we *assume* them.
- Still, we need to check whether the assumptions are reasonable by checking *conditions* that provide information about the assumptions.
- The corresponding conditions to check before using the Normal to model the distribution of sample proportions:
 - **Randomization Condition:** The sample should be a simple random sample of the population.
 - **10% Condition:** If sampling has not been made with replacement, then the sample size, n , must be no larger than 10% of the population.

- **Success/Failure Condition:** The sample size has to be big enough so that both np and nq are at least 10.

With these sampling distributions, we can apply standardization in order to find the probability for the proportion of Successes. The **z value for a value of \hat{p}** is:

$$z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}},$$

which is well approximated by the standard normal distribution.

Example:

A study showed that the proportion of people in the 20 to 34 age group with an IQ (on the Wechsler Intelligence Scale) of over 120 is about 0.35.

Let \hat{p} = proportion of the sample with an IQ of at least 120.

a) Find the mean and standard deviation of sample proportion

- b) What can you say about the distribution of sample proportion?
- c) Find the probability for the event that in a sample of 50 there are more than 30 people with an IQ of at least 120.

Example:

In an experiment, 32 subjects made a total of 60,000 guesses on a set of 5 symbol cards.

Pure chance would give around 12,000 correct guesses, but the subjects had a total of 12,489 correct guesses.

- a) Find the mean and standard deviation of the sample proportion of correct guesses.
- b) What can you say about the distribution of sample proportion of correct guesses?
- c) Could this excess of 489 good guesses just be good luck? In other words, calculate the probability for the event that in the total guesses, there are more than 12489 correct guesses.

Example: (Please try it on your own)

Suppose that the true proportion of people who have failed a professional exam is 0.87. A sample consists of 158 people is randomly drawn.

- a) Find the mean and standard deviation of the sample proportion of people failed to pass the professional exam?

$$\mu_{\hat{p}} = 0.870 \quad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.87(1-0.87)}{158}} = 0.0268$$

- b) What can you say about the distribution of sample proportion?

$$np = 158(0.87) = 137.46 \quad n(1-p) = 158(1-0.87) = 20.54$$

Since both values are > 15 , the distribution of \hat{p} should be well approximated by a normal curve.

- c) Find the probability that the sample proportion of people failed to pass the professional exam exceeds 0.94.

$$\begin{aligned} P(\hat{p} > 0.94) &\rightarrow P\left(\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} > \frac{0.94 - 0.87}{0.0268}\right) = P(Z > 2.64) \\ &= 1 - P(Z < 2.64) = 1 - 0.9959 = 0.0041 \end{aligned}$$

14.3 The Sampling Distribution of Other Statistics

What about Quantitative Data?

- Proportions summarize categorical variables.
- The Normal sampling distribution model looks like it will be very useful.
- Can we do something similar with quantitative data?
- We can indeed. Even more remarkable, not only can we use all of the same concepts, but almost the same model.

Example:

The population is a class of 5 Stat 151 students. Let μ be the population mean of their weight. Select a random sample of size 3 and observe the average weight \bar{y} . We must be careful to distinguish this number \bar{y} from μ .

How can \bar{y} , based on a sample of a small percentage of the class, be an accurate estimate of μ ? After all, a second sample would give a different value of \bar{y} .

This basic fact is called **sampling variability** (the value of a statistic varies in repeated random sampling).

Now suppose you look at every possible random sample from this Stat 151 population and the corresponding sample mean. For these numbers, you can create the **sampling distribution**.

Population Data

Student	Weight
A	70
B	75
C	75
D	75
E	80

All possible samples of size 3

Possible Samples	Weight in the sample	\bar{y}
ABC	70, 75, 75	73.3333
ABD	70, 75, 75	73.3333
ABE	70, 75, 80	75
ACD	70, 75, 75	73.3333
ACE	70, 75, 80	75
ADE	70, 75, 80	75
BCD	75, 75, 75	75
BCE	75, 75, 80	76.6667
BDE	75, 75, 80	76.6667
CDE	75, 75, 80	76.6667

NOTE: The total number of samples: ${}_5C_3 = 10$

Freq and Rel. Freq Dist^{ns} of \bar{y} when the sample size is 3

\bar{y}	Freq	Weight in the sample
73.3333	3	3/10
75	4	4/10
76.6667	3	3/10

Sampling Distribution of \bar{y} when the sample size is 3

\bar{y}	$P(\bar{y})$
73.3333	.3
75	.4
76.6667	.3
Total	1

NOTE: $\mu = 75$

$$\mu_{\bar{y}} = 73.3333 \times .3 + 75 \times .4 + 76.6667 \times .3 = 75 = \mu$$

Remark:

1. The value of \bar{y} differs from one random sample to another (sample variability).

2. Some samples produced \bar{y} values larger than μ , whereas other produce \bar{y} smaller than μ .
3. They can be fairly close to the mean μ , or also quite far off from the population mean μ (even it rarely happens).

A histogram of all these different \bar{y} values would give some insight into the accuracy of this estimation procedure.

Consider another Stat 151 class consists of 50 students. How many different samples of size of 3 we can take from a total of 50 students, we have ${}_{50}C_3 = 19600$, and this process is very cumbersome. Fortunately, there are mathematical theorems that help us to obtain information about the sampling distributions.

The Sampling Distribution of a Sample Mean

If \bar{y} is the sample average of an SRS of size n drawn from a population with mean μ and standard deviation σ , then the sampling distribution of \bar{y} has a mean of μ and a standard deviation of $\frac{\sigma}{\sqrt{n}}$.

NOTATION:

1. The population mean of \bar{y} , denoted $\mu(\bar{y})$, is equal to μ .

$$\mu(\bar{y}) = \mu$$

2. The population standard deviation of \bar{y} , denoted $SD(\bar{y})$, is

$$SD(\bar{y}) = \frac{\sigma}{\sqrt{n}},$$

where σ is the population standard deviation

NOTE: The standard deviation of the sampling distribution declines *only* with the square root of the sample size (the denominator contains the square root of n).

Therefore, the variability decreases as the sample size increases.

Now that we learned about the mean and the standard deviation of the sampling distribution of a sample mean, we might ask, “do we know anything about the shape of the density curve of this distribution?”

The following section explains why the normal distribution is so important in statistics.

Sampling from a Normally Distributed Population

If random samples of n observations are drawn from any normal distributed population with mean μ and standard deviation σ ($y \sim N(\mu, \sigma)$), then the sampling distribution of the mean \bar{y} is

normal distributed, with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$

$$(\bar{y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)).$$

What if the distribution of y is not normal?

Central Limit Theorem (CLT):

If random samples of n observations are drawn from any population with mean μ and standard deviation σ ($y \sim ?(\mu, \sigma)$), then for large n , the sampling distribution of the mean \bar{y} is normal distributed, with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$

$$(\bar{y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)).$$

Basically, it states that under rather general conditions, means of random samples drawn from one population tend to have an

approximately normal distribution. We find that it does not matter which kind of distribution we find in the population, it even can be discrete or extremely skewed, but if n is *large enough* the distribution of the mean is approximately normal distributed. That is, under all the possible distributions, we find one family of distributions that describes approximately the distribution of a sample mean, if only n is large enough.

Remarks:

- 1) If the *original* population is normal, then \bar{y} is exactly normal distributed for any value of n , so that n does not have to be large.
- 2) When the sampled population has a symmetric distribution, the sampling distribution of \bar{y} becomes quickly normal.
- 3) If the distribution is skewed, usually for $n = 30$ the sampling distribution is already close to a normal distribution.

Assumptions and Condition:

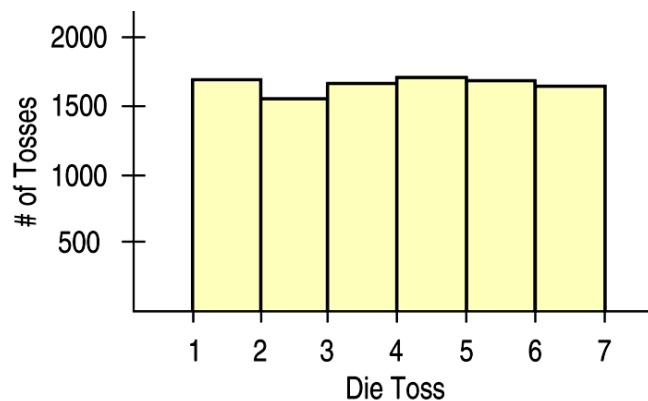
- The CLT requires essentially the same assumptions we saw for modeling proportions:

- Independence Assumption: The sampled values must be independent of each other.
 - Sample Size Assumption: The sample size must be sufficiently large.
- We can't check these directly, but we can think about whether the Independence Assumption is plausible. We can also check some related conditions:
- Randomization Condition: The data values must be sampled randomly.
 - Large Enough Sample Condition: The CLT doesn't tell us how large a sample we need. For now, you need to think about your sample size in the context of what you know about the population.

Example:

Consider tossing n unbiased dice and recording the average number of the upper faces.

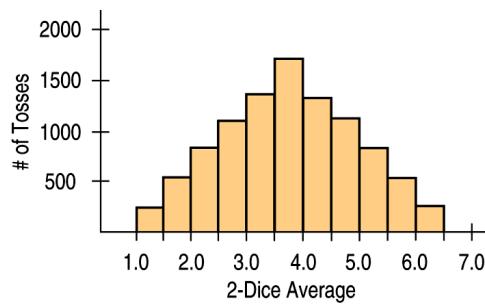
- Let's start with a simulation of 10,000 tosses of a die. A histogram of the results is:



The sampling distribution of the average of one die:

Number	1	2	3	4	5	6
Probability	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$

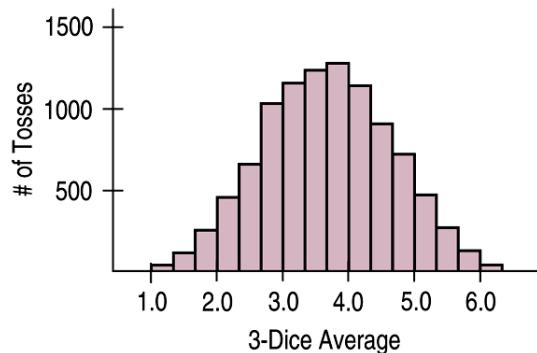
- Looking at the average of two dice after a simulation of 10,000 tosses:



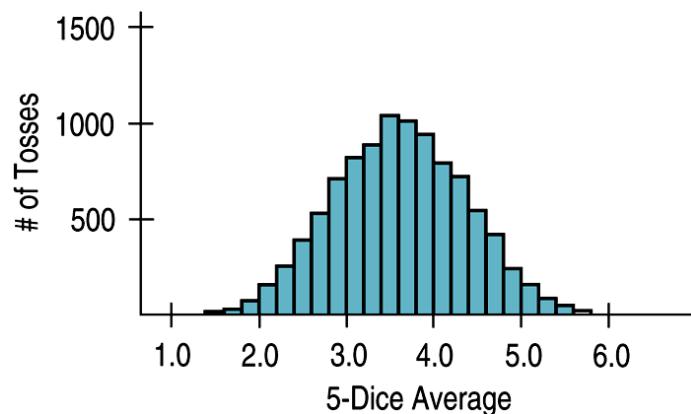
The sampling distribution of the average of two dice:

Sample Mean	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6
Probability	$1/36$	$2/36$	$3/36$	$4/36$	$5/36$	$6/36$	$5/36$	$4/36$	$3/36$	$2/36$	$1/36$

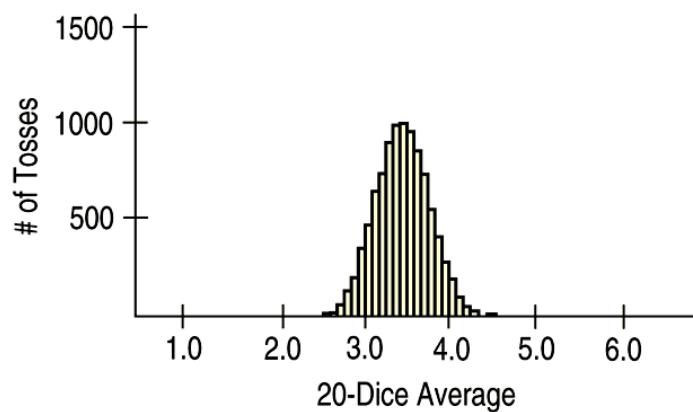
- The average of three dice after a simulation of 10,000 tosses looks like:



- The average of 5 dice after a simulation of 10,000 tosses looks like:



- The average of 20 dice after a simulation of 10,000 tosses looks like:



- As the sample size (number of dice) gets larger, each sample average is more likely to be closer to the population mean.
 - So, we see the shape continuing to tighten around 3.5
- And, it probably does not shock you that the sampling distribution of a mean becomes Normal.
- The sampling distribution of *any* mean becomes more nearly Normal as the sample size grows.
 - All we need is for the observations to be independent and collected with randomization.
 - We don't even care about the shape of the population distribution!
- The Fundamental Theorem of Statistics is called the Central Limit Theorem (CLT).
- The CLT is surprising and a bit weird:
 - Not only does the histogram of the sample means get closer and closer to the Normal model as the sample size grows, but *this is true regardless of the shape of the population distribution.*
- The CLT works better (and faster) the closer the population model is to a Normal itself. It also works better for larger samples.

With these sampling distributions, we can apply standardization in order to find the probability for the mean. The **z value for a value**

of \bar{y} is:
$$z = \frac{\bar{y} - \mu_{\bar{y}}}{\sigma_{\bar{y}}}$$

Example:

The scores of students on the ACT college entrance exam has a normal distribution with $\mu = 18.6$ and $\sigma = 5.9$.

- a) What is the probability that 1 randomly chosen student scores 21 or higher?

- b) Now take a mean of an SRS of 50 students who took the test. What are the mean and standard deviation of \bar{y} and describe the shape of its sampling distribution?

- c) What is the probability that the mean score \bar{y} is 21 or higher?

Example:

The duration of a disease from the onset of symptoms until death ranges from 3 to 20 years. The mean is 8 years and the standard deviation is 4 years.

Looking at the average duration for 30 randomly selected patients, calculate the mean and standard deviation of \bar{y} and describe the shape of its sampling distribution. What is the probability that the average duration of those 30 patients is less than 7 years?

Example:

A coke filling machine is supposed to fill cans of coke with 12 fluid ounces. Suppose that the fills are actually normally distributed with a mean of 12.1 oz and a standard deviation of .2 oz.

- a) What is the probability that an individual bottle contains less than 12 oz?

- b) What is the probability that the average fill for a 6-pack of coke is less than 12 oz?

- b) The coke bottler claims that only 5% of the coke cans are underfilled. A quality control technician randomly samples 200 cans of coke. What is the probability that more than 10% of the cans are underfilled?

Example: (Please try it on your own)

If the weight of individual eggs follow a normal distribution with mean $\mu = 65$ gram and standard deviation of $\sigma = 5$ gram. What is the probability that:

- a) a randomly selected egg will weigh between 63 and 66 grams?

Let y = the weight of an egg

$$\begin{aligned} P(63 < y < 66) &= P\left(\frac{63 - 65}{5} < \frac{y - \mu}{\sigma} < \frac{66 - 65}{5}\right) = P(-0.4 < z < 0.2) \\ &= P(z < 0.2) - P(z < -0.4) = 0.5793 - 0.3446 = 0.2347 \end{aligned}$$

- b) a dozen eggs will weigh between 756 and 792 grams?

Let total = total weight of a dozen eggs

$$\begin{aligned}
P(756 < \text{total} < 792) &= P\left(\frac{756}{12} < \frac{\text{total}}{12} < \frac{792}{12}\right) = P(63 < \bar{y} < 66) \\
&= P\left(\frac{63 - 65}{5/\sqrt{12}} < \frac{\bar{y} - \mu_{\bar{y}}}{\sigma_{\bar{y}}} < \frac{66 - 65}{5/\sqrt{12}}\right) = P(-1.39 < z < 0.69) \\
&= P(z < 0.69) - P(z < -1.39) = 0.7549 - 0.0823 = 0.6726
\end{aligned}$$

NOTE 1: this question is the same as asking “what is the probability of the average weight of a dozen eggs is between 63 and 66 grams?”

NOTE 2: Can we standardize total directly?

Answer: You can standardize total only with the mean of the total (μ_t) and the standard deviation of the total (σ_t), where

$$\mu_t = n \mu$$

$$\sigma_t^2 = n \sigma^2 \rightarrow \sigma_t = \sqrt{n} \sigma$$

Example (Please try this question on your own):

The number of accidents per week at a hazardous intersection varies with mean $\mu = 2.2$ and standard deviation $\sigma = 1.4$. This distribution takes only whole numbers, so it is certainly not normal

a) Let \bar{y} be the mean number of accidents per week at the intersection during a year. What is the approximate distribution of \bar{y} according to the CLT?

The population distribution is NOT normal, but we have a sample size of 52, so its sampling distribution is normal.

The CLT says that the distribution of \bar{y} is *approximately* normal with

$$\text{- mean of } \bar{y} : \mu_{\bar{y}} = \mu = 2.2$$

$$\text{- standard deviation of } \bar{y} : \sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{1.4}{\sqrt{52}} = 0.1941$$

b) What is the approximate probability that the average number of accidents per week is less than 2?

$$P(\bar{y} < 2) = P\left(\frac{\bar{y} - \mu_{\bar{y}}}{\sigma_{\bar{y}}} < \frac{2 - 2.2}{0.1941}\right) \text{ standardize}$$

$$= P(z < -1.03) = 0.1515 \text{ (Table z)}$$

c) What is the approximate probability that there are fewer than 100 accidents in a year?

NOTE: The average number of accidents is the total number

of accidents divided by 52. That is: $\bar{y} = \frac{\text{total}}{52}$

$$P(\text{total} < 100) = P(\bar{y} < 100/52)$$

$$= P\left(\frac{\bar{y} - \mu_{\bar{y}}}{\sigma_{\bar{y}}} < \frac{100/52 - 2.2}{0.1941}\right) \text{ standardize}$$

$$= P(z < -1.43) = 0.0764 \text{ (Table z)}$$

Where are we?

- Ch. 2–5: Data set and its distribution, statistics, Normal model.
- Ch. 9–10: Collecting data, random samples, randomized experiments.
- Ch. 11–13: Probability, random variables, probability distribution models.
- Ch. 14–24: Statistical inference: what does sample data tell us about the underlying population. Inferences about parameters (proportions, means, etc.) in a model for the population distribution.

Quick Review of Ch 14:

The Sampling Distribution of a Sample Proportion

Suppose we are just interested in one characteristic occurred in the population of interest. For convenience, we will call the outcome we are looking for “Success”. The **population proportion**, p , is obtained by taking the ratio of the number of successes in a population to the total number of elements in the population.

1) $\mu_{\hat{p}}$ is the **mean** of the sampling distribution of \hat{p} equals p :

$$\mu_{\hat{p}} = p$$

Notation: $\mu_{\hat{p}}$ is also denoted as $\mu(\hat{p})$

2) The **standard deviation** of the sampling distribution of \hat{p} ,

$\sigma_{\hat{p}}$, is:

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

Notation: $\sigma_{\hat{p}}$ is also denoted as $SD(\hat{p})$.

3) When n is large and p is not too close to 0 or 1, the sampling distribution of \hat{p} is approximately normal.

Rule of thumb: The sample size is considered to be sufficiently large if

$$np \geq 10 \text{ and } n(1-p) \geq 10$$

The Sampling Distribution of a Sample Mean

If \bar{y} is the sample average of an SRS of size n drawn from a population with mean μ and standard deviation σ , then:

1. The population mean of \bar{y} , $\mu_{\bar{y}}$, is equal to μ .

$$\mu_{\bar{y}} = \mu$$

NOTATION: textbook denoted $\mu_{\bar{y}}$ as $\mu(\bar{y})$

2. The population standard deviation of \bar{y} , denoted $\sigma_{\bar{y}}$, is

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}},$$

NOTATION: textbook denoted $\sigma_{\bar{y}}$ as $SD(\bar{y})$

3. If random samples of n observations are drawn from any normal distributed population with mean μ and standard deviation σ , then the sampling distribution of the mean \bar{y} is normal distributed, with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

4. CLT: If random samples of n observations are drawn from any population with mean μ and standard deviation σ , then for large n (ie. $n \geq 30$), the sampling distribution of the mean \bar{y} is normal distributed, with mean μ and standard deviation

$$\frac{\sigma}{\sqrt{n}}.$$

Ch 15 Confidence Intervals for Proportions

We will be starting now to cover inferential statistics! Its objective is to use sample data to obtain results about the whole population.

In a first step, the goal is to describe an underlying population. Since the populations are described in form of models that are characterized by parameters (mean μ and standard deviation σ or probability p for the event of interest).

At this time we will estimate those characteristics. There are two different approaches for estimating: Point Estimation and Interval Estimation.

For Point Estimation, you give one value for a characteristic, which is hopefully close to the true unknown value.

For Interval Estimation, you give an interval of likely values, where the width of the interval will depend on the confidence you require to have in this interval.

Since we base our statement just on a sample, we see later how to give a measure of accuracy or confidence for the estimate.

Point Estimation

A **point estimate** of a possible characteristic is a single number that is based on sample data and represents the population parameter.

Example:

- The *sample mean* \bar{y} is a point estimate for the *population mean* μ .
- The *sample proportion* \hat{p} is a point estimate of p the *population probability for Success*.

A point estimate gives a single value that is supposed to be close to the true value of the characteristic but it does *NOT* tell how close the estimate is.

Considering we know that we would observe different values of a point estimate from sample to sample, point estimates are not enough to describe a parameter. Thus, we introduce the second type of estimate – *interval estimate*.

15.1 A Confidence Interval

As an alternative to point estimation we can report not just a single value for the population characteristic, but an entire interval of reasonable values based on sample data. These intervals take into account of error and uncertainty. We often

associate interval estimate with some level of confidence and the result is called a *confidence interval*.

- Recall: Both of the sampling distributions for proportions and means are Normal.

- For proportions : $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

- For means: $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$

- When we don't know p or σ , we're stuck, right?
 - No, we will use sample statistics to estimate these population parameters.
 - Whenever we estimate the standard deviation of a sampling distribution, we call it a **standard error**.
 - The standard error for a sample proportion:

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

- The standard error for the sample mean: $SE(\bar{y}) = \frac{s}{\sqrt{n}}$

By the 68-95-99.7% Rule, when sampling distribution \hat{p} is approximately Normal, we know

- about 68% of all samples will have \hat{p} 's within 1 *SE* of p
 - about 95% of all samples will have \hat{p} 's within 2 *SEs* of p
 - about 99.7% of all samples will have \hat{p} 's within 3 *SEs* of p
- We can look at this from \hat{p} 's point of view...

Consider the $C = 95\%$ level:

- There's a 95% chance that p is no more than 2 *SEs* away from \hat{p} .
 - So, if we reach out 2 *SEs*, we are 95% sure that p will be in that interval. In other words, if we reach out 2 *SEs* in either direction of \hat{p} , we can be 95% confident that this interval contains the true proportion.
- This is called a **95% confidence interval**.

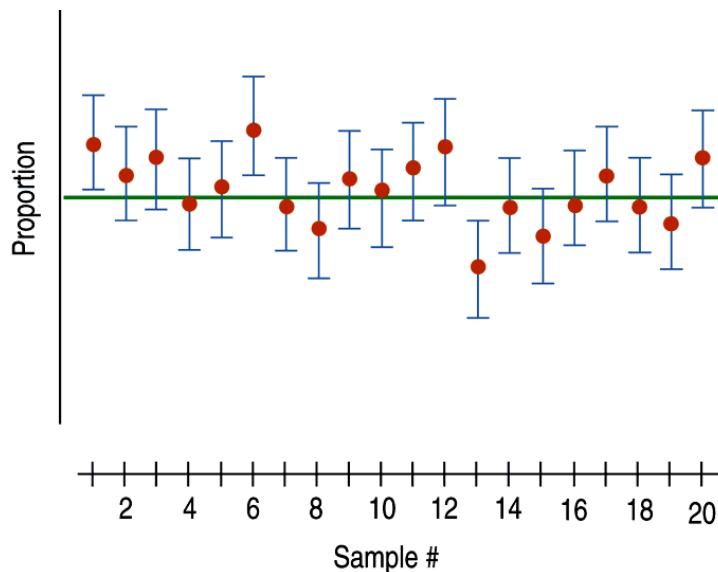
15.2 What does “95% Confidence” really mean?

Being "95% confident" means, if you were to construct 100 95% confidence intervals from 100 different samples. Of the 100

intervals, you expect 95 to capture the true parameter, and 5 not to capture the parameter.

How can this happen?

- Each confidence interval uses a sample statistic to estimate a population parameter, but since samples vary, the statistics we use, and thus the confidence intervals we construct, vary as well.
- In conclusion, you cannot be sure that a specific confidence interval captures the true proportion p .
- Our **confidence** is in the *process* of constructing the interval, not in any one interval itself.
- The following figure shows that some of our confidence intervals (from 20 random samples) capture the true proportion (the horizontal line), while others do not:



15.3 Margin of Error: Certainty vs. Precision

- Most confidence intervals are of the form:

Point estimate \pm margin of error

$$= \text{point estimate} \pm \text{critical value} \times \text{SE}(\text{estimate})$$

- The more confident we want to be, the larger our margin of error needs to be (makes the interval wider).
- We need more values in our confidence interval to be more certain.
- Because of this, every confidence interval is a balance between certainty and precision.
- The tension between certainty and precision is always there.
- Fortunately, in most cases we can be both sufficiently certain and sufficiently precise to make useful statements.
- The most commonly chosen confidence levels (C) are 90%, 95%, and 99% (but any percentage can be used).

Critical Values

- The critical value is how far we need to deviate from the estimate to capture the central $100C\%$ of the values on the sample distribution.

- The ‘2’ in $\hat{p} \pm 2SE(\hat{p})$ (our 95% confidence interval) came from the 68-95-99.7% Rule.
- Using a table or technology, we find that a more exact value for our 95% confidence interval is 1.96 instead of 2.
- We call 1.96 the critical value and denote it z^* .

Example: To find the central 95% region on a standard normal curve, you need to cut off 2.5% at each end.

The z^* value for $C = 0.95$ has 97.5% of the area to the left. Using z -table, we find $z^* = 1.96$.

- For any confidence level, we can find the corresponding critical value.
- Commonly used critical values

Confidence Coefficient C	1 – C	(1 – C)/2	z^*
0.90	0.1	0.05	1.645
0.95	0.05	0.025	1.96
0.99	0.01	0.005	2.58

Example: Show that for a 90% confidence interval, the critical value is 1.645.

Example:

Consider flipping an unbiased coin 1000 times. The results showed that you flipped 400 heads. Based on this result, what interval captures the most likely 95% of the values of the actual proportion of heads? Then check if the coin is fair.

15.4 A 100C% Large Sample Confidence Interval for a Population Proportion p .

Assumption:

- Here are the assumptions and the corresponding conditions you must check before creating a confidence interval for a proportion:
 - 1) Independence Assumption: You cannot check this by looking at the data. Instead, we check two conditions to decide whether independence is reasonable.
 - Randomization Condition: Were the data sampled at random or generated from a properly randomized experiment? Proper randomization can help ensure independence.
 - 10% Condition: Is the sample size no more than 10% of the population?

2) Sample Size Assumption: The sample needs to be large enough for us to be able to use the CLT.

- Success/Failure Condition: We must expect at least 10 “successes” and at least 10 “failures”
- When the conditions are met, we are ready to find the confidence interval for the population proportion, p .
- The confidence interval is
$$\hat{p} \pm z^* SE(\hat{p}) = \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$
- The critical value, z^* , depends on the particular confidence level, C , that you specify.

Example (cont):

To answer this question, we will calculate a 95% confidence interval from this data and check if 0.5 (the probability for HEAD, when tossing an unbiased coin) is in the confidence interval.

NOTE:

Do NOT say:

- There is a probability of 0.95 that p is between 0.37 and 0.43.
- Between 37% and 43% of all tosses appear as HEADS.
- 95% of all random samples of coin tosses will show between 37% and 43% of HEADS.
- We can be 95% confident that all random samples will show 40% of HEADS.
- There is a 95% chance that the true proportion of HEADS is between 37% and 43%.

Example:

For a project, a student randomly sampled 182 other students at a large university to determine if the majority of students were in favor of a proposal to build a field house. He found that 75 were in favor of the proposal.

a) Find the 95% confidence interval for p .

b) Find the 99% confidence interval for p .

Remark: In order to have a higher confidence, we need to accept a larger margin of error, ie. a wider interval.

c) Find the 95% confidence interval for p if this student randomly sampled 364 students at a large university and found that 148 were in favor of the proposal instead.

Example: Which of the following statements about a confidence interval is generally **true**?

- a. If you take random samples over and over again from the same population and make 95% confidence intervals for a population parameter, about 95% of the intervals should contain the population parameter.
- b. A 95% confidence interval obtained from a random sample of 1000 people has a better chance of containing the population parameter than a 95% confidence interval obtained from a random sample of 500 people.
- c. A 95% confidence interval obtained from a random sample of 500 people has a better chance of containing the population parameter than a 95% confidence interval obtained from a random sample of 1000 people.
- d. A 95% confidence interval is narrower than a 90% confidence interval.
- e. A 95% confidence interval obtained from a random sample of 10 people is wider than a 90% confidence interval obtained from a random sample of 5 people.

Example: A 95% confidence interval for a population proportion is (0.08, 0.09). Using the same sample, which of following statements is definitely **true**?

- a. A 90% C.I certainly includes 0.09.
- b. **A 90% C.I. certainly does not include 0.09.**
- c. A 90% C.I. may or may not include 0.09.
- d. A 99% C.I. certainly does not include 0.09.
- e. A 98% C.I. certainly does not include 0.08.

A Confidence Interval for Small Samples

- A simple adjustment to the confidence interval when the Success/Failure Condition fails.
- All we do is add four observations, two successes and two failures.
- So instead of $\hat{p} = \frac{y}{n}$, we use the adjusted proportion

$$\tilde{p} = \frac{y + 2}{n + 4}$$

- Now the adjusted interval is $\tilde{p} \pm z^* \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{n}}$
- The adjusted form gives better performance overall and works much better for proportions of 0 or 1.

Choosing the sample size

Recall: the margin of error in the CI for p is: $ME = z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

We may like to choose the sample size n to achieve a certain margin of error, so we solve for n :

$$n = \left(\frac{z^*}{ME} \right)^2 \hat{p}(1-\hat{p})$$

- \hat{p} is based on a pilot study or on past experience, but we may not have prior sampling done!
 - o use $\hat{p} = 0.50$. This is conservative as it gives a margin of error bigger than the true margin of error.

Example:

If a TV executive would like to find a 95% confidence interval estimate within 0.03 for the proportion of all households that watch NYPD Blue regularly. How large a sample is needed if a prior estimate for \hat{p} was 0.15?

Example (Revisited):

Suppose a TV executive would like to find a 95% confidence interval estimate within 0.03 for the proportion of all households that watch NYPD Blue regularly. How large a sample is needed if we have no reasonable prior estimate for \hat{p} ?

Example (Please try it on your own):

To conduct a political poll that is 99% sure of finding the level of support for the Conservative party to within 0.01 of margin of error, how large a sample would we need?

$$n = \left(\frac{z^*}{ME} \right)^2 p^*(1-p^*) = \left(\frac{2.576}{0.01} \right)^2 0.5(1-0.5) = 16589.44$$

Thus, to be 99% confidence of finding the level of support for the Conservative party to within 0.01, we need 16590 samples.

Ch 16 Testing Hypotheses About Proportions

Previously, population parameters were described, now we will be checking if claims about the population parameters are true, or plausible to a given degree.

Recall:

Background for a Large Sample Test concerning a Probability Proportion p

For developing a test again the facts we know from the CLT have to be considered. The point estimate for a probability of success is the sample proportion \hat{p} . From the CLT, we know about the sampling distribution of \hat{p} that:

1. $\mu_{\hat{p}} = p$
2. $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
3. If n is large, the sampling distribution of \hat{p} is approximately normal.

Example:

Suppose p is the probability to get a HEAD when flipping a coin. Eric claims that the coin is fair, but Amy wants to test if Eric's claim is true. She took a random sample of size 1000 showed 400 HEADS.

Based on this result, can we state that:

- i) the proportion of heads occurring is less than 0.5 (so Eric's claim is invalid)? OR
- ii) the difference between 0.5 (the population proportion of heads) and 0.4 (the sample proportion of heads) may have occurred because of sampling error?

In order to answer this question, we need to perform a hypothesis test.

Hypothesis Test

A **hypothesis test** is a method for using sample statistics to decide between two competing claims on hypotheses about a **population parameter**. It follows the following procedure:

- 1) Define the variable, the parameter(s) of interest, and any relevant assumptions.
- 2) State the null hypothesis H_0 and alternative hypothesis H_a .

- 3) Gather the evidence (sample). Based on the data in the sample, we will calculate a test statistic.
- 4) Assess the strength of the evidence against the null hypothesis in favor of the alternative. This will be done by finding p-value.
- 5) Make a decision based on Step 4.
- 6) State the conclusion.

State the hypotheses:

The **null hypothesis** H_0 is a claim about a *population parameter* that is assumed to be true until it is declared false. It is generally the hypothesis of “no effect.”

- We usually write down the null hypothesis in the form H_0 :
parameter = hypothesized value.

The **alternative hypothesis** H_a is a claim about a *population parameter* that will be true ONLY when we reject the null hypothesis. In another words, this is the hypothesis that we are trying to find evidence for.

Remark:

Common choices of hypotheses are:

- Two-tailed Test:
 - H_0 : population characteristic = specific value versus
 - H_a : population characteristic \neq specific value
- Upper-tailed Test:
 - H_0 : population characteristic = specific value versus
 - H_a : population characteristic $>$ specific value
- Lower-tailed Test:
 - H_0 : population characteristic = specific value versus
 - H_a : population characteristic $<$ specific value

Examples:

- $H_0: \mu = 100$ versus $H_a: \mu < 100$
- $H_0: p = 0.25$ versus $H_a: p \neq 0.25$
- We **cannot** test $H_0: p = 0.5$ versus $H_a: p > 0.6$
- We **cannot** test $H_0: \bar{y} = 100$ versus $H_a: \bar{y} \neq 100$
- We **cannot** test $H_0: \hat{p} = 0.65$ versus $H_a: \hat{p} < 0.65$

Example:

According to a poll, 40% of Canadians eat chocolate every day. Is this proportion higher for Edmontonians? In a random sample of 50 Edmontonians, 25 reported that they eat chocolate every day. Which type of hypothesis test would you use?

- A. One-tail upper tail
- B. One-tail lower tail
- C. Two-tail
- D. Both A and B

Example:

A statistics professor wants to see if more than 80% of her students enjoyed taking her class. At the end of the term, she takes a random sample of students from her large class and asks, in an anonymous survey, if the students enjoyed taking her class. Which set of hypotheses should she test?

- A. $H_0: p < 0.80$ $H_A: p > 0.80$
- B. $H_0: p = 0.80$ $H_A: p > 0.80$**
- C. $H_0: p > 0.80$ $H_A: p = 0.80$
- D. $H_0: p = 0.80$ $H_A: p < 0.80$

Example:

An online catalog company wants on-time delivery for 90% of the orders they ship. They have been shipping orders via UPS and FedEx but will switch to a new, cheaper delivery service (ShipFast) unless there is evidence that this service cannot meet the 90% on-time goal. As a test the company sends a random sample of orders via ShipFast, and then makes follow-up phone calls to see if these orders arrived on time. Which hypotheses should they test?

- A. $H_0: p < 0.90$ $H_A: p > 0.90$
- B. $H_0: p = 0.90$ $H_A: p > 0.90$
- C. $H_0: p > 0.90$ $H_A: p = 0.90$
- D. $H_0: p = 0.90$ $H_A: p < 0.90$**

Testing H_0 vs. H_a :

- H_0 will be rejected only if the sample evidence strongly suggests that H_0 is false.
- Otherwise H_0 will not be rejected.

So there are **two** possible conclusions:

- reject H_0 (accept H_a)
- do **not** reject H_0

Note: When H_0 is not being rejected, it doesn't mean strong support for H_0 , but lack of strong evidence for H_a . These decisions are not symmetric, there is *NO* way you can say you accept H_0 .

Idea: Compare the process to a criminal trial.

The fact is that a person accused of a crime is either guilty or not guilty.

- To prove someone is guilty, we start by *assuming* they are innocent.
 - H_0 : innocent
- We retain that hypothesis until the facts make it unlikely beyond a reasonable doubt.
 - H_a : guilty

Rejection of H_0 :

Nonrejection of H_0 :

Example (con't):

Suppose p is the probability to get a HEAD when flipping a coin. Eric claims that the coin is fair, but Amy wants to test if Eric's claim is true. She took a random sample of size 1000 and it showed 400 HEADS. State the hypothesis. Interpret rejection and nonrejection of H_0 for this example.

H_0 : versus H_a :

Rejection of H_0 :

Nonrejection of H_0 :

How to make the decision (reject H_0 or do not reject H_0)

The decision to reject, or not to reject H_0 is based on information contained in a sample drawn from the population of interest.

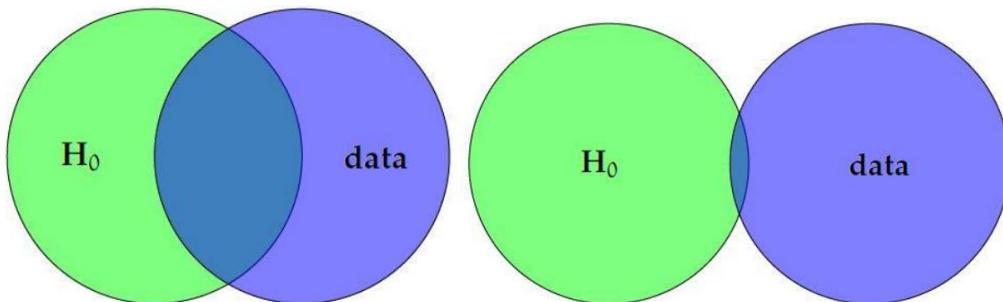
Use the sample to:

- Calculate a **test statistic** (a number that measures how many standard deviations away the estimate in the sample is from the hypothesized value of the parameter in H_0),

- **p-value:**
 - Use the value of the test statistic and its distribution to calculate the **p-value** (the probability of observing the value of the test statistic as extreme or more extreme than the one observed, if H_0 is true). In other words, we try to find out how likely the observed results could have happened if the null hypothesis were true.

In general:

- When the data are consistent with the model from the null hypothesis, *the P-value is high and we fail to reject the null hypothesis.*
- *If the P-value is low enough, we'll "reject the null hypothesis,"* since what we observed would be very unlikely were the null model true.



- $H_A: \text{parameter} \neq \text{value}$ (a two-sided alternative)
 - we are equally interested in deviations on either side of the null hypothesis value.

- For two-sided alternatives, the P-value is the probability of deviating in *either* direction from the null hypothesis value.
- The other two alternative hypotheses are called **one-sided alternatives**.
 - A one-sided alternative focuses on deviations from the null hypothesis value in only one direction.
 - Thus, the P-value for one-sided alternatives is the probability of deviating *only in the direction of the alternative* away from the null hypothesis value.

P-Values and Decisions:

- How small should the P-value be in order for you to reject the null hypothesis?
- It turns out that our decision criterion is context-dependent.
 - When we're screening for a disease and want to be sure we treat all those who are sick, we may be willing to reject the null hypothesis of no disease with a fairly large P-value.

- A longstanding hypothesis, believed by many to be true, needs stronger evidence (and a correspondingly small P-value) to reject it.
- Your conclusion about any null hypothesis should be accompanied by the P-value of the test.
 - If possible, it should also include a confidence interval for the parameter of interest.
- Don't just declare the null hypothesis rejected or not rejected.
 - Report the P-value to show the strength of the evidence against the hypothesis.
 - This will let each reader decide whether or not to reject the null hypothesis.

17.3 Alpha Levels

- Sometimes we need to make a firm decision about whether or not to reject H_0
- When the p-value is small, it tells that our data are rare given the H_0
- How rare is “rare”?

- If our P-value falls below a threshold point, we'll reject H_0 . We call such results *statistically significant*.
- The threshold is called an *alpha level*, denoted by α .
- Common alpha levels are 0.10, 0.05, and 0.01.
- You have the option—almost the *obligation*—to consider your alpha level carefully and choose an appropriate one for the situation.
- The alpha level is also called the *significance level*.

What does it mean to say a test is statistically significant?

- All it means is the test statistic had a P-value lower than our alpha level.
- Don't be lulled into thinking that statistical significance carries with it any sense of practical importance or impact.

17.5 Decision Errors

There are two different types of errors you can make in statistical testing.

Idea: Use the criminal trials to demonstrate this,

H_0 : the person is innocent

H_a : the person is guilty

In a trial, the jury might make the following mistake:

- a) convict an innocent person
- b) to set a guilty person free

		Reality	
		Not Guilty	Guilty
Court's decision	Not Guilty	OK	Type II error
	Guilty	Type I error	OK

Definition:

- a) **type I error** – the error of rejecting H_0 even though H_0 is true
- b) **type II error** – the error of failing to reject H_0 even though H_0 is false

		Truth	
		H_0 is true	H_0 is false
Test	Do not reject H_0	OK	Type II error
	Reject H_0	Type I error	OK

Definition:

- 1) **Probability of a type I error (α)** is the probability of rejecting H_0 even though H_0 is true.

$$\alpha = P(H_0 \text{ is rejected} \mid H_0 \text{ is true})$$

2) **Probability of a type II error (β)** is the probability of failing to reject H_0 even though H_0 is false.

$$\beta = P(H_0 \text{ is not rejected} \mid H_0 \text{ is false})$$

3) **Power of the test ($1 - \beta$)** is the probability of correctly rejecting a false H_0 .

$$1 - \beta = P(H_0 \text{ rejected} \mid H_0 \text{ is false})$$

Example:

The proportion of people surviving is 30% after a specific cancer treatment. A new treatment has been proposed to be effective. To show the efficacy of the new treatment, a study has to be designed to test the following hypotheses for p (the proportion of people surviving under the new treatment).

$$H_0: p = 30\%$$

versus

$$H_a: p > 30\%$$

Describe the type I and type II error.

Type I Error (error of rejecting H_0 even though it is true):

Type II Error (error of failing to reject H_0 even though H_0 is false):

If the scientist wants to make sure that this new treatment is only used if it is really improving the survival rate as it may cause some side effects, then would he choose $\alpha = 0.01$ or $\alpha = 0.05$?

He has to limit the probability for the type I error, so he chooses $\alpha = 0.01$.

Remark:

- a) The two types of errors that occur in the tests of hypotheses depend on each other. Lowering the value of α will increase the value of β , and vice versa.
 - It makes sense that the more we're willing to accept a Type I error, the less likely we will be to make a Type II error.
- b) The value of α is controlled by the experimenter.
- c) The value of β is difficult, if not impossible to calculate, because we don't know what the value of the parameter really is.
- d) The power increases with the effect size, where effect size $= | p - p_0 |$
 - The larger the effect size, the easier it should be to see it.

- Obtaining a larger sample size decreases the probability of a Type II error, so it increases the power.
- e) The only way to reduce *both* types of errors is to collect more data. Otherwise, we just wind up trading off one kind of error against the other.

NOTE: increase alpha → decrease beta → increase the power of the test

Example:

Suppose that a manufacturer is testing one of its machines to make sure that the machine is producing more than 97% good parts ($H_0: p = 0.97$ and $H_A: p > 0.97$).

- If the test results in a P-value of 0.102. In reality, the machine is producing 99% good parts. What probably happens as a result of our testing?
 - We correctly fail to reject H_0 .
 - We correctly reject H_0 .
 - We reject H_0 , making a Type I error.
 - We fail to reject H_0 , making a Type I error.
 - We fail to reject H_0 , making a Type II error.**

- b) We conclude that it is producing more than 97% good parts when it is not. What probably happens as a result of our testing?
- A. We correctly fail to reject H_0 .
 - B. We correctly reject H_0 .
 - C. We reject H_0 , making a Type I error.**
 - D. We reject H_0 , making a Type II error.
 - E. We fail to reject H_0 , committing Type II error.

One Proportion z-Test

1. Assumption & Conditions:

- Randomization
- Independence
- The sample size n is large, that is:

$$np_0 \geq 10 \text{ and } n(1 - p_0) \geq 10.$$

where p_0 comes from the hypotheses.

2. Determine the type of test

- a) two tailed:

a. $H_0 : p = p_0$ versus $H_a : p \neq p_0$

b) lower tailed:

$$\text{a. } H_0 : p = p_0 \quad \text{versus} \quad H_a : p < p_0$$

c) upper tailed:

$$\text{a. } H_0 : p = p_0 \quad \text{versus} \quad H_a : p > p_0$$

3. Test statistic:

Let p_0 be a value between zero and one, and define the test statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

4. p-value:

Test Type	p-value
Upper Tail	$P(z > z_o)$
Lower Tail	$P(z < z_o)$
Two Tails	$2 \times P(z > z_o)$ $= 2 \times P(z < - z_o)$

5. Decision: Reject H_0 , if and only if $p\text{-value} \leq \alpha$.

Example (con't):

Suppose p is the probability to get a HEAD when flipping a coin. Eric claims that the coin is fair, but Amy wants to test if Eric's claim is true. She took a random sample of size 1000 showed 400 HEADS.

Recall: The 95% CI:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{400}{1000} \pm 1.96 \sqrt{\frac{0.4 \cdot 0.6}{1000}} = 0.4 \pm 0.03 = [0.37, 0.43]$$

Confidence Intervals and Hypothesis Tests

- Confidence intervals and hypothesis tests are built from the same calculations.
 - They have the same assumptions and conditions.
- You can approximate a hypothesis test by examining a confidence interval.
 - Just ask whether the null hypothesis value is consistent with a confidence interval for the parameter at the corresponding confidence level.
- Because confidence intervals are two-sided, they correspond to two-sided tests.
 - A confidence interval with a confidence level of $C\%$ corresponds to a *two-sided hypothesis test* with an α -level of $100 - C\%$.
 - A confidence interval with a confidence level of $C\%$ corresponds to a *one-sided hypothesis test* with an α -level of $\frac{1}{2}(100 - C)\%$.

Example:

Suppose Amy wants to test if TAILS occurs more often than HEADS now. She took a random sample of size 1000 showed 400 HEADS. Carry out a hypothesis test at level of significance 0.05.

Let p = proportion of HEADS	Let p = proportion of TAILS

Calculate the 90% CI for p :

Let p = proportion of HEADS	Let p = proportion of TAILS

Example:

A company claims to have 40% of the market for some product.

You suspect this number, so you conduct a survey and find 38 out of 112 buyers (~33.9%) purchased this brand. Are these data consistent with the company's claim at $\alpha=0.05$ level?

Let p be the true market share of this product.

Example:

Suppose that the proportion of adults above 40 who are participating in fitness activities is 0.8 one year ago. An advertising campaign that promotes fitness activities is launched this year and you want to test whether the proportion is higher now. Assume you take a random sample of $n = 100$ and the number of people sampled who participate in those activities equals 85. Carry out a hypothesis test at the level of significance of 0.01 to test your claim.

Ch 21 Comparing Two Proportions

Estimating the Difference between Two Proportions

You may want to compare:

- the proportion of people who play computer games in the age groups of 20 to 30 and 30 to 40.
- the proportion of defective items manufactured in two production lines

The statistic for estimating the difference in two population proportions ($p_1 - p_2$) that comes to mind is the difference in the sample proportions ($\hat{p}_1 - \hat{p}_2$).

Notation:

	Population	Sample	
	proportion	size	Proportion
Population 1	p_1	n_1	\hat{p}_1
Population 2	p_2	n_2	\hat{p}_2

Properties of the Sampling Distribution of the Difference between two Independent Sample Proportions ($\hat{p}_1 - \hat{p}_2$)

Consider that you have two independent samples of sizes n_1 and n_2 from two populations with parameters p_1 and p_2 , respectively.

The sampling distribution of ($\hat{p}_1 - \hat{p}_2$) has these properties:

1. The mean of ($\hat{p}_1 - \hat{p}_2$) is $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$, so $\hat{p}_1 - \hat{p}_2$ is an unbiased estimate for $p_1 - p_2$

2. The variance $\sigma_{\hat{p}_1 - \hat{p}_2}^2 = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$, so the

standard deviation is: $\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

3. If the sample sizes n_1 and n_2 are both large, that is when

$$n_1 p_1 \geq 10 \text{ and } n_1(1-p_1) \geq 10$$

$$\text{and } n_2 p_2 \geq 10 \text{ and } n_2(1-p_2) \geq 10.$$

4. With these properties, then the sampling distribution of $\hat{p}_1 - \hat{p}_2$ approximately normal.

Two-Proportion z-Interval

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

For this we must have the following assumptions and conditions meet:

- Independence Assumptions:
 - Randomization Condition: The data in each group should be drawn independently and at random from a homogeneous population or generated by a randomized comparative experiment.
 - The 10% Condition: If the data are sampled without replacement, the sample should not exceed 10% of the population.
 - Independent Groups Assumption: The two groups we're comparing must be independent *of each other*.
- Sample Size Condition:
 - *Each* of the groups must be big enough
 - Success/Failure Condition: Both groups are big enough that at least 10 successes and at least 10 failures have been observed in each, that is:

$$n_1 p_1 \geq 10 \text{ and } n_1(1 - p_1) \geq 10$$

$$\text{and } n_2 p_2 \geq 10 \text{ and } n_2(1 - p_2) \geq 10.$$

Example:

Suppose we want to compare therapies. The criterion for the comparison is the probability to survive at least 5 years after therapy.

The study produced the following data:

	Therapy 1	Therapy 2
<i># of people sampled</i>	100	80
<i># of people survive at least 5 years after therapy</i>	90	70

Find the 95% confidence interval for the difference in proportions.

Example: APGAR score and gestational age at birth

APGAR scores provide measures of health for newborn infants.

Low scores indicate problems. A 5-minute APGAR less than 7 provides a relatively stable indicator of poor health.

The data below are a SRS of size 400 from a Royal Alexandra Hospital database.

	Full term	Preterm	Total
High APGAR	268	93	361
Low APGAR	14	25	39
Total	282	118	400

Let

p_1 = proportion with low APGAR for population of preterm births

p_2 = proportion with low APGAR for population of full term births

- Verify independence and sample size conditions:
 - Randomization condition: We have a simple random sample from each population. Or a randomized experiment with two treatments.
 - The two samples are independent.

- Success/Failure: Expected number of successes and failures in each sample is at least 10.
 - a) Construct a 95% CI for the proportion of low APGAR scores for preterm births (less than 37 weeks gestation).
 - b) Construct a 95% CI for the proportion of low APGAR scores for full term births.
 - c) Do the intervals for preterm births and full term births overlap? What do you think this means about the difference in measures of health between preterm and full term births?

d) Construct a 95% CI for the difference in proportions of low APGAR between the preterm and full term births.

Interpret.

e) Does the interval contain zero?

f) Are the results in parts c and e contradictory?

Two-Proportion z Test

1. Determine the type of test

a) lower tailed:

$$\circ \quad H_0 : p_1 - p_2 = p_0 \quad \text{versus} \quad H_a : p_1 - p_2 < p_0$$

b) upper tailed:

$$\circ \quad H_0 : p_1 - p_2 = p_0 \quad \text{versus} \quad H_a : p_1 - p_2 > p_0$$

c) two tailed:

$$\circ \quad H_0 : p_1 - p_2 = p_0 \quad \text{versus} \quad H_a : p_1 - p_2 \neq p_0$$

2. Assumption & Condition:

- a) The sample is randomly chosen and the sampled values are independent
- b) Assume the two samples are independent of each other
- c) Check the condition: Both sample sizes are large, that is:

$$n_1 p_1 \geq 10 \text{ and } n_1(1 - p_1) \geq 10$$

$$\text{and } n_2 p_2 \geq 10 \text{ and } n_2(1 - p_2) \geq 10.$$

3. Test statistic:

$$z_0 = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

If $p_1 - p_2 = 0$, then we are hypothesizing that there is no difference between the two proportions, which implies the standard deviations for each proportion are the same. Since this is the case, we combine (or pool) the counts to get an overall

$$\text{estimate } \hat{p}_{\text{pooled}} = \frac{\text{total number of successes}}{\text{total observations}} = \frac{(n_1 \hat{p}_1 + n_2 \hat{p}_2)}{n_1 + n_2},$$

which is the combined estimate of the common population

proportion p_{pooled} , and we put this pooled value into the formula substituting it for both sample proportions:

$$\begin{aligned} z_0 &= \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{\hat{p}_{pooled}(1 - \hat{p}_{pooled})}{n_1} + \frac{\hat{p}_{pooled}(1 - \hat{p}_{pooled})}{n_2}}} \\ &= \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}_{pooled}(1 - \hat{p}_{pooled})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \end{aligned}$$

4. P-value:

Test Type	p-value
Upper Tail	$P(z > z_o)$
Lower Tail	$P(z < z_o)$
Two Tails	$2 \times P(z > z_o)$

5. Decision:

- $p\text{-value} \leq \alpha \rightarrow H_0$ is rejected.
- $p\text{-value} > \alpha \rightarrow H_0$ is not rejected.

Example:

Find if the proportions of red M&M's in the plain and peanut variety differ at a significance level of 0.05.

The sample

	Plain(1)	Peanut(2)
Sample Size	56	64
Number of Red M&Ms	12	16

Example:

Suppose you want to compare the infestation by a specific pest of two forests. Let p_1 be the proportion of trees in forest 1 that are affected and p_2 be the proportion of trees in forest 2 that are affected. Find if the proportions of trees that are affected in forest 1 to a lesser degree than in forest 2 at a significance level of 0.05.

We obtain the sample:

	Forest 1	Forest 2
Sample Size	100	100
Number of trees affected	10	15

Ch 18 Inference About Mean

18.1 The CLT Revisited

Now that we know how to create confidence intervals and test hypotheses about proportions, it'd be nice to be able to do the same for means.

Just as we did before, we will base both our confidence interval and our hypothesis test on the sampling distribution model.

Recall: If we use the statistic \bar{y} for estimating the population mean μ , we can use the following information from the CLT in order to obtain a confidence interval for μ .

- ❖ $\mu_{\bar{y}} = \mu$,
- ❖ $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$ standard deviation of \bar{y} ,
- ❖ The standard error of \bar{y} is $SE(\bar{y}) = \frac{s}{\sqrt{n}}$ AND
- ❖ If the population distribution is originally normal, then the sampling distribution is also normal OR
- ❖ If the population distribution is non normal, but it has $n \geq 30$, then we can assume that the sampling distribution of \bar{y} is approximately normal.

18.2 Gosset's t

Until now, all statistical tools that were introduced were based on the assumption that population standard deviation σ is known. In practice, this assumption is very artificial and is never fulfilled in any real live situation.

All procedures introduced until now are based on the *normal* distribution, which requires the population standard deviation σ . In most situations, σ is unknown and has to be replaced by the sample standard deviation s , it causes variability in the result. In order to calculate a confidence interval, we need to fix the problem of variability by introducing another distribution called the ***Student's t-distribution***.

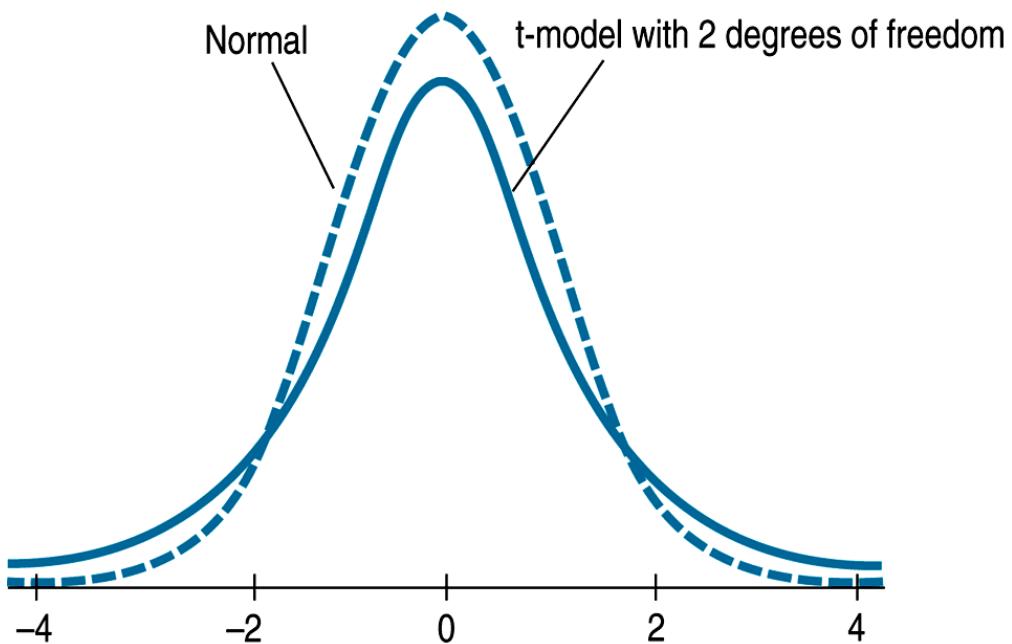
The t -distribution only depends on one parameter, which is called the **degrees of freedom** (df).

Properties of the t-distribution:

- its density curves look quite similar to the standard normal curve. They are symmetric about 0, single-peaked, and bell-shaped.
- The spread of the t -distributions is a bit larger than that of the standard normal curve. (As we are now using an

estimate for the population standard deviation, we must accept slightly more error in our estimation.)

- As degrees of freedom (d.f.) gets bigger, the t -density curve gets closer to the standard normal density curve. (NOTE: Table t) In other words, as degrees of freedom increases, the spread of the corresponding t density curve decreases.
- In fact, the t -model with infinite df is exactly normal.



Remark: The structure of the table is different than the table for the standard normal distribution.

- It is giving you the upper tail probabilities!
- the probabilities are the label of the columns instead of being inside the table.

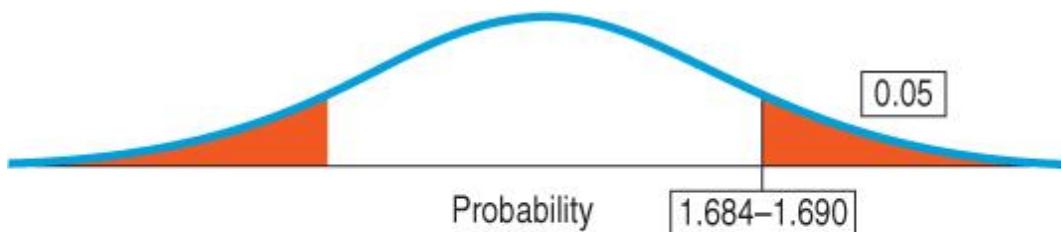
- This table will not provide the critical values for all the degrees of freedom. We need approximate the critical value in such cases or use software.

Example: Find t^* (the critical value).

- a) The t-distribution with 5 df has probability 0.05 to the right of t^* .
- b) The t-distribution with 5 df and confidence level of 90%.
- c) The one sample t statistic from an SRS with 20 observations has a probability of 0.9 to the left of t^* .
- d) The t-distribution with 39 df has probability 0.05 to the right of t^* .

NOTE: The correct value lies between 1.684 and 1.690.

Either be conservative and go with the bigger value, 1.690, or use software



	Two-tail	0.20	0.10	0.05	0.02	0.01
	One-tail	0.10	0.05	0.025	0.01	0.005
28		1.313	1.701	2.048	2.467	2.763
29		1.311	1.699	2.045	2.462	2.756
30		1.310	1.697	2.042	2.457	2.750
32		1.309	1.694	2.037	2.449	2.738
35		1.306	1.690	2.030	2.438	2.725
40		1.303	1.684	2.021	2.423	2.704
45		1.301	1.679	2.014	2.412	2.690
50		1.299	1.676	2.009	2.403	2.678
60		1.296	1.671	2.000	2.390	2.660

18.3 A Confidence Interval for a Population Mean μ (when σ is unknown)

Assumptions for using the t-statistics:

Independence Assumption:

- Independence Assumption. The data values should be independent.
- Randomization Condition: The data arise from a random sample or suitably randomized experiment. Randomly sampled data (particularly from an SRS) are ideal.

- 10% Condition: When a sample is drawn without replacement, the sample should be no more than 10% of the population.

Normal Population Assumption:

- We can never be certain that the data are from a population that follows a Normal model, but we can check the Nearly Normal Condition: The data come from a distribution that is unimodal and symmetric.
 - o Check by making a histogram or Normal probability plot.
- You can also ensure normality by checking sample size is large enough

A confidence interval for the population mean μ (when σ is unknown) is given by

$$\bar{y} \pm t^* \frac{s}{\sqrt{n}}$$

where t^* is the critical value for the t distribution with $df = n - 1$ confidence level C . In other words, t^* is the upper $\frac{1-C}{2}$ critical value for the $t(n-1)$ distribution.

NOTE:

- When Gosset corrected the model for the extra uncertainty, the margin of error got bigger.
- Your confidence intervals will be just a bit wider and your P-values just a bit larger than they were with the Normal model.
- By using the t -model, you've compensated for the extra variability in precisely the right way.

Example: (Using the battery lifetime example from Ch3)

We have a random sample of $n = 4$ observations on y = battery lifetime (hrs): 5.9, 7.3, 6.6, 5.7

NOTE: $\bar{y} = 6.375$, $s = 0.7274$ (calculated in Ch3)

Find the 95% confidence interval for the mean battery lifetime.

Example:

A scientist interested in monitoring chemical contaminants in food, and thereby the accumulation of contaminants in human diets, selected a random sample of $n = 50$ male adults. It was found that the average daily intake of dairy products was $\bar{y} = 756$ grams with a standard deviation of $s = 35$ grams.

Find a 95% confidence interval for the mean daily intake of dairy products for men.

Example: IQ test scores

The SRS IQ test scores of 31 girls in Region A as follows:

113 102 105 ... 95

This has a sample mean $\bar{y} = 105.84$ and a sample standard deviation of $s = 15$. The shape of the population distribution is unimodel and relatively symmetric.

- a) Give a 99% confidence interval for the true mean IQ μ of all girls in the district.

b) Give a 90% confidence interval for the true mean IQ μ of all girls in the district.

- c) If the sample mean of IQ test scores of 20 girls in Region A is 105.84 with $s = 15$, give a 90% confidence interval for the true mean IQ μ of all girls in the district.

Remark:

Margin of error $m = t^* \frac{s}{\sqrt{n}}$ gets smaller when

- t^* gets smaller, which is the same as smaller $(1 - \alpha)$. To obtain a smaller margin of error, you must accept lower confidence.
- n gets larger. Increasing the sample size gives more accuracy.
- σ gets smaller. The less inherent variation in the population you are studying, the more accurate your estimate will be.

NOTE: we can control t^* and n , but we cannot control σ .

Example: (Please try it on your own)

Bank	Mean Number of Phone Calls/hour
A	15.6
B	11.9
C	11.7

Suppose that each sample mean was based on an SRS of $n = 50$ working hours and that $s = 5$ is known.

- a) Compute a 95% CI for μ_A , the true mean number of phone calls per hour to Bank A.

With $C = 95\%$ and $df = 49$ (round down to 45), $t^* = 2.014$.

$$\bar{y} \pm t^* \frac{s}{\sqrt{n}} = 15.6 \pm 2.014 \frac{5}{\sqrt{50}} = 15.6 \pm 1.424$$

We are 95% confident that μ_A is between 14.176 and 17.024 phone calls per hour.

- b) Compute a 95% CI for μ_B and μ_C .

95% CI for μ_B :

$$\bar{y} \pm t^* \frac{s}{\sqrt{n}} = 11.9 \pm 2.014 \frac{5}{\sqrt{50}} = 11.9 \pm 1.424 = (10.476, 13.324)$$

95% CI for μ_C :

$$\bar{y} \pm t^* \frac{s}{\sqrt{n}} = 11.7 \pm 2.014 \frac{5}{\sqrt{50}} = 11.7 \pm 1.424 = (10.276, 13.124)$$

- c) A survey claims that Bank A receives more phone calls than the other Banks. Based on the confidence intervals from parts (a) and (b), do you agree?

The \bar{y} value for Bank A is so large that its confidence interval lies entirely to the right of all other CIs. Even taking random variation into account, the number of phone calls received is clearly larger than other Banks.

Example: A researcher found that a 98% confidence interval for the mean hours per week spent studying by college students was (13, 17). Which is true?

- a) There is a 98% chance that the mean hours per week spent studying by college students is between 13 and 17 hours
- b) We are 98% confident that the mean hours per week spent studying by college students is between 13 and 17 hours**
- c) Students average between 13 and 17 hours per week studying on 98% of the weeks
- d) 98% of all students spend between 13 and 17 hours studying per week.

Previously, population parameters were described, now we will be checking if claims about the population parameters are true, or plausible to a given degree.

Example:

A company is advertising that the mean lifetime of their light bulbs is 1000 hours. A person suspects the mean lifetime of the light bulbs is less than 1000 hours (company is lying in their advertisement), so he picks a sample of 100 light bulbs and finds the average lifetime of these 100 light bulbs is $\bar{y} = 998$.

Based on this result, can we state that:

- i) the mean lifetime of this company's light bulb, on average, is less than 1000 hours (so this company is lying in their advertisement)? OR
- ii) the difference between 1000 hours (the average lifetime for the population) and 998 hours (the average lifetime for the sample) may have occurred because of sampling variability?

Recall: How to write the hypothesis:

Example: You are considering moving to Richmond Hill, and are concerned about the average one-way commute time to downtown Toronto. Does the average one-way commute time exceed 25 minutes? You take a random sample of 50 Richmond Hill residents and find an average commute time of 29 minutes with a standard deviation of 7 minutes. Which set of hypotheses should you test?

- | | | |
|--------------------|----|--------------------|
| A) $H_0: \mu = 25$ | vs | $H_A: \mu > 25$ |
| B) $H_0: \mu = 25$ | vs | $H_A: \mu < 25$ |
| C) $H_0: \mu = 29$ | vs | $H_A: \mu > 29$ |
| D) $H_0: \mu = 25$ | vs | $H_A: \mu \neq 25$ |

Example: You want to see if the number of minutes cell phone users use each month has changed from its mean of 120 minutes 2 years ago. You take a random sample of 100 cell phone users and find an average of 135 minutes used. Which set of hypotheses should you test?

- | | | |
|---------------------------------------|-----------|---------------------------------------|
| A) $H_0: \mu = 120$ | vs | $H_A: \mu > 120$ |
| B) $H_0: \mu = 120$ | vs | $H_A: \mu \neq 120$ |
| C) $H_0: \mu = 120$ | vs | $H_A: \mu < 120$ |
| D) $H_0: \mu = 135$ | vs | $H_A: \mu \neq 135$ |

18.4 Hypothesis Test for the Mean

Example (con't):

A company is advertising that the average lifetime of their light bulbs is 1000 hours. A random sample of size 100 showed the average lifetime of their light bulbs is 998 hours with $s = 5$.

You want to test $H_0: \mu = 1000$ versus $H_a: \mu < 1000$.

But can the difference between μ and \bar{y} be explained by the sampling variability?

To find this out, we calculate the test statistic that will relate the sample value \bar{y} with the claimed value from the **null hypothesis** μ_0 .

$$t_0 = \frac{\bar{y} - \mu}{s_{\bar{y}}} = \frac{\bar{y} - \mu_0}{s / \sqrt{n}}$$

If H_0 is true, this test statistic is approximately standard normal distributed (CLT: sample size of 100 is large enough), so that the value from a random sample can be judged by the standard normal distribution.

For this sample,

$$t_0 = \frac{\bar{y} - \mu}{s_{\bar{y}}} = \frac{\bar{y} - \mu_0}{s / \sqrt{n}} = \frac{998 - 1000}{5 / \sqrt{100}} = -4$$

That is, if the null hypothesis $H_0: \mu = 1000$ is true, $\bar{y} = 998$ is 4 standard deviations less than what we would expect it to be. We know that this t test statistic is very low and is unlikely to occur. Let's calculate the probability to observe such a small or even smaller value, if H_0 is in fact true (i.e. we calculate the p-value):

$$\text{p-value} = P(t < -4) \approx 0$$

There is virtually **no** chance of observing this value of the test statistic t this extreme as a result of chance variation alone when H_0 is true. Hence, there is almost no chance of seeing a sample mean \bar{y} value as extreme as that observed. The evidence by the sample is compelling for H_0 not to be true. We will reject H_0 in favor of H_a .

One Sample t-Test for a Population Mean μ

1. The assumptions and conditions for the one-sample t -test for the mean are the same as for the one-sample t -interval.
2. Hypothesis:

Test Type	
Upper Tail	$H_0: \mu = \mu_o$ $H_a: \mu > \mu_o$
Lower Tail	$H_0: \mu = \mu_o$ $H_a: \mu < \mu_o$
Two Tails	$H_0: \mu = \mu_o$ $H_a: \mu \neq \mu_o$

3. Test statistic:

$$t_0 = \frac{\bar{y} - \mu_0}{\frac{s}{\sqrt{n}}}$$

with $n - 1$ degrees of freedom.

4. P-value:

Test Type	p-value
Upper Tail $H_0: \mu = \mu_o$ $H_a: \mu > \mu_o$	$P(t > t_o)$

Lower Tail	$H_0: \mu = \mu_o$	$P(t < t_o)$
	$H_a: \mu < \mu_o$	
Two Tails	$H_0: \mu = \mu_o$	$2 \times P(t > t_o)$
	$H_a: \mu \neq \mu_o$	

5. Decision: Reject H_0 , if and only if $p\text{-value} \leq \alpha$.

Example:

We have a random sample of $n = 4$ observations on $y = \text{battery lifetime (hrs)}$: 5.9, 7.3, 6.6, 5.7

NOTE: $\bar{y} = 6.375$, $s = 0.7274$

A scientist is interested in knowing if the average battery lifetime can last more than 6.3 hours at a level of significance of 0.025. Carry out a hypothesis test.

Remark:

- Confidence intervals and hypothesis tests are built from the same calculations.
- In fact, they are complementary ways of looking at the same question.
- The confidence interval contains all the null hypothesis values we can't reject with these data.
- More precisely, a level C confidence interval contains *all* of the possible null hypothesis values that would not be rejected by a two-sided hypothesis test at alpha level $(1 - C)$.
 - So a 95% confidence interval matches a 0.05 level test for these data.
 - Confidence intervals are naturally two-sided, so they match exactly with two-sided hypothesis tests.
 - When the hypothesis is one sided, the corresponding alpha level is $(1 - C)/2$.
 - So a 95% confidence interval matches a 0.025 level test for these data.

Example:

Poisoning by the pesticide DDT affects the nervous system and ought to slow the “absolutely refractory period,” the time required for a nerve to recover after a stimulus. This period is known to be 1.3ms in normal rats and follows a normal distribution. Measurements were taken on 4 rats exposed to DDT, and we get $\bar{y} = 1.75$ and $s = 0.13$. Do we have evidence that DDT poisoning slows nerve recovery at $\alpha = 0.05$ level? The parameter of interest is the true mean absolutely refractory period μ in poisoned rats.

Can I conclude that every rat exposure to DDT poisoning has a nerve recovery period greater than 1.3ms?

Example:

A health center reports that the mean systolic blood pressure for males over 35 years of age is 128. The medical director of a large company looks at the records of 72 executives and finds the mean blood pressure in this group is $\bar{y} = 126.07$ with $s = 15$. Is this evidence that the company's executives have a different mean blood pressure than the general population at $\alpha = 0.05$ level?

18.5 Determining the Sample Size

One of the important decisions, before drawing a sample, is how many experimental units from the population should be sampled. That is: what is the appropriate sample size?

The answer depends on the specific object of investigation and the precision or accuracy one wants to insure. A measure for the accuracy in estimation is the margin of error.

In general, the researcher chooses the largest value ME that is acceptable for the margin of error. Then the researcher determines what confidence level C he wants to attain in his claims in the study. From this, the necessary sample size can be determined.

- To find the sample size needed for a particular confidence level with a particular margin of error (ME), solve this

$$\text{equation for } n: \quad ME = t^* \frac{s}{\sqrt{n}}$$

- We don't know most of the values. To overcome this:
 - We can use s from a small pilot study.
 - We can use z^* in place of the necessary t value.

$$\blacksquare \text{ Thus, } n \geq \left(\frac{z^* s}{ME} \right)^2$$

Example:

Suppose you want to estimate the average daily yield μ of a chemical process and you want to insure with 95% confidence that the estimate is not more than 4 tons of the true mean yield μ . Assume a previous sample would have shown a sample standard deviation of $s = 21$ tons. Find the minimum sample size needed.

Example:

The financial aid office wishes to estimate the mean cost of textbooks per quarter for students at a particular college. For the estimate to be useful, it should be within \$20 of the true population mean. How large a sample should be used to be 95% confident of achieving this level of accuracy if the financial aid uses a standard deviation of \$100?

19 Comparing Means

19.1 Independent populations

The two populations under investigation don't have any impact on each other are independent.

Example: comparing the height of males and females in a certain company

Example: comparing the grades of university students in Canada and US.

Since we are investigating two populations, we will introduce the following notations:

Notation:

	Population		sample		
	mean	Std dev	size	mean	Std dev
Population 1	μ_1	σ_1	n_1	\bar{y}_1	s_1
Population 2	μ_2	σ_2	n_2	\bar{y}_2	s_2

In the presence of two populations, it is usually the goal to compare them. To decide if the sample means are significantly different, we cannot just compare the difference of the mean.

Instead, we will see that the observed difference between the two samples depends on how big the difference is compared to the inherent variability in the populations.

In order to do inferential statistics using this difference we have to investigate the distribution of this statistic.

Sampling Distribution of $\bar{y}_1 - \bar{y}_2$ from two independent samples

Assuming unequal variances	Assuming equal variances
$\mu_{\bar{y}_1 - \bar{y}_2} = \mu_1 - \mu_2$	$\mu_{\bar{y}_1 - \bar{y}_2} = \mu_1 - \mu_2$
$\sigma_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$\sigma_{\bar{y}_1 - \bar{y}_2} = \sigma_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
$s_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$s_{\bar{y}_1 - \bar{y}_2} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, where the pooled estimate of standard deviation is $s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$
If both populations are normal distributed or n_1 and n_2 are both large, then the sampling distribution of $\bar{y}_1 - \bar{y}_2$ is (approximately) normal	

NOTE: Since σ_1 and σ_2 are unknown, we need to use the t -distribution as the sampling distribution.

How to tell whether there is equal or unequal variance?

For this course, we will use the following condition to check whether we have equal variability:

- the informal standard deviation ratio is less than two or look for similar IQR using boxplots

This condition must hold for equal variability, otherwise, you can assume unequal variability.

	2-sample t test assuming equal variance (Pooled t-test)	2-sample t test assuming not equal variance (Nonpooled t-test)
Assumption	1) Independent Groups 2) Independent Observations - Randomization 3) both populations have distribution approximately normal distributed, or n_1 and n_2 are large enough 4) Equal variability	4) Unequal variability

Hypothesis	Upper Tail $H_0: \mu_1 - \mu_2 = d_o$ $H_a: \mu_1 - \mu_2 > d_o$	
	Lower Tail $H_0: \mu_1 - \mu_2 = d_o$ $H_a: \mu_1 - \mu_2 < d_o$	
	Two Tails $H_0: \mu_1 - \mu_2 = d_o$ $H_a: \mu_1 - \mu_2 \neq d_o$	
Test statistic	$t_0 = \frac{\bar{y}_1 - \bar{y}_2 - d_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$t_0 = \frac{\bar{y}_1 - \bar{y}_2 - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
Distribution	$\sim t$ -distribution with $df = n_1 + n_2 - 2$ and $s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$	$\sim t$ -distribution with $df = \min(n_1 - 1, n_2 - 1)$ or $df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{s_1^2}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right)\left(\frac{s_2^2}{n_2}\right)^2}$
P-value	Upper tail $P(t > t_o)$	
	Lower tail $P(t < t_o)$	
	Two tails $2 \times P(t > t_o)$	
Decision	$p\text{-value} \leq \alpha \rightarrow H_0$ is rejected.	
	$p\text{-value} > \alpha \rightarrow H_0$ is not rejected.	

The (1-α)100% level Confidence Interval	$(\bar{y}_1 - \bar{y}_2) \pm t^* s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ <p>where t^* represents the $100\left(1 - \frac{\alpha}{2}\right)th$ percentile of the t-distribution with $df = n_1 + n_2 - 2$</p>	$\bar{y}_1 - \bar{y}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ <p>with t^* represents the $100\left(1 - \frac{\alpha}{2}\right)th$ percentile of the t-distribution with $df = \min(n_1 - 1, n_2 - 1)$ or</p> $df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{s_1^2}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right)\left(\frac{s_2^2}{n_2}\right)^2}$
--	---	--

Example: A company wants to show that a vitamin supplement decreases the recovery time from a common cold. They selected randomly 70 adults with a cold. 35 of those were randomly selected to receive the vitamin supplement. The data on the recovery time for both samples is shown below.

Population	1	2
	no vitamin	Vitamin
sample size	35	35
sample mean	6.9	5.8
sample std deviation	2	1.2

Now test the claim of the company that vitamin supplement decreases the average recovery time from a common cold at level of significant 0.025.

Calculate a 95% Confidence Interval for the difference in average recovery time $\mu_1 - \mu_2$.

Refer to the previous “Vitamin Supplement” Example with the following revised data:

Population	1 no vitamin	2 Vitamin
sample size	35	35
sample mean	6.9	5.8
sample std deviation	2.9	1.2

Now test the claim of the company that vitamin supplement decreases the average recovery time from a common cold at level of significant 0.025.

Continue Example from above:

Calculate a 95% Confidence Interval for the difference in average recovery time $\mu_1 - \mu_2$.

Example:

In an attempt to determine if two competing brands of cakes contain, on the average, the same amount of baking powder, twelve different cakes from each of the two competing brands were randomly selected and tested for the amount of baking powder each contains. The results (in milligrams) follow the normal distribution. Use a significance level of 0.01.

<u>Brand A</u>	<u>Brand B</u>
517, 495, 503, 491	493, 508, 531, 521
503, 493, 505, 495	569, 572, 500, 515
498, 481, 499, 494	536, 490, 519, 515

State and perform an appropriate hypothesis test.

μ_1 = the mean amount of baking powder in cakes of brand A

μ_2 = the mean amount of baking powder in cakes of brand B

Example:

A philosophy professor wants to find out whether the mean age of men in his large lecture class is equal to the mean age of the women in his class. After collecting data from his students, the professor tested the hypothesis $H_0: \mu_M - \mu_W = 0$ against the alternative $H_a: \mu_M - \mu_W \neq 0$. The p-value for the test was 0.003. Which is true?

- A. There is a 0.3% chance that the mean age for the men is equal to the mean age for the women.
- B. There is a 0.3% chance that the mean age for the men is different from the mean age of women.
- C. **It is very unlikely that the professor would see results like these if the mean age of men was equal to the mean age of women.**
- D. There is a 0.3% chance that another sample will give these same results.

20 Paired Samples and Blocks

- Data are **paired** when the observations are collected in pairs or the observations in one group are naturally related to observations in the other group.
- Paired data arise in a number of ways. Perhaps the most common is to compare subjects with themselves before and after a treatment.
 - When pairs arise from an experiment, the pairing is a type of *blocking*.
 - When they arise from an observational study, it is a form of *matching*.
- If you know the data are paired, you can (and must!) take advantage of it.
 - To decide if the data are paired, consider how they were collected and what they mean.
 - There is no test to determine whether the data are paired.
- Once we know the data are paired, we can examine the *pairwise* differences.

- Because it is the *differences* we care about, we treat them as if they were the data and ignore the original two sets of data.

Example 1:

Compare the resting pulse and pulse after exercise.

To control for all other influences, you take both measurements on every individual in one sample.

We are interested in the difference in the population means $\mu_d = \mu_1 - \mu_2$.

For statistical inference, the difference of each of the paired observations in the sample:

Sample 1 value – Sample 2 value.

This creates one sample of size n of measurement differences.

Sample 1 Observation Value	Sample 2 Observation Value	Difference for pairs
y_{1_1}	y_{2_1}	$y_{d_1} = y_{1_1} - y_{2_1}$
y_{1_2}	y_{2_2}	$y_{d_2} = y_{1_2} - y_{2_2}$
\vdots	\vdots	\vdots
y_{1_n}	y_{2_n}	$y_{d_n} = y_{1_n} - y_{2_n}$

Now that we have only one set of data to consider, we can return to the simple one-sample t -test.

- Mechanically, a paired t -test is just a one-sample t -test for the mean of the pairwise differences.
 - The sample size is the number of pairs.

Notation:

- \bar{d} is the mean of the pairwise differences
- s_d is the standard deviation of the pairwise differences.
- n is the number of pairs.

20.1 Paired t-Test for Comparing Two Population Means

Assumption:

- Paired Data Assumption:
 - Paired data Assumption: The data must be paired.
- Independence Assumption:
 - Independence Assumption: The differences must be independent of each other.

- Randomization Condition: Randomness can arise in many ways. What we want to know usually focuses our attention on where the randomness should be.
- 10% Condition: When a sample is obviously small, we may not explicitly check this condition.
- Normal Population Assumption: We need to assume that the population of *differences* follows a Normal model.
- Nearly Normal Condition: Check this with a histogram or Normal probability plot of the differences.

Hypothesis:

Test Type
Upper Tail $H_0: \mu_d = d_o$ vs. $H_a: \mu_d > d_o$
Lower Tail $H_0: \mu_d = d_o$ vs. $H_a: \mu_d < d_o$
Two Tails $H_0: \mu_d = d_o$ vs. $H_a: \mu_d \neq d_o$

Test statistic:

$$t_0 = \frac{\bar{d} - d_0}{\frac{s_d}{\sqrt{n}}} \quad \text{with df} = n - 1.$$

Critical Value and P-value:

Test Type	p-value
Upper Tail	$P(t > t_o)$
Lower Tail	$P(t < t_o)$
Two Tails	$2 \times P(t > t_o)$

Decision:

- $p\text{-value} \leq \alpha \rightarrow H_0$ is rejected. You report that the results are statistically significant at level α .
- $p\text{-value} > \alpha \rightarrow H_0$ is not rejected. You report that the results are *not* significant at level α .

20.2 Paired t-Confidence Interval for μ_d

Assumption: Same as the paired t -test.

The level 100C% Confidence Interval for μ_d :

$$\bar{d} \pm t^* \frac{s_d}{\sqrt{n}}$$

where t^* is the critical value of the t-distribution with $n - 1$ degrees of freedom.

Example 2:

The effect of exercise on the amount of lactic acid in the blood was examined.

Blood lactate levels were measured in eight males before and after playing three games of racquetball.

Player	Before	After	Difference
1	13	18	-5
2	20	37	-17
3	17	40	-23
4	13	35	-22
5	13	30	-17
6	16	20	-4
7	15	33	-18
8	16	19	-3

Let's test if the lactate level before exercise is lower than the lactate level after exercise at a level of significance of 0.05.

NOTE1: $d = y_{\text{before}} - y_{\text{after}}$

NOTE2: $\bar{d} = -13.63$, $s_d = 8.28$

Let's give an estimate (90% Confidence interval) for μ_d , mean difference of lactate level.

Example 3: Sidestream smoke

A researcher believes that sidestream (second hand) smoke may be more dangerous than mainstream smoke (inhaled directly from the cigarette). Here's some data on tar yield.

Brand	Side	Main	Difference
A	15.8	18.5	-2.7
B	16.9	17.0	-0.1
C	21.6	17.2	4.4
D	18.8	19.4	-0.6
E	29.3	15.6	13.7
F	20.7	16.4	4.3
G	18.9	13.3	5.6

Carry out a hypothesis test and see whether the data support the researcher's belief at level of significance of 0.05.

NOTE: What if at the level of significant of 0.1?

Example:

Random samples of 50 men and 50 women are asked to imagine buying a birthday present for their best friend. We want to estimate the difference in how much they are willing to spend.

We would use a

- a. Two-sample t hypothesis test
- b. Two-sample t confidence interval**
- c. Paired t hypothesis test
- d. Paired t confidence interval

Example:

Are parents equally strict with boys and girls? In a random sample of families, researchers asked a brother and sister from each family to rate how strict their parents were. We would use a

- a. Two-sample t hypothesis test
- b. Two-sample t confidence interval
- c. Paired t hypothesis test**
- d. Paired t confidence interval

22 Comparing Counts

This chapter will cover 2 types of tests:

1. Tests of hypotheses for experiments with more than 2 categories, called goodness-of-fit tests.
2. Tests of hypotheses about contingency tables, called independence and homogeneity tests.

Goodness-of-Fit Test or Univariate χ^2 Test

- a test of whether the frequency distribution of a categorical variable with more than 2 categories from a sample matches the probability distribution predicted by a model.

Example:

A company filling grass seed bags wants to evaluate their filling machine. The following distribution is advertised on the bags:

Kinds of seeds	Proportion (expected frequency)
K1	0.5
K2	0.25
K3	0.15
K4	0.05
K5	0.05

The company wants to check if the seed distribution in the bags fits the advertised distribution.

They take a random sample of size 1000 and find the following summarized data:

Kinds of seeds	Count
K1	480
K2	233
K3	160
K4	63
K5	64

The χ^2 Goodness-of-Fit Test

Given a distribution p_{01}, \dots, p_{0k}

Hypotheses:

$$H_0 : p_1 = p_{01}, \dots, p_k = p_{0k} \text{ vs } H_a : H_0 \text{ is not true.}$$

Assumptions and Conditions:

- Counted Data Condition: Check that the data are *counts* for the categories of a categorical variable.
- Independence Assumption: The counts in the cells should be independent of each other.

- Randomization Condition: The individuals who have been counted and whose counts are available for analysis should be a random sample from some population.
- Sample Size Assumption: We must have enough data for the methods to work.
 - Expected Cell Frequency Condition: We should expect to see at least 5 individuals in each cell.

Test Statistics:

- The test statistic, called the **chi-square** (or chi-squared) **statistic**, is found by adding up the sum of the squares of the deviations between the observed and expected counts divided by the expected counts:

$$\chi^2_0 = \sum_{\text{all categories}} \frac{(Observed - Expected)^2}{Expected}$$

- Where the expected value is the product of the total number of observations times this proportion.

NOTE:

- The chi-square value follows a χ^2 distribution (like t distribution, it is identify by a value called degrees of freedom)

- The number of degrees of freedom for a goodness-of-fit test is $c - 1$, where c is the number of categories.
- The chi-square statistic is used only for testing hypotheses, not for constructing confidence intervals.
- If the observed counts don't match the expected, the statistics will be large.
 - this statistic χ^2 measures how far apart the observed and expected counts are
 - When $Observed - Expected$ is small, the claim is right.
(because Observed is closed to Expected value)
 - When it is large, bad evidence.

P-value = $P(\chi^2 > \chi_0^2)$ is the upper-tail area for a χ^2 distribution with $c - 1$ df.

- The mechanics may work like a one-sided test, but the interpretation of a chi-square test is in some ways many-sided.
- There are many ways the null hypothesis could be wrong.
- There is no direction to the rejection of the null model – all we know is that it doesn't fit.

Continue Example:

Carry out a hypothesis test and see whether the advertised label is giving a true description of the contents of the seed bag.

$$H_0: p_{01} = 0.5; p_{02} = 0.25; p_{03} = 0.15; p_{04} = 0.05; p_{05} = 0.05$$

H_a : at least 1 kind of seeds is not in the claiming proportions

Kinds of seeds	Advertised Proportion	Observed Count	Expected Count (np_o)
K1	0.5	480	
K2	0.25	233	
K3	0.15	160	
K4	0.05	63	
K5	0.05	64	

Example:

On a bag of candies with 6 different colors, it states that:

Color	Proportion
Brown	0.3
Red	0.2
Yellow	0.2
Orange	0.1
Green	0.1
Blue	0.1

You randomly take one bag and started counting and find the following summarized data:

Color	Number
Brown (br)	50
Red (r)	32
Yellow (y)	20
Orange (o)	18
Green (g)	22
Blue (bl)	25

Carry out an appropriate hypothesis test to test whether the color distribution of the candies comes in the proportions claimed by the company.

$H_0: p_{br} = 0.3; p_r = 0.2; p_y = 0.2; p_o = 0.1; p_g = 0.1; p_{bl} = 0.1$ (the color distribution comes in the proportions claimed by the company)

$H_a:$ at least 1 color of candies is not in the claiming proportions

χ^2 Test for Homogeneity and Independence in a 2-Way-Table

A Test of Homogeneity

- A test comparing the distribution of counts for two or more groups on the same categorical variable is called a *chi-square test of homogeneity*.
- A test of homogeneity is actually the generalization of the two-proportion *z*-test.
- The statistic that we calculate for this test is *identical* to the chi-square statistic for goodness-of-fit.
- In this test, however, we ask whether choices have changed (i.e., there is no model).
- The expected counts are found directly from the data and we have different degrees of freedom.

Assumptions and Conditions

- The assumptions and conditions are the same as for the chi-square goodness-of-fit test:
 - Counted Data Condition: The data must be counts.
 - Randomization Condition.
 - Expected Cell Frequency Condition: The expected count in each cell must be at least 5.

Test statistics:

- To find the expected counts, we multiply the row total by the column total and divide by the grand total.
- We calculated the chi-square statistic as we did in the goodness-of-fit test:

$$\chi_0^2 = \sum_{\text{all categories}} \frac{(Observed - Expected)^2}{Expected}$$

- In this situation we have $(R - 1)(C - 1)$ degrees of freedom, where R is the number of rows and C is the number of columns.
- We'll need the degrees of freedom to find a P-value for the chi-square statistic.

p-value = $P(\chi^2 > \chi_0^2)$ is the upper-tail area for a χ^2 distribution with $(R - 1)(C - 1)$ degrees of freedom

Example:

Suppose we define 3 income strata: high income group (with income $> \$100,000$), medium income group (with income of $\$50,000$ to $\$100,000$), and low income group (with an income of less than $\$50,000$). Furthermore, assume that we take one sample of 250 households from California and another sample of 150 households from Wisconsin, and the collected the info on the incomes of these households are shown in the following table:

	California	Wisconsin	Total
High Income	70	34	104
Medium Income	80	40	120
Low Income	100	76	176
Total	250	150	400

Use appropriate test to test the null hypothesis that the distribution of income is the same for California and Wisconsin at level of significance 0.025.

H_0 : *the proportions of households that belong to different income groups are the same in both states*

H_a : *these proportions of households that belong to different income groups are not the same in both states*

Heart Disease Example:

A study has been conducted and resulted in the following data:

		Smoker		Total
		yes	No	
Heart disease	Yes	23	15	38
	no	69	259	328
Total		92	274	366

- a) Conduct a χ^2 test to see whether the proportion of having heart disease is the same for smoker and nonsmoker at level of significance 0.025.

- b) Carry out a two proportion z-test to see whether the proportion of having heart disease is the same for smoker and nonsmoker at level of significance 0.025.
- c) Compare the two tests in part (a) and (b).

The χ^2 Test for Independence

- Contingency tables categorize counts on two (or more) variables so that we can see whether the distribution of counts on one variable is contingent on the other.
- Tests of independence examine counts from a single group for evidence of an association between two categorical variables.
- A chi-square test of independence uses the same calculation as a test of homogeneity; the only difference is *what you think*.

Assumptions and Conditions:

- We still need counts and enough data so that the expected values are at least 5 in each cell.
- If we're interested in the independence of variables, we usually want to generalize from the data to some population.
 - In that case, we'll need to check that the data are a representative random sample from that population.

Example:

A survey was conducted to evaluate the effectiveness of a new flu vaccine that had been administered in a small community. It consists of a two-shot sequence of two weeks.

A survey of 1000 residents the following spring provided the following information:

	No vaccine	One shot	Two shot	Total
Flu	24	9	13	46
No Flu	289	100	565	954
Total	313	109	578	1000

Is there any evidence that the number of flu shots and the incidence of the flu are independent?

Example (Please try it on your own):

Some month ago there was a report in the news that an AIDS vaccine tested in Thailand didn't show any effect. The data quoted in the news is presented in the 2-way table below:

	Placebo	Vaccine	Total
HIV +	105	106	211
HIV -	1168	1167	2335
Total	1273	1273	2546

Does the data provide evidence that the HIV infection rate and the vaccine are independent?

Assumption:

- 1) Each cell has expected count more than 5
- 2) Randomization is satisfied.

H_0 : The HIV infection rate and the vaccine are independent.

H_a : The HIV infection rate and the vaccine are not independent.

$$\chi^2 = 0.00237 + 0.00237 + 0.000214 + 0.000214 = 0.005168$$

$\sim \chi^2$ distribution with $df=(2-1)(2-1)=1$

$p\text{-value} = P(\chi^2 > 0.0052)$, so that $p\text{-value} > 0.1$.

Do not reject H_0 . The data doesn't provide evidence that the HIV infection rate was impacted by the vaccine.

24 ANOVA (ANalysis Of VAriance)

Are the Means of Several Groups Equal?

- We already know how to test whether *two* groups have equal means (Ch 19)
- When we want to test whether more than 2 groups have equal means, we could compare each pair of groups with a *t*-test.
- However, we'd wind up increasing the probability of a Type I error, since each test would bring with it its own α .
- Fortunately, there is a test that generalizes the *t*-test to *any* number of treatment groups.
- For comparing several means, there is yet another sampling distribution model, called the *F*-model.

Here, we introduce the tool, ANOVA, for comparing the means of at least two populations. It will provide us with the opportunity to make general conclusions by finding an overall error probability α .

There are 2 possible conclusions:

1. There are some differences among the means at a level of significant α .
2. There are no differences among the means at a level of significant α .

Example: We want to compare three brands of gasoline.

Scenario 1			
	Brand1	Brand2	Brand3
	15	19	22
	19	17	17
	14	16	19
	16	20	18
Average	16	18	19
St. Dev.	2.16	1.83	2.16

- It looks like these observations could have occurred from treatments with the same means.

Let's consider another scenario:

Scenario 2			
	Brand1	Brand2	Brand3
	15.2	18.5	19.6
	16.1	17.5	19.3
	16.8	18.2	18.4
	15.9	17.8	18.7
Average	16	18	19
St. Dev.	0.66	0.44	0.55

- It's easy to see that the means in the second set differ.
 - o It's hard to imagine that the means could be that far apart just from natural sampling variability alone.

NOTE: the two sets of mean travel distances in both scenarios are the same. (They are 16, 18, and 19, respectively.) Then why do the figures look so different?

- In order to assess whether these sample means differed because their respective population means actually are different, we need to introduce variability.
- Two types of variability:
 - 1) variability *within (inside)* the groups
 - 2) variability *between* the groups (ie. Difference between the means of the groups)
- In the first scenario, the differences among the means look as though they could have arisen just from natural sampling variability from groups with equal means, so there's not enough evidence to reject H_0 .
- In the second scenario, the variation *within* each group is so small that the differences *between* the means stand out.
- And it's the central idea of the F -test.

$$F = \frac{\text{variation between the groups}}{\text{variation within the groups}}$$

- We compare the difference *between* the means of the groups with the variation *within* the groups
- When the differences between means are large compared with the variation within the groups, we reject the null hypothesis and conclude that the means are (probably) not equal.
- We have an estimate from the variation *within* groups.
 - Traditionally, it's called the **Error Mean Square** (or sometimes **Within Mean Square**) and written MS_E .
 - It's just the variance of the residuals.
 - Because it's a pooled variance, we write it s_p^2
- We've got a *separate* estimate from the variation *between* the groups.
 - We call this quantity the **Treatment Mean Square** (or sometimes **Between Mean Square**) denoted by MS_T .
 - We expect it to estimate σ^2 , if we assume the null hypothesis is true.

The **F**-Statistic

- When the null hypothesis is true and the treatment means are equal, both MS_E and MS_T estimate σ^2 , and their ratio should be

close to 1.

- We can use their ratio, the F -statistic $F = \text{MS}_T/\text{MS}_E$ to test the null hypothesis.
- Just like Student's t , the F -models are a family of distributions. However, since we have 2 variance estimates, we have 2 degrees of freedom parameters:
 - MS_T estimates the variance of the treatment means and has $\text{df}_1 = k - 1$ when there are k groups
 - MS_E is the pooled estimate of the variance within groups. If there are a total of N observations and there are k groups, MS_E has $\text{df}_2 = N - k = k(n - 1)$, where n is the number of observations in each of the k groups.
- You'll often see the Mean Squares and other info put into a table called the ANOVA table.
- For the soap example in the book, the ANOVA table is:

Source	Sum of Squares	Analysis of Variance Table			F-ratio	P-value
		DF	Mean Square			
Soaps	29882	3	9960.64		7.0636	0.0011
Error	39484	28	1410.14			
Total	69366	31				

- The ANOVA table was originally designed to organize the calculations.

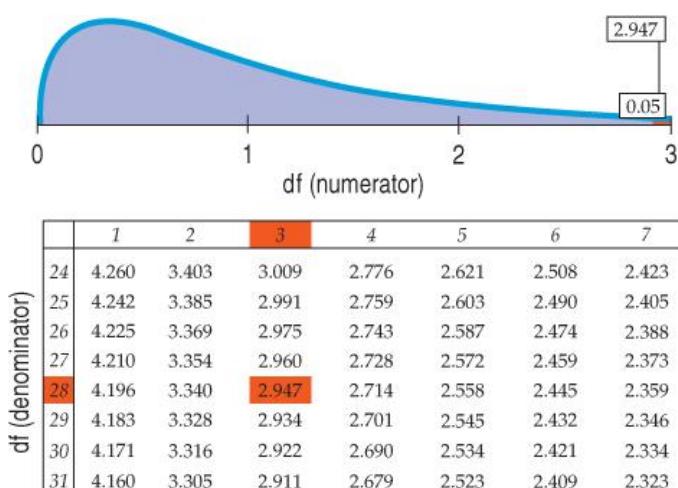
- With advances in technology, we get all of this info, but we only need to look at the F -ratio and the p -value
- Usually, you'll get the P-value for the F -statistic from technology. Any software program performing an ANOVA will automatically "look up" the appropriate one-sided p-value for the F -statistic.
- If you want to do it yourself, you'll need an F -table (Appendix C)

The F -table:

- give the critical value of the F -statistic with the appropriate number of degrees of freedom determined by your data, for the α -level that you select.
- If your F -statistic is greater than that value, you know that its P-value is less than that α level.
- So, you'll be able to tell whether the P-value is greater or less than 0.05, 0.01, or 0.001, but to be more precise, you'll need technology.

Figure 25.4

Part of an F -table showing critical values for $\alpha = 0.05$, and highlighting the critical value, 2.947, for 3 and 28 degrees of freedom. We can see that only 5% of the values will be greater than 2.947 with this combination of degrees of freedom.



ANOVA assumptions

1. We have k *independent* random samples, one from each of the k populations.
2. The data *within* each treatment group must be independent.
3. The i^{th} population has a *normal distribution* with unknown mean μ_i , where $i = 1, \dots, k$. The means may be different.
(Check by a histogram or a Normal probability plot of all the residuals together; Check for outliers for each treatment group)
4. All the populations have the *same standard deviation* σ , whose value is unknown.

The **hypotheses** in ANOVA are:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

$$H_a: \text{the means are not all equal}$$

The **F-statistic** is: $F\text{-statistic} = \frac{MST}{MSE} = \frac{SST / I - 1}{SSE / n - I}$

The F distributions

1. are a family of distributions with two parameters: the degrees of freedom in the numerator and denominator;

2. interchanging the degrees of freedom changes the distributions;
3. are right-skewed;
4. have no probability below 0;
5. the peak of the density curve is near 1;

P-value

p-value = $P(F > F_0)$ with $df_1 = k - 1$ and $df_2 = N - k$;

(ie. The p-value for H_0 is the probability of obtaining a sample with an F-statistic greater than the observed one if H_0 is true.)

NOTE: the alternative in the ANOVA is always the same. It is not one-sided or two-sided, but multi-sided. P-values are always and only calculated as the probability in the right tail of an F-distribution.

ANOVA Table for k Independent Random Samples

Source	df	SS	MS	F	p-value
Treatments	$k-1$	SS_T	$MS_T = SS_T/(k-1)$	MS_T/MS_E	p-value
Error	$N-k$	SS_E	$MS_E = SS_E/(N-k)$		
Total	$N-1$	Total SS			

Example (Continued):

Using the Data from Scenario 1, the following ANOVA summary is resulted:

ANOVA

<i>Source of Variation</i>	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>
Between Groups	2	18.66667	9.333333	2.210526	0.165597
Within Groups	9	38	4.222222		
Total	11	56.66667			

Example:

In an effort to improve the quality of recording tapes, the effects of four kinds of coatings A, B, C, D on the reproducing quality of sound are compared.

The following values on distortion are obtained:

	A	B	C	D
	10	14	17	12
	15	18	16	15
	8	21	14	17
	12	15	15	16
	15		17	15
			15	15
			18	
Average	12	17	16	15
Variance	9.5	10	2	2.8

The following ANOVA output is obtained:

ANOVA					
<i>Source of Variation</i>	SS	df	MS	F	P-value
Between Groups	68	3	22.66667	4.340426	0.018136
Within Groups	94	18	5.222222		
Total	162	21			

With the help of such a sample, we want to decide if the four different coatings result in different mean distortions at $\alpha = 0.05$