Guided Notes and Practice Problems Part 1

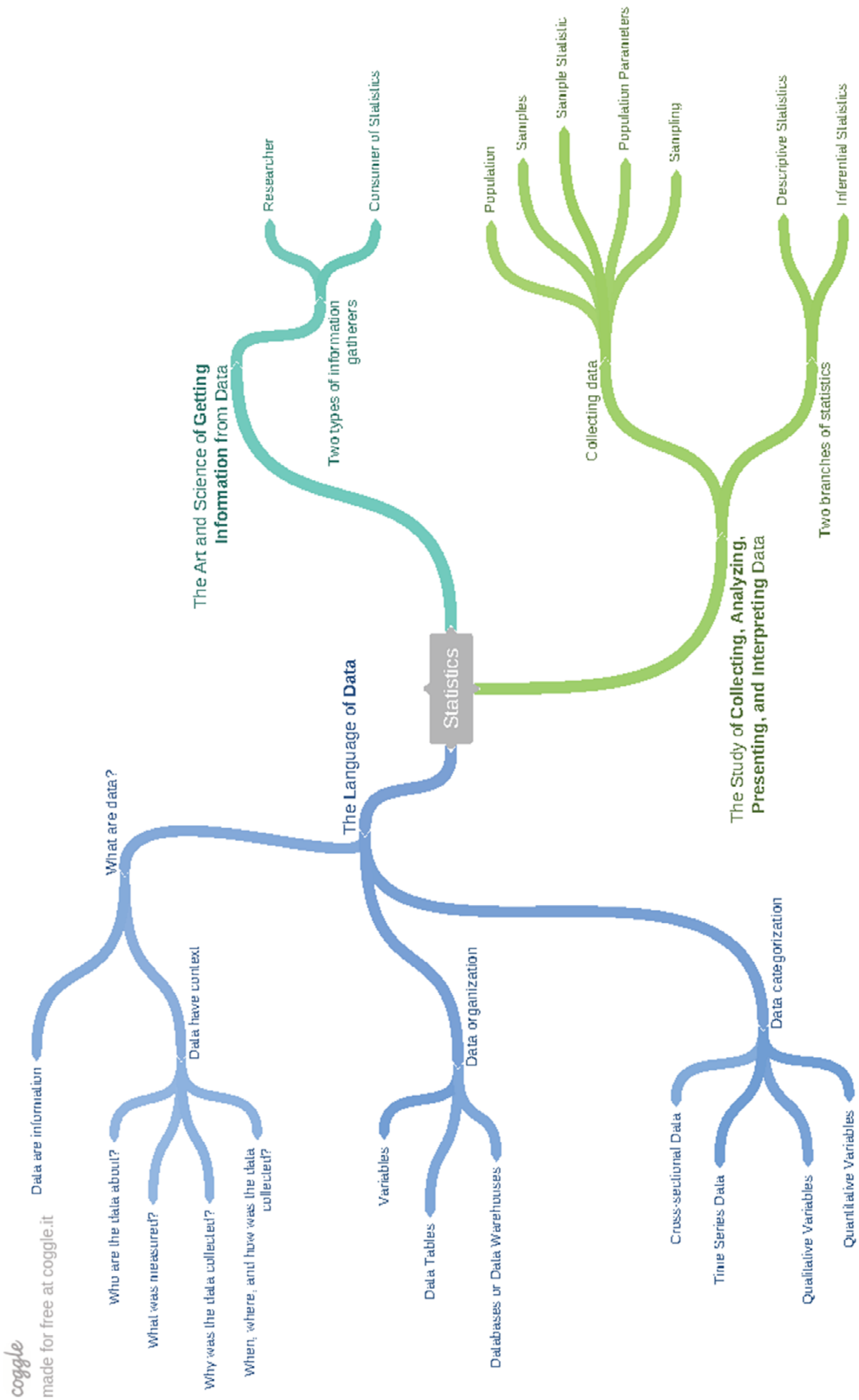
QSIS 285: Business Statistics

Lisa Over

ABSTRACT

This packet includes guided notes, examples, and exercises for chapters 1 through 6 from the course text "Business Statistics: Communicating with Numbers" Second Edition by Jaggia/Kelly.

Table of Contents


Statistics
The Art and Science of Getting Information from Data
Two types of information gatherers
Researcher
Consumer of Statistics
The Study of Collecting, Analyzing, Presenting, and Interpreting Data
Collecting data
Population
Samples
Sample Statistic
Population Parameters
Sampling
Two branches of statistics
Descriptive Statistics
Inferential Statistics
The Language of Data
What are data?
Data are information
Data have context
Who are the data about?
What was measured?
Why was the data collected?
When, where, and how was the data collected?
Data organization
Variables
Data Tables
Databases or Data Warehouses
Data categorization
Cross-sectional Data
Time Series Data
Qualitative Variables
Quantitative Variables
coggle
made for free at coggle.it

1 Statistics and Data

1.1 Objectives

- Describe the field of statistics using the terms population, sample, parameter, statistic, sampling, descriptive statistics, and inferential statistics.
- Evaluate the context of data by asking questions about the who, what, why, where, when, and how of the data.
- Describe the organization of data using the terms variables, data tables, databases, and data warehouses.
- Distinguish between qualitative and quantitative variables, cross-sectional and time series data, and their applications in business and economics.

1.2 What is Statistics?

- The language of data
- The Art and Science of Getting Information from Data
- The Study of Collecting, Analyzing, and Interpreting Data

1.3 Data Context

Data are information, which is comprised of facts or characteristics about a subject of interest. You must know the context of the data before referencing or analyzing it.

- *Who are the data about?* Who are the subjects of the data? Subjects can be people, objects, events, etc. This includes any demographic information that describes the subject, such as country, gender, or any other group label.
- *What was measured?* What characteristics are measured about each subject, e.g., demographic information, sales information, opinions, costs, rates, time, etc.
- *Why was the data collected?* What business problem was the data collected to support?
- *Where, when, and how was the data collected?* The source of the data can make the difference between insight and nonsense.
    - When were the data collected? Time – day, year, etc.
    - Where were the data collected? From websites, in person interview, etc.
    - How were the data collected? How were subjects selected? For surveys, how were subjects contacted – phone call, email?

1.3.1 Example Problem: Data Context

The following table shows the Fortune 500 rankings of America's largest Corporations for 2010. Next to each corporation are its market capitalization (in billions of dollars as of March 26, 2010) and its total return to investors for the year 2009. This data was obtained from Fortune.com and from each corporation's annual reports. [1]

Company	Mkt Cap. (in $ billions)	Return
Wal-Mart	209	-2.7
Exxon Mobil	314	-12.6
Chevron	149	8.1
General Electric	196	-0.4
Bank of America	180	7.3
ConocoPhillips	78	2.9
AT&T	155	4.8
Ford Motor	47	336.7
JP Morgan Chase	188	19.9
Hewlett-Packard	125	43.1

1. Who are the data about?
   The data are about large American Fortune 500 corporations.
2. What was measured?
   Marketing capital in billions as of March 26, 2010
   Return to investors for the year 2009
3. Why was the data collected?
   What are some questions you could answer with this data? The problem scenario does not specify.
4. When, where, and how was the data collected?
   Obtained from Fortune.com and from annual reports.
5. What are the variables?
   Marketing capital – Mkt Cap.
   Return to investors – Return

[1] Chapter 3, Section 4, Problem 26, Page 82

1.4 Data Organization

A characteristic observed about people, objects, or events is called a variable because the values often **differ in kind or degree among the various subjects**. The values of the characteristics are organized into a data table with each row representing a subject and each column representing a variable. Data are organized and stored in a form that supports efficient movement or processing, i.e., electronically in databases and data warehouses.

WHAT

WHO

Company	**Mkt Cap. (in $ billions)**	**Return**
Wal-Mart	209	-2.7
Exxon Mobil	314	-12.6
Chevron	149	8.1
General Electric	196	-0.4
Bank of America	180	7.3
ConocoPhillips	78	2.9
AT&T	155	4.8
Ford Motor	47	336.7
JP Morgan Chase	188	19.9
Hewlett-Packard	125	43.1

DATA TABLE

1.5 Types of Data

1.5.1 Cross-Sectional

Cross-sectional data is a set of data points collected by observing many subjects (such as individuals, firms, countries, or regions) at the same point of time, or without regard to differences in time.

Example

The following table shows the Fortune 500 rankings of America's largest Corporations for 2010. Next to each corporation are its market capitalization (in billions of dollars as of March 26, 2010) and its total return to investors for the year 2009. This data was obtained from Fortune.com and from each corporation's annual reports. [2]

- No time column
- Values represent one time period
- Cannot compare the characteristics over a period of time

Company	**Mkt Cap. (in $ billions)**	**Return**
Wal-Mart	209	-2.7
Exxon Mobil	314	-12.6
Chevron	149	8.1
General Electric	196	-0.4
Bank of America	180	7.3
ConocoPhillips	78	2.9

[2] Chapter 3, Section 4, Problem 26, Page 82

1.5.2 Time Series

Time series data is a set of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive, equally spaced points in time such as daily, weekly, monthly, quarterly, annually, etc.

Example

Elizabeth feels she is ready to invest some of her earnings. She investigates two mutual funds from Janus Capital Group using data her financial planner obtained directly from the company. The following table compares the annual returns (in percentages) of the two mutual funds over the past 10 years. [3]

- Has a time column
- One year represents one time period
- Can compare the characteristics over the one-year periods of time

Year	Janus Balanced Fund	Janus Overseas Fund
2000	-2.16	-18.57
2001	-5.04	-23.11
2002	-6.56	-23.89
2003	13.74	36.79
2004	8.71	18.58
2005	7.75	32.39
2006	10.56	47.21
2007	10.15	27.76
2008	-15.22	-52.75
2009	24.28	78.12

[3] Chapter 3, Case Studies, Case Study 3.2, Page 102

1.5.3 Practice Problems: Types of Data

Note: Cross-sectional data and time series data are equally valuable in different types of research.

Classify the following data scenarios as cross-sectional (C) or (T) time series…

_________ The test scores of students in a class

_________ The current average prices of regular gasoline in different states

_________ The sales prices of single-family homes sold last month in California

_________ GDP of the United States from 1990-2010

_________ Daily price of DuPont stock during the first quarter

_________ Quarterly housing starts collected over the last 60 years

_________ Results of market research testing consumer preferences for soda

_________ The 2011 year-end book value per share for all companies listed on the New York Stock Exchange

_________ The stock price for Google at the end of the past four quarters

_________ The price of oil over the past 10 years

_________ The sale prices of townhouses sold last year

_________ Starting salaries of recent business graduates at Penn State University

C, C, C, T, T, T, C, C, T, T, C, C

1.6 Variables

There are two types of variables:

Qualitative and **quantitative**

There are four scales of measurement, two for each type of variable:

Nominal, **ordinal**, **interval**, and **ratio**

Variable types and scales of measurement describe the nature of information assigned to the variables.

A qualitative (Categorical) variable assumes **labels or names** to identify the characteristic. Qualitative variables are described as either nominal or ordinal.

A quantitative variable assumes **numeric values**. Quantitative variables are described as either interval or ratio.

1.6.1 Types of Variables

Qualitative	Quantitative	
	Continuous	Discrete
Description Assumes labels or names to identify the characteristics	**Description** Assumes numeric values Infinitely uncountable within an interval, i.e., can take any value within an interval	**Description** Assumes numeric values Countable number of values
Examples Last name Gender Marital status Religious affiliation Zip code Social security number Opinion yes/no Names of companies Class status Performance rating 1 to 5 Poor to excellent	**Examples** Time spent studying Amount of money spent on groceries Price of cars Temperature Distance Speed	**Examples** Number of items sold Number of social media sites you use Number of children Number of cars you own Number of bathrooms in your house

1.6.2 Scales of Measurement

Nominal	Ordinal	Interval	Ratio
Type Qualitative	**Type** Qualitative	**Type** Quantitative	**Type** Quantitative
Description • Categorize a characteristic Least sophisticated scale	**Description** • Categorize a characteristic • Rank	**Description** • Categorize a characteristic • Rank • Meaningful difference between values	**Description** • Categorize a characteristic • Rank • Meaningful difference between values • Has a true zero point Most sophisticated scale
Examples Last Name Gender Marital status Opinion yes/no	**Examples** Performance rating • 1 to 5 • Excellent to poor Class status • Freshman, sophomore, junior, senior	**Examples** Temperature • Diff between 20° & 30° is 10° • Diff between 60° & 70° is 10° Cannot construct ratios: 80° is not twice as hot as 40°	**Examples** Distance Price of gasoline Price of gold • Diff between \$20 & \$30 is \$10 • Diff between \$60 & \$70 is \$10 Construct ratios: \$40 is twice as much as \$20

1.7 Researcher

A researcher **studies a problem using statistical methods** and reports or presents the information obtained.

1.8 Consumer of Statistics

A consumer of statistics **reads statistical reports to obtain information** about a problem.

Note: Only trust research that is adequately supported by valid statistical methods and theories.

1.9 Population vs. Sample

Population	Sample

• A **population** consists of all items or subjects of interest in a statistical problem.
• The population is typically too large to study directly.
• A **population parameter** describes a characteristic of the population.
  ○ Average salary of all high school teachers
  ○ Median family income for all residents in a particular state
	• A **sample** is a representative subset of the population.
	• A **sample statistic** describes a characteristic of a sample and estimates a population parameter.
	  ○ Average salary for a sample of high school teachers
	  ○ Median family income for a sample of families from a particular state

1.10 Sampling

- The process of selecting a subset of a population to study.
- It is cost prohibitive and/or infeasible to study the entire population so statisticians study smaller subsets of the population, i.e., samples.
- Used to estimate population parameters.
- Used heavily in manufacturing and service settings to ensure high quality products and services.

1.11 Branches of Statistics

Descriptive statistics is the branch of statistics concerned with **numeric (averages, percentages, etc.) and graphical summaries of data**.

Inferential statistics is the branch of statistics concerned with the problem of **estimating population parameters and testing hypotheses about the parameters**.

1.11.1 Practice Problems: Sampling

Sampling is appropriate in settings where processes can be standardized. Select the settings below in which sampling would be appropriate…

- ❑ Computer assembly
- ❑ Custom cabinet making
- ❑ Cell phone manufacturing
- ❑ Technical support by phone

State whether the following data scenarios require sampling (Yes) or not (No)…

__________ US Unemployment rate

__________ Average salary for American high school teachers

__________ Total rainfall in Phoenix, Arizona in 2010

__________ The Cleveland Indians' hitting average in 2010

__________ The average SAT score of incoming Freshman at a university

__________ The average life of light bulbs produced by a manufacturer

__________ The percentage of US school teachers who support democrats

__________ The average height of NBA players

__________ The average content of cereal boxes produced by a manufacturer

Computer assembly, Cell phone manufacturing, Technical support by phone

Yes, Yes, No, No, No, Yes, Yes, No, Yes

1.11.2 Diagram: The Big Idea of Statistics

2 Tabular and Graphical Methods

2.1 Objectives

- Construct and interpret frequency distributions, pie charts, and bar graphs to summarize qualitative data.
- Construct and interpret grouped frequency distributions, cumulative frequency distributions, and histograms to summarize quantitative data.
- Construct side-by-side bar charts to summarize the relationship between two qualitative variables.
- Construct scatterplots to summarize the relationship between two quantitative variables.

2.2 Qualitative Data

The numerical sample statistics for summarizing one or more categorical variables is a proportion or a percent. Frequency distribution tables, pie charts, and bar charts are visual summarizations of qualitative data.

Frequency distribution tables, pie charts, and bar charts must divide a "whole" into categories. All frequencies must sum to the total number of observations, all relative frequencies must sum to 1, and all percent frequencies must sum to 100.

2.2.1 Frequency Distribution Tables

A frequency distribution summarizes qualitative data by grouping the data into categories and recording the number of observations in each category and the relative frequency (proportion) of observations in each category.

Frequency = count of elements

Relative frequency = (count of elements/total number of elements)

Example

Color	Frequency	Relative Frequency
Brown	112	0.0892
Yellow	105	0.0837
Red	109	0.0869
Orange	327	0.2606
Green	314	0.2502
Blue	288	0.2295
Total	1255	1

Table 1 Frequency and Relative Frequency Distribution for the Colors of Plain M&Ms

2.2.2 Pie Charts

A pie chart is a segmented circle whose segments portray the relative frequencies or percent frequencies of the categories of some qualitative variable. The size of the segments are proportional to the values depicted.

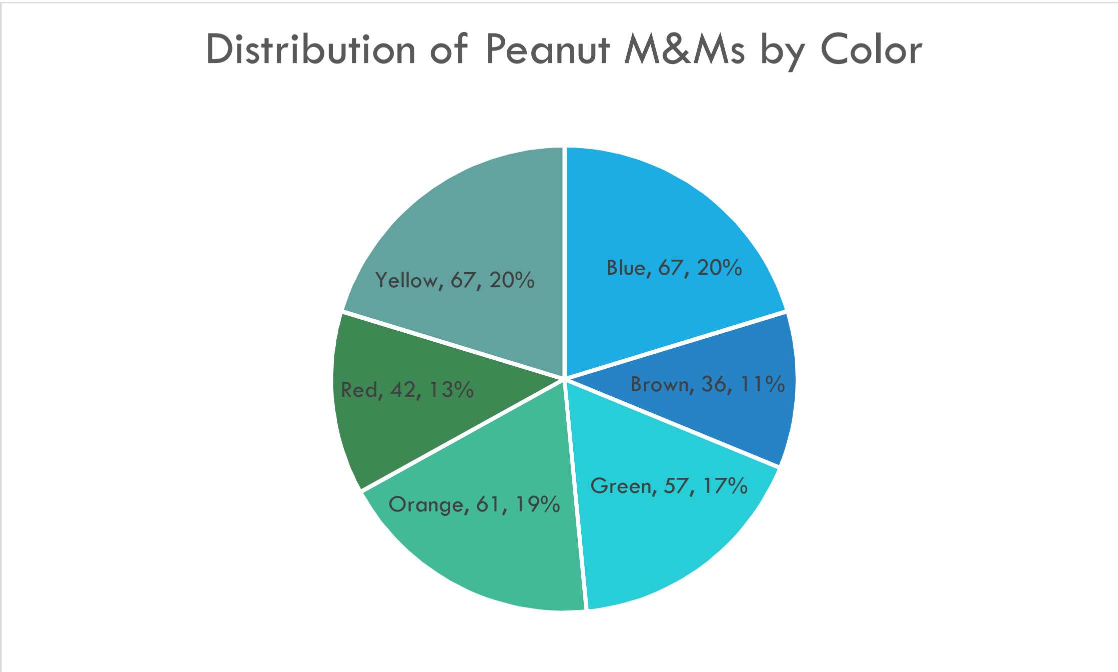


Figure 1 Pie Chart Illustrating the Distribution of Peanut M&Ms by Color

Important Considerations for Creating Pie Charts

- Colors must be distinguishable from one another.
- Legends or data labels (one per wedge) must be clear and easy to read.
- Use a bar chart if two or more categories are equal or differ only slightly. It is difficult to tell which pie slice is greatest or smallest or if they are equal. A bar chart better portrays slight differences in size and equality better than a pie chart.
- Use the simplest graph possible to convey the information in the data. Be clear, clean, and professional.
- Any graphical display of the distribution for a categorical variable should include two important statistics:
  - Percent or frequency for each category
  - Total sample size

  Sample size can be incorporated into a title or caption if it does not appear on the body of the chart.
- The chart must have a title.

2.2.3 Bar Charts

A bar chart is a series of horizontal or vertical bars where the bars portray the frequency or relative frequency for each category of some qualitative variable. The length of the bars are proportional to the values depicted.

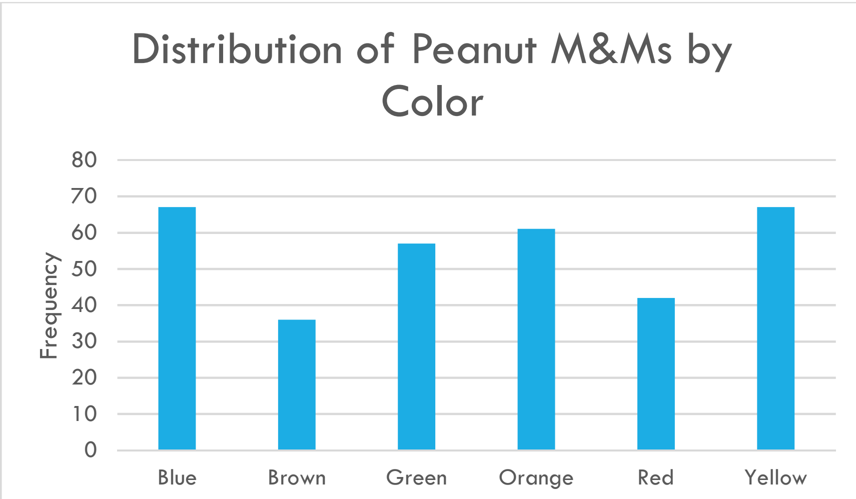


Figure 2 Bar Chart Illustrating the Distribution of Peanut M&Ms by Color

Important Considerations for Creating Bar Charts

- The axes must be marked clearly with numbers or category names. The axis with the frequency (proportion or percentage) measure has a title: Frequency, Relative Frequency (or Proportion), or Percent Frequency (or Percentage).
- The bars must have the same width. Visual distortions will distort the interpretation. Excel should do this automatically.
- The bars of a bar chart must be separated by spaces to create a visual separation of the categories because the items in each category are clearly separate from one another.
- The vertical axis must portray the differences between categories. A high upper limit of the vertical axis compresses the graph so that differences are not noticeable. A low upper limit that creates very small increments between marks stretches the graph so that differences are pronounced.
- Use the simplest graph possible to convey the information in the data. Be clear, clean, and professional.
- Any graphical display of the distribution for a categorical variable should include two important statistics:
  - Percent or frequency for each category
  - Total sample size

  Sample size can be incorporated into a title or caption if it does not appear on the body of the chart.
- The chart must have a title.

2.2.4 Contingency Tables (Cross-tabulation)

A contingency table shows the distribution of one variable in rows and another in columns and is used to study the relationship between the two variables.

When studying two categorical variables, the proportions or percentages should be represented as either row or column percentages to allow comparison of how one variable may influence the responses of the other variable.

Row Labels	Peanut M&M	Plain M&M	Grand Total
Blue	67	288	355
Brown	36	112	148
Green	57	314	371
Orange	61	327	388
Red	42	109	151
Yellow	67	105	172
Grand Total	**330**	**1255**	**1585**

Table 2 Contingency Table Summarizing the Relationship Between M&M Type and Color

2.3 Practice Problems: Qualitative Data

The following exercises are also in the Excel file:
Qualitative Data Practice Problems.xlsx

Answers to the following exercises are in the Excel file:
Qualitative Data Practice Problems KEY.xlsx

2.3.1 Exercise 1: Auto Parts Chain

An auto parts chain asked customers to complete a survey rating the chain's customer service as average, above average, or below average. The following table shows the survey results.

Average	Below Average	Average
Above Average	Above Average	Above Average
Below Average	Average	Average
Above Average	Average	Below Average
Below Average	Below Average	Average

Rating	Frequency	Relative Frequency	Percent Frequency
Below Average	5	0.33	33
Average	6	0.40	40
Above Average	4	0.27	27

1. The proportion of customers who felt the customer service was Average is closest to _______.

2. A rating of Average or Above Average accounted for what number of responses to the survey?

2.3.2 Exercise 2: Professor Smith

Students in Professor Smith's business statistics course have evaluated the overall effectiveness of the professor's instruction on a five-point scale, where a score of 1 indicates very poor performance and a score of 5 indicates outstanding performance. The following table shows the results.

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

Rating	Frequency	Relative Frequency	Percent Frequency
1	1	0.03	3.33
2	5	0.17	16.67
3	7	0.23	23.33
4	8	0.27	26.67
5	9	0.30	30.00

1. What is the most common score given in the evaluations?

2. What percentage of students gave professor Smith an evaluation of at least 4?

3. What percentage of students gave Professor Smith an evaluation of 2 or less?

4. What is the relative frequency of the students who gave Professor Smith an evaluation of 3?

2.3.3 Exercise 3: US Poverty Level

The Statistical Abstract of the United States, 2010 provided the following frequency distribution of the number of people who live below the poverty level by region.

Region	Number of People (in 1000s)	Relative Frequency	Percent Frequency
Northeast	7,174	0.18	17.77
Midwest	8,137	0.20	20.16
South	16,457	0.41	40.77
West	8,593	0.21	21.29

1. What is the percentage of people who live below the poverty level in the West or Midwest?

2.3.4 Exercise 4: City Building Repair

A city in California spent $6 million repairing damage to its public buildings in 2010. The following table shows the categories where the money was directed.

Cause	Frequency	Relative Frequency	Percent Frequency
Termites	1,560,000	0.26	26%
Water Damage	480,000	0.08	8%
Mold	540,000	0.09	9%
Earthquake	1,320,000	0.22	22%
Other	2,100,000	0.35	35%

1. How much did the city spend to fix damage caused by mold?

2. How much more did the city spend to fix damage caused by termites compared to the damage caused by water?

2.3.5 Exercise 5: CBS News Survey

A survey conducted by CBS news asked 1,026 respondents: "What would you do with an unexpected tax refund?" The responses are summarized in the following table.

Category	Frequency	Relative Frequency	Percent Frequency
Pay off debts	523	0.51	51%
Put it in the bank	257	0.25	25%
Spend it	92	0.09	9%
I never get a refund	92	0.09	9%
Other	62	0.06	6%

1. How many people will either put it in the bank or spend it?

2.3.6 Exercise 6: Busy Airports

The world's busiest airports by passenger traffic for 2010…

Name	Location	# of Passengers (in millions)	Relative Frequency	Percent Frequency
Hartsfield-Jackson	Atlanta, Georgia, United States	93	0.26	25.55
Capital International	Beijing, China	76	0.21	20.88
London Heathrow	London, United Kingdom	70	0.19	19.23
O'Hare	Chicago, Illinois, United States	65	0.18	17.86
Tokyo	Tokyo, Japan	60	0.16	16.48

1. The percentage of passenger traffic in the five busiest airports that occurred in Asia is the closest to ______.

2. How many more millions of passengers flew out of Atlanta than flew out of Chicago?

2.3.7 Exercise 7: Children's Library

	Number of Unique Titles	Relative Frequency
Rick Riordan	6	0.08
CS Lewis	12	0.16
J.K. Rolling	8	0.10
Orson Scott Card	9	0.12
Erin Hunter	32	0.42
Lois Lowry	4	0.05
Suzanne Collins	3	0.04
Veronica Roth	3	0.04

1. Create pie and bar charts for the data.

2. What is the scale of measurement?

3. Which author authored the most unique titles in the library?

2.3.8 Exercise 8: Marital Status

	1960	2010
Married	0.71	0.52
Single	0.15	0.28
Divorced	0.05	0.14
Widowed	0.09	0.06

1. Create pie and bar charts for the data.

2. What percentage of adults were married in 1960?

3. What percentage of adults were married in 2010?

2.4 Quantitative (Numeric) Data

Frequency distribution tables, histograms, and box plots are visual summarizations of quantitative data.

2.4.1 Grouped Frequency Distribution Tables

A frequency distribution summarizes quantitative data by grouping the data into user specified **classes** and recording the number of observations that fall within the range of each **class** and the relative frequency of observations in each **class**.

A frequency distribution for quantitative data also shows the cumulative frequency and cumulative relative frequency for the ordered classes.

Class and Data Grouping Requirements

- All classes are mutually exclusive - no data point can be in more than one class.
- All classes are collectively exhaustive - all data points must fit into one of the class designations.
- The total number of classes range between 5 and 20.

Width of a class = (Max value - Min value) / (Number of Classes)

All bins must be the same width. It may be necessary to start the first bin with a number lower than the lowest in the dataset or to end the last bin with a number that is greater than the largest number in the dataset.

Example

Weights	**Frequency**	**Cumulative Frequency**	**Relative Frequency**	**Cumulative Relative Frequency**
0.35 up to 0.4	12	12	0.0992	0.099173554
0.4 up to 0.45	33	45	0.2727	0.371900826
0.45 up to 0.475	37	82	0.3058	0.67768595
0.475 up to 0.5	4	86	0.0331	0.710743802
>0.5	35	121	0.2893	1
Grand Total	**121**		**1**	

Table 3 Grouped Frequency, Cumulative Frequency, Relative Frequency, and Cumulative Relative Frequency Distribution for the Weight of 121 Packages of M&Ms

2.4.2 Histograms

A **histogram** is a series of adjacent rectangles (no spaces between) where the width and height of each rectangle represent the class width and frequency (or relative frequency) of the respective class. A histogram displays the overall distribution of the data not the distribution of individual values.

A histogram is the best graphical tool for displaying the relative frequency of grouped quantitative data.

There is **no space** between bars because the items in each class are not that distinct from one another. For example, define the classes "6.0-10.9" and "11.0-15.9." There is not much difference between 10.9 and 11.0.

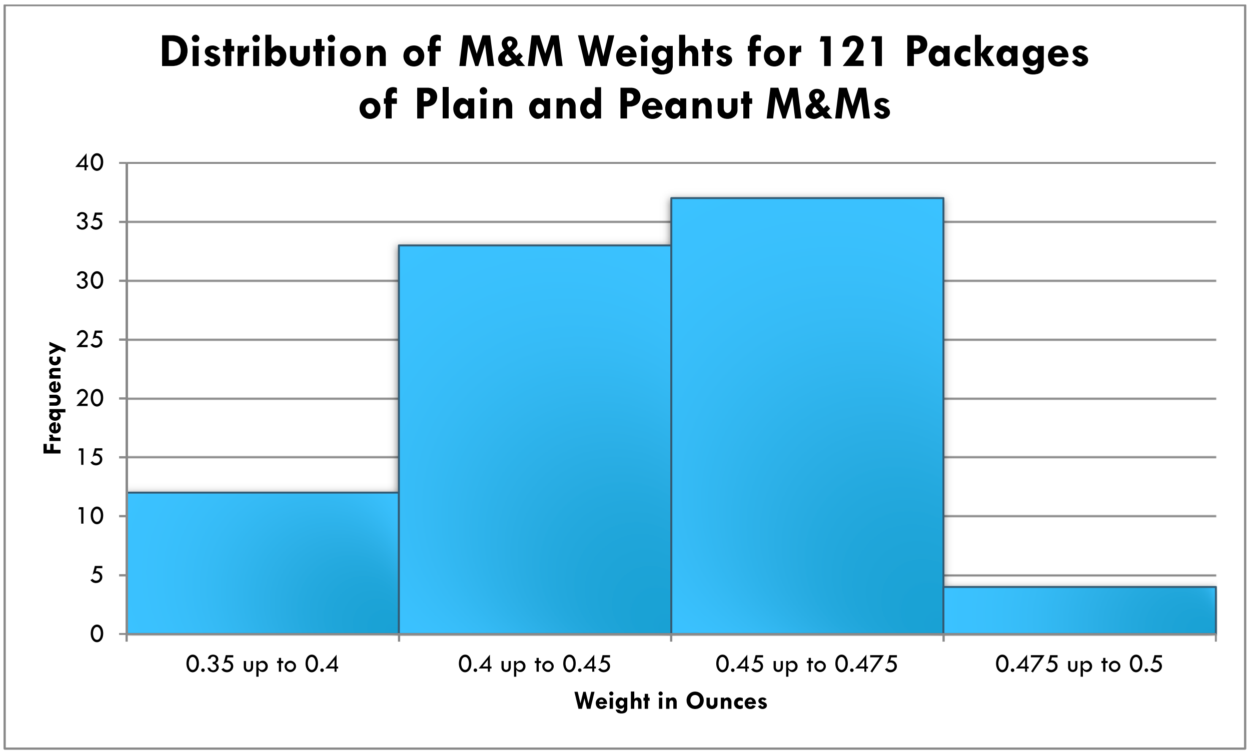


Figure 3 Histogram Illustrating the Distribution of Weight for 121 Packages of Plain and Peanut M&Ms

2.4.3 Scatterplots

A scatterplot is a graphical tool that illustrates how two variables are related, i.e., how one variable affects another. The two variables are graphed as coordinate pairs (x, y).

DATA

Avg. Daily Auto Traffic (000s)	Annual Sales ($000s)
62	1121
35	766
36	701
72	1304
41	832
39	782
49	977
25	503
41	773
39	839
35	893
27	588
55	957
38	703
24	497
28	657
53	1209
55	997
33	844
29	883

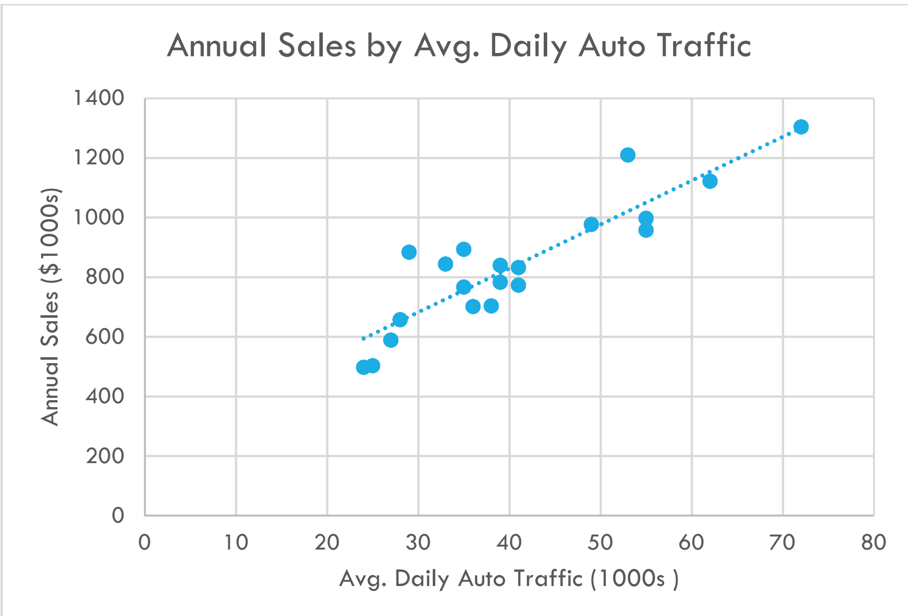


Figure 4 Scatterplot Illustrating the Relationship Between Annual Sales and Daily Automobile Traffic

2.5 Practice Problems: Quantitative Data

The following exercises are also in the Excel file:
Quantitative Data Practice Problems.xlsx

Answers to the following exercises are in the Excel file:
Quantitative Data Practice Problems KEY.xlsx

2.5.1 Exercise 1: Midwestern Homes

The following data represent the recent sales price (in $1,000s) of 24 homes in a Midwestern city.

187	125	165	170	230	139	195	229
239	135	188	210	228	172	127	139
122	181	196	237	115	199	170	239

1. Sort the data in ascending order.
2. Determine the number of classes. Usually 5 to 20 classes is the rule.
3. Count and record the number of data points that fall into each class (frequency).
4. Calculate and record the cumulative frequency.
5. Calculate and record the relative frequency and cumulative relative frequency.

Grouped Sales Prices	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency

1. Suppose the data on house prices will be grouped into five classes. The width of the classes for a frequency distribution or histogram is the closest to _______.
2. Suppose the data are grouped into five classes, and one of them will be "115 up to 140." - that is, {x; 115 ≤ x < 140}. The relative frequency of this class is ______.
3. Suppose the data are grouped into five classes, and one of them will be "165 up to 190." - that is, {x; 165 ≤ x < 190}. The frequency of this class is _____.

2.5.2 Exercise 2: Statistics Quiz

The following data represent scores on a pop quiz in a statistics section.

16	16	17	32	32	33	37	44	45	47
55	56	56	62	66	70	72	74	82	84

Note: The data are already sorted.

Instructions

1. Create a frequency distribution for this data.
2. Answer the questions below.

Grouped Scores	**Frequency**	**Cumulative Frequency**	**Relative Frequency**	**Cumulative Relative Frequency**

1. Suppose the data on quiz scores will be grouped into five classes. The width of the classes for a frequency distribution or histogram is ______.

2. Suppose the data are grouped into five classes, and one of them will be "30 up to 44." that is, $\{x; 30 \le x < 44\}$. The frequency of this class is _____.

3. Suppose the data are grouped into five classes, and one of them will be "30 up to 44" —that is, $\{x; 30 \le x < 44\}$. The relative frequency of this class is _____.

2.5.3 Exercise 3: Eastside HS

Thirty students at Eastside High School took the SAT on the same Saturday. Their raw scores are given next.

1,450	1,480	1,490	1,530	1,590	1,620	1,620	1,640	1,650	1,710
1,740	1,780	1,800	1,800	1,820	1,830	1,830	1,840	1,870	1,900
1,910	1,950	1,950	1,980	2,000	2,010	2,100	2,260	2,350	2,390

Note: The data are already sorted.

Instructions

1. Create a frequency distribution for this data.
2. Answer the questions below.

Grouped Scores	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency

1. Consider a frequency distribution of the data that groups the data in classes of 1400 up to 1600, 1600 up to 1800, 1800 up to 2000, and so on. How many students scored at least 1800 but less than 2000?

2. Consider a frequency distribution of the data that groups the data in classes of 1400 up to 1600, 1600 up to 1800, 1800 up to 2000, and so on. What percent of students scored less than 2200?

3. Consider a frequency distribution of the data that groups the data in classes of 1400 up to 1600, 1600 up to 1800, 1800 up to 2000, and so on. What is the approximate relative frequency of students who scored more than 1600 but less than 1800?

2.6 Summary of Tables and Graphs

2.6.1 Summarizing One Qualitative (Categorical) Variable

- Frequency distribution table
  - Category name
  - Frequency
  - Relative frequency
  - Percent frequency
- Pie chart
  - Portrays the frequency, relative frequency, or percent frequency of each category
  - May not be best graph when the values of two or more categories are close or equal in value (difficult to distinguish the relative size of the wedges)
- Bar chart
  - Portrays the frequency, relative frequency, or percent frequency of each category
  - Bars are separated by space

2.6.2 Summarizing One Quantitative (Numeric) Variable

- Frequency distribution table
  - Class name
  - Frequency
  - Relative frequency
  - Percent frequency
  - Cumulative frequency
  - Cumulative relative frequency
- Histogram
  - Portrays the frequency, relative frequency, or percent frequency of each class
  - Bars are adjacent (touching)
  - Classes are mutually exclusive and exhaustive
  - Displays the overall distribution, which is classified by its shape…
    - Symmetric
    - Skewed right
    - Skewed left

2.6.3 Summarizing Two Qualitative (Categorical) Variables

- Contingency table
  - Shows relationship between two categorical variables
  - Shows counts by category pairs in cross-tabulation format, i.e., categories of one variable make up the rows while categories of the other variable make up the columns
  - Used to study the relationship between two categorical variables
  - Shows the frequencies or percent frequencies of grand total, row, or column
- Side-by-side bar chart
  - Portrays the frequency or relative frequency of each category pair
  - Bar pairs are separated by space

2.6.4 Summarizing Two Quantitative (Numeric) Variables

- Scatterplot
  - Shows relationship between two quantitative variables
  - Graphical display consists of two axes and dots
    - Each axis is marked and labeled to represent one of the variables
    - Each dot represents a pair of observed values
- Types of relationships include…
  - Linear relationship (straight line)
    - Positive (positive slope) – as x increases, y increases
    - Negative (negative slope) – as x increases, y decreases
  - Non-linear relationship (curved)
  - No relationship (scattered, no pattern)

[This page intentionally left blank]

3 Numeric Descriptive Measures

3.1 Objectives

- Compute and interpret measures of central location: mean, median, mode.
- Calculate and interpret percentiles and quartiles.
- Compute and interpret measures of variability: range, interquartile range, variance, and standard deviation.
- Construct and interpret a boxplot.
- Determine outliers using the 1.5*IQR rule.
- Convert data values to Z-scores and interpret the relative location.
- Convert Z-scores to data values.
- Apply the empirical rule to determine the percentage of the data within a specified number of standard deviations from the mean.
- Apply the empirical rule to determine percentile ranks of data values that are within a specified number of standard deviations from the mean.
- Determine outliers using z-scores: outlier < -3 or outlier > 3.
- Calculate and interpret the correlation between two quantitative variables.

3.2 Measures of Central Location

Content examples are in the Excel file:
Measures of Location Practice Problems.xlsx

Quantitative data tends to cluster around some middle or central value. There are three measures of central location: mean, median, and mode.

Measures of Central Location		
Mean	**Median**	**Mode**
Arithmetic Mean $$\frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$ Primary measure of central location Weakness: Influenced by outliers, extremely large or small values Notation Population mean: μ Sample mean: $\bar{x}$	Middle value in the dataset **after sorting** Based only on the number of values, not the magnitude of the values 50% of the data points are less than the median and 50% of the data points are greater than the median. Not influenced by outliers Notation M	Most frequent value or class Seen as the peak(s) on a histogram – a clear rise and fall in frequency or relative frequency Zero modes One mode – unimodal Two modes – bimodal Three or more modes – multimodal The value of this measure diminishes in datasets with more than three modes.

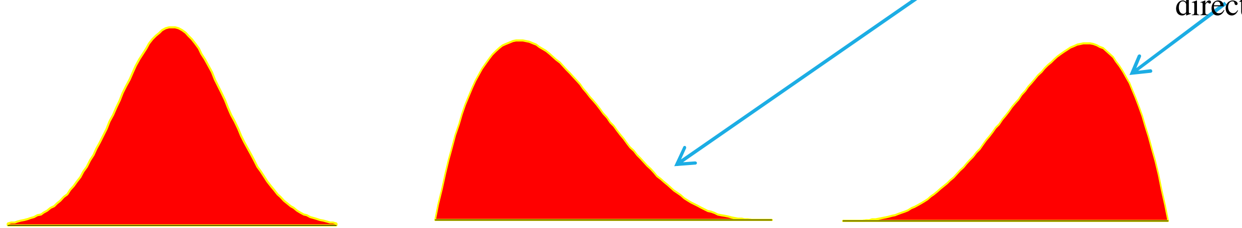
3.2.1 Example Problem: Measures of Central Location

	Employee Salaries	All Get 10% Raise	Only Highest Salary Changes
	39	42.9	39
	45	49.5	45
	65	71.5	65
	67	73.7	67
	78	85.8	78
	78	85.8	78
	85	93.5	200
MEAN	**65.29**	**71.81**	**81.71**
MEDIAN	**67**	**73.7**	**67**
MODE	**78**	**85.8**	**78**

Because the mean involves every data value in the set, it is sensitive to extreme values. The median is not calculated from the values, rather, it is calculated from the number of values and so is not influenced by extreme values.

The mean is pulled away from the median in the positive or right direction.

The mean is pulled away from the median in the negative or left direction.



Mean ≈ Median	**Mean > Median**	**Mean < Median**
Symmetric: Mean and median are approximately equal.	Right or Positively Skewed: Mean is greater than the median.	Left or Negatively Skewed: Mean is less than the median.

3.3 Measures of Location by Position (Percentiles and Quartiles)

Content examples are in the Excel file:
Measures of Location Practice Problems.xlsx

Measures of location by position, percentiles and quartiles, divide the dataset into equal parts (after sorting) and reflect where a data value is relative to the other values.

Measures of Location (Position)	
Percentiles	**Quartiles**
Divides the dataset into 100 equal parts **after sorting**	Divides the dataset into 4 equal parts **after sorting**
Notation $P_1, P_2, P_3, \ldots, P_{100}$	Notation Q_1, Q_2, Q_3, Q_4
P_{30} is the 30th percentile Each percentile cuts the data into two parts, some data point values are less than the given percentile and some are greater than it. The percentage of values in each part depends on the percentile…	Q_1 is the first quartile and the 25th percentile Q_2 is the second quartile, the 50th percentile, and the Median Q_3 is the third quartile and the 75th percentile Q_4 is the fourth quartile and the 100th percentile
P_{30} means 30% of the data points are less than the value at the 30th percentile and 70% of the data points are greater than or equal to the value.	5-Number Summary Min Q_1 M Q_3 Max
P_{50} is the Median	

3.3.1 Example Problem: Measures of Location by Position

	Data	Percentile	Data	Percentile	Data	Percentile	Data	Percentile
	100	100%	500	100%	700	100%	100	100%
	90	90%	90	90%	92	90%	90	90%
	80	80%	80	80%	85	80%	80	80%
	70	70%	70	70%	79	70%	70	70%
	60	60%	60	60%	59	60%	60	60%
MEDIAN	50	50%	50	50%	50	50%	50	50%
	40	40%	40	40%	43	40%	40	40%
	30	30%	30	30%	39	30%	30	30%
	20	20%	20	20%	15	20%	20	20%
	10	10%	10	10%	7	10%	10	10%
	0	0%	0	0%	0	0%	-500	0%
AVERAGE	50.00		86.36		106.27		4.55	

3.4 Practice Problems: Measures of Central Location and Location by Position

The exercises are only in the Excel file:
Measures of Location Practice Problems.xlsx

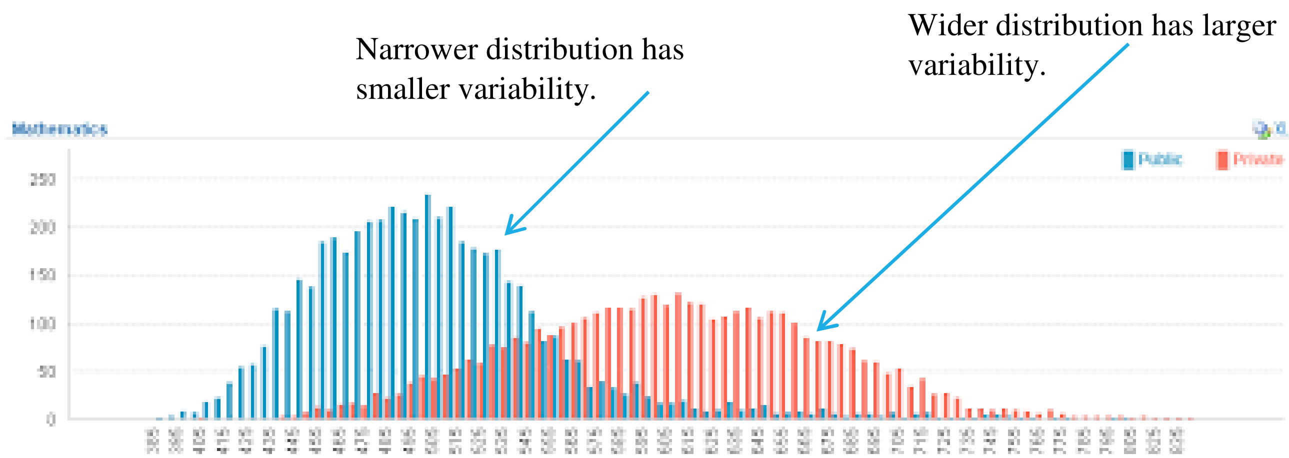
Answers to the exercises are in the Excel file:
Measures of Location Practice Problems KEY.xlsx

3.5 Measures of Variability (Spread)

A measure of variability tells how “close together” or “far apart” values are in a dataset. How different are the values in the dataset?

The farther apart the values are…the wider the distribution and the larger the variability.

The closer together the values are…the narrower or more compact the distribution and the smaller the variability.



There are two measures of variability:

1. The variance and standard deviation are associated with the mean
2. The 5-Number Summary is associated with the median

3.5.1 Variance and Standard Deviation

Variance is the average distance data points are from the mean. Variance is calculated as the average of the squared deviations from the mean. The standard deviation is the square root of the variance. The standard deviation is the most widely used measure of variability.

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

Steps to Calculate the Variance

1. Compute the mean, $\bar{x}$.
2. Subtract the mean from each point, $x_i - \bar{x}$, $(i = 1\ to\ n)$.
3. Square each difference (deviation), $(x_i - \bar{x})^2$.*
4. Sum the squared deviations, $\sum_{i=1}^{n}(x_i - \bar{x})^2$.
5. Divide by (n-1) for sample data or N for population, $\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$.**

Steps to Calculate the Standard Deviation

1. Calculate the variance.
2. Take the positive square root of the variance.***

*For step 3 in calculating the variance, why do we square the deviations?

> The sum of the deviations, without squaring in step 4, will always be equal to zero for all datasets.

**For step 5 in calculating the variance, why do we divide by (n-1) for the sample variance?

> When using sample data, the variance underestimates the population by a predictable amount. The (n-1) is a correction to account for this.

***Why do we take the square root of the variance if variance is the average distance of the points from the mean?

> Variance is in units squared. Taking the positive square root of the variance provides a measure of variability that is in the same units as the data.

3.5.2 Properties of the Standard Deviation

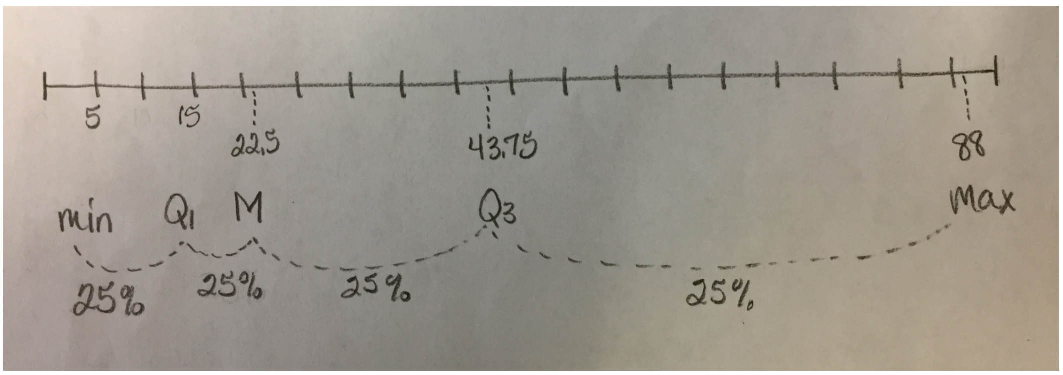
- S measures spread about the mean
- S is always greater than or equal to zero
- S has the same units of measurement as the original observations
- Like the mean, s is sensitive to outliers. A few extreme values in either direction can make s large

3.5.3 5-Number Summary

The 5-number summary consists of the following 5 numbers:

Min Q_1 M Q_3 Max

1. Min → Minimum value
2. Q_1 → First quartile
3. M → Median
4. Q_3 → Third quartile
5. Max → Maximum value



Other important measures related to the median and 5-number summary include the range and interquartile range.

3.5.4 Range

The range is the difference between the maximum and minimum values in the dataset.

Range = Max – Min

3.5.5 Interquartile Range (IQR)

The interquartile range is the difference between the 3rd and 1st quartiles.

Interquartile Range = $Q_3 - Q_1$

3.5.6 Example Problem: Measures of Variability (Spread)

Content examples are in the Excel file:
Measures of Variability Practice Problems KEY.xlsx

Use the following two datasets to calculate measures of center and spread.

Travel Times for Workers in Minutes

North Carolina
5
10
10
10
10
12
15
20
20
25
30
30
40
40
60

New York	
5	25
10	30
10	30
15	40
15	40
15	45
15	60
20	60
20	65
20	88

	North Carolina	New York
MEAN	22.47	31.40
VARIANCE	231.98	496.04
STANDARD DEVIATION	15.23	22.27
MIN	5	5
Q1	10	15
M	20	22.5
Q3	30	43.75
MAX	60	88

3.6 Practice Problems: Measures of Variability (Spread)

The exercises are only in the Excel file:
Measures of Variability Practice Problems.xlsx

Answers to the exercises are in the Excel file:
Measures of Variability Practice Problems KEY.xlsx

3.7 Outliers

An **outlier** is a data point that is extremely different from the other data points in the dataset. The interquartile range (IQR) is used for identifying outliers.

General Rule: A data point is an outlier if it falls more than (1.5 × IQR) above the third quartile (Q_3) or less than (1.5 × IQR) below the first quartile (Q_1).

Lower boundary = $Q_1 - 1.5(IQR)$

Upper boundary = $Q_3 + 1.5(IQR)$

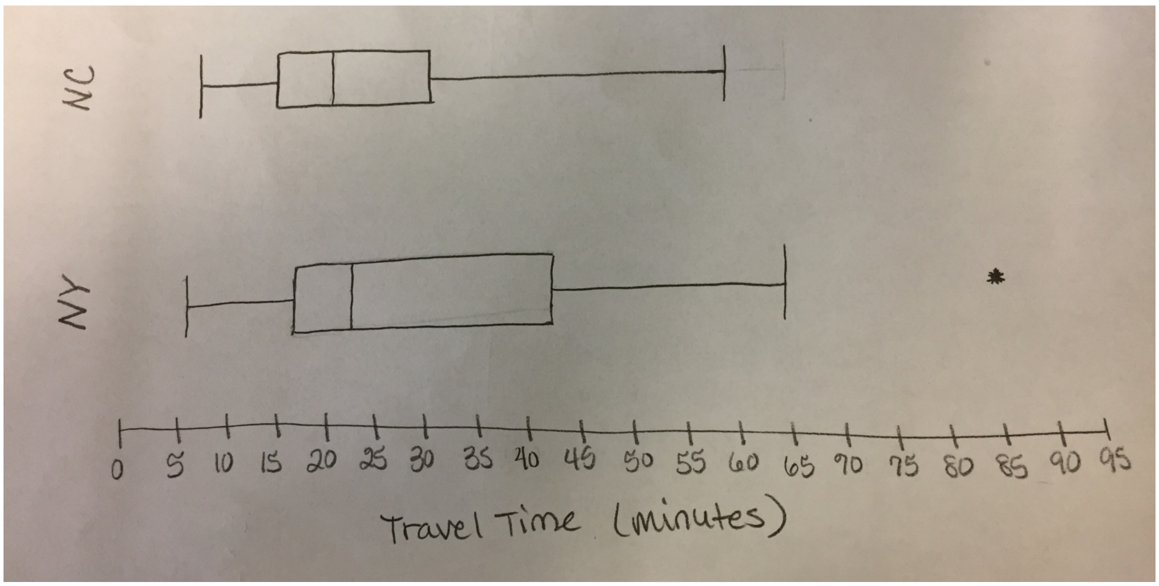
North Carolina Travel Times IQR = 20	**New York Travel Times** IQR = 28.75
1.5(IQR) = 1.5(20) = 30	**1.5(IQR) = 1.5(28.75) = 43.125**
$Q_1 - 1.5(IQR)$ 10 – 30 = -20	$Q_1 - 1.5(IQR)$ 15 – 43.125 = -28.13
$Q_3 + 1.5(IQR)$ 30 + 30 = 60	$Q_3 + 1.5(IQR)$ 43.75 + 43.125 = 86.9
Any points less than **-20** or greater than **60** are suspected outliers.	Any points less than **-28.13** or greater than **86.9** are suspected outliers.

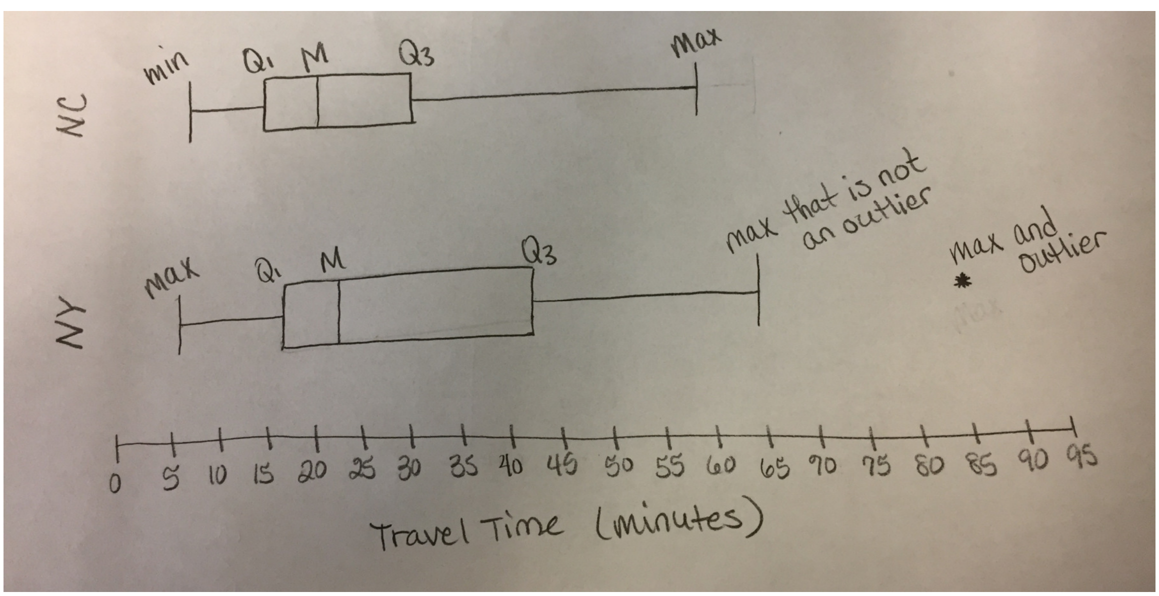
Make sure your results make sense. We would not report the lower values (negatives) because they are impossible. We can only have high outliers with this dataset.

3.8 Box Plots (Box and Whisker)

- A graphical representation of the 5-number summary
- Especially useful when the data are **NOT** approximately bell-shaped
- Also useful in visualizing outliers

Draw two box plots side by side to illustrate and compare the distribution of travel times for each state.





3.9 Choosing Measures of Center and Spread

You have a choice between two descriptions of the center and spread of a distribution:

- Mean and Standard Deviation
- Median and Five-Number Summary

Mean and Standard Deviation

Use mean and standard deviation only for reasonably symmetric distributions that don't have outliers.

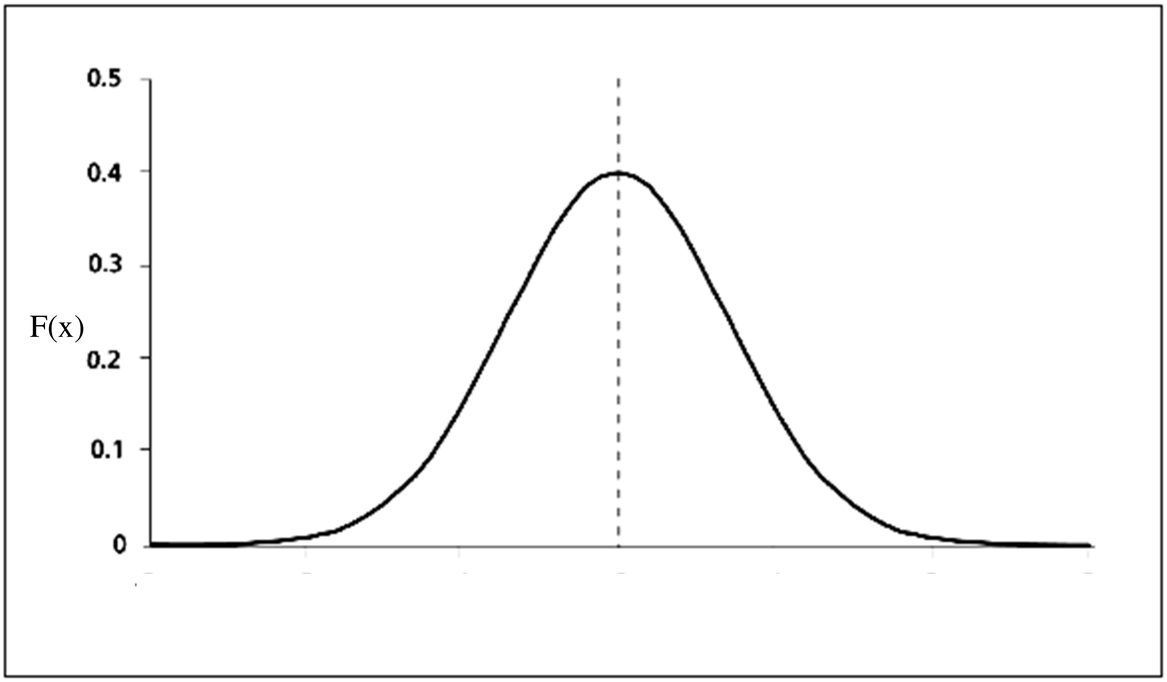
Median and 5-Number Summary

The median and Five-Number Summary are usually better than the mean and standard deviation for describing a skewed distribution or a distribution with outliers.

NOTE: Numerical summaries do not fully describe the shape of a distribution. *ALWAYS GRAPH YOUR DATA!*

3.10 Measures of Relative Location (Relative to the Mean)

3.10.1 The Normal Distribution



Properties

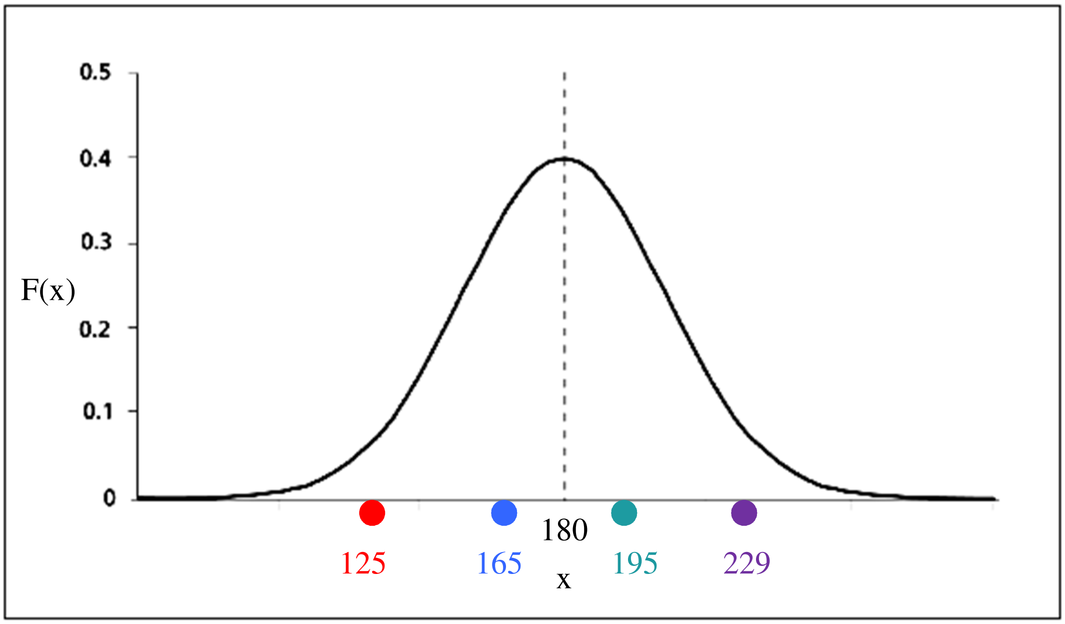
- All Normal curves are symmetric, single-peaked, and bell-shaped.
- The x-values can take on positive and/or negative values (depends on the data context).
- The tails approach the horizontal axis but never touch or cross it (asymptotic).
- The y-axis is the frequency or relative frequency, which are never zero or negative.
- The entire '*family of Normal distributions*' is differentiated by two parameters: the mean, **μ** and the standard deviation, $\boldsymbol{\sigma}$ or the mean and the variance, $\boldsymbol{\sigma^2}$.
- Mathematical notation: X ~ N(μ, σ) or X ~ N(μ, σ^2).

3.10.2 Standard Normal Scores (Z-scores)

The **Z-score** for a value in a dataset is the number of standard deviations the observation is from the mean of the distribution. In other words, the Z-score tells you how far a value is from the mean in terms of standard deviations. Z-scores **standardize** approximately bell-shaped distributions of variables so the distributions can be compared even when the data values are measured on different scales. $Z = \frac{x - \mu}{\sigma}$

Consider the data:

x	125	139	165	187	170	195	229	230
z	-1.44	-1.07	-0.39	0.18	-0.26	0.39	1.28	1.31

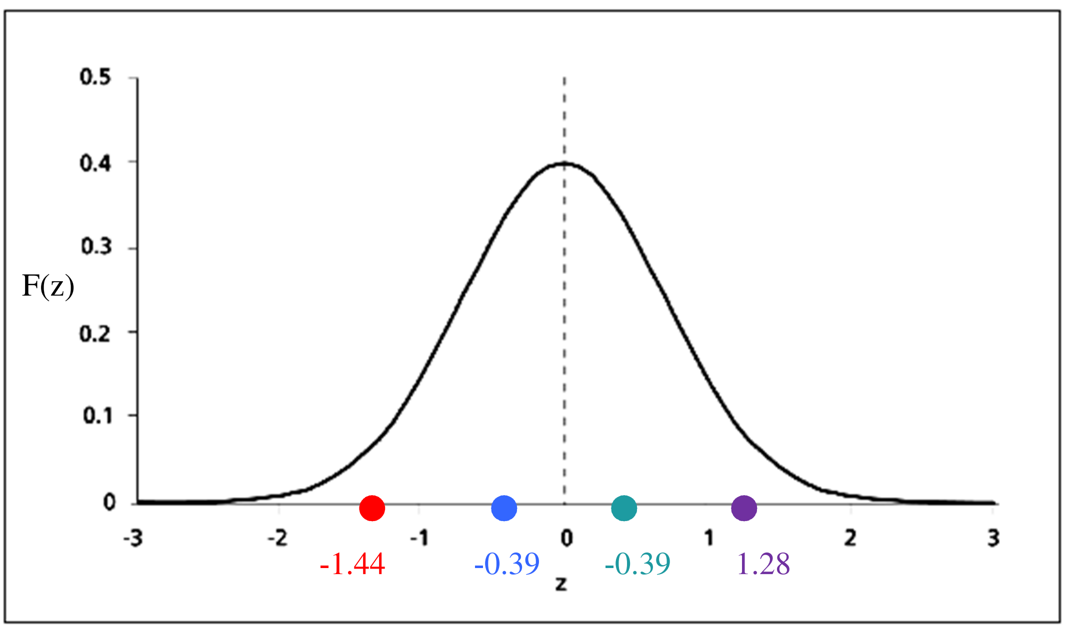


$$z = \frac{125 - 180}{38.18} = -1.44$$

$$z = \frac{165 - 180}{38.18} = -0.39$$

$$z = \frac{195 - 180}{38.18} = 0.39$$

$$z = \frac{229 - 180}{38.18} = 1.28$$



Content examples are in the Excel file:
Z-Scores.xlsx

3.10.3 Example Problem: Calculate and Compare Z-Scores

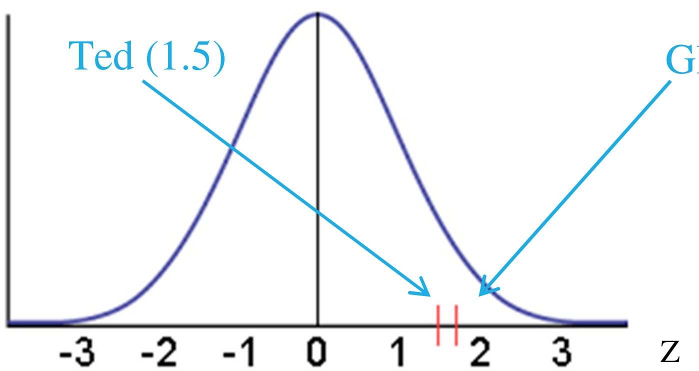
Gloria scores 680 on the verbal SAT. Ted scored 27 on the verbal ACT. Both scored above the national average, but who scored higher on their respective scales?

The distribution of SAT verbal scores is normal, with mean equal to 500 and standard deviation equal to 100. The distribution of ACT verbal scores is normal, with mean equal to 18 and standard deviation equal to 6.

$$Z = \frac{x - \mu}{\sigma} \qquad Z_{Gloria} = \frac{680-500}{100} = 1.8 \qquad Z_{Ted} = \frac{27-18}{6} = \frac{9}{6} = 1.5$$

Gloria scored higher!

By calculating each student's Z-score, we take away the units of measurement and put each score relative to its distance from the mean. **The normal curve, "bell-shaped" distribution:**



3.10.4 Use Z-Scores to Determine if Outliers Exist

There are two methods for determining if outliers exist:

1. 5-Number Summary and the 1.5xIQR rule
2. Z-Scores and the 3 standard deviations rule

Values more than 3 standard deviations from the mean are considered outliers. That is, if the z-score of a value, x, is greater than 3 or less than -3, ($z < -3$ or $z > 3$), the value is an outlier.

3.11 Practice Problems: Z-Scores

Exercises 1 and 2 are in the Excel file:
Z-Scores.xlsx

Answers to exercises 1 and 2 are in the Excel file:
Z-Scores KEY.xlsx

3.11.1 Exercise 1: Determine Outliers

Consider a sample with 10 observations of 0, –5, 12, 4, –4, –2, –4, 5, 12, and 4. Use *z*-scores to determine if there are any outliers in the data; assume a bell-shaped distribution.

3.11.2 Exercise 2: Compare Batting Averages

Compare batting averages. Three landmarks of baseball achievement are Ty Cobb's batting average of .420 in 1911, Ted William's .406 in 1941 and George Brett's .390 in 1980. These batting averages cannot be compared directly because the distribution of major league batting averages has changed over the years. The distributions follow a Normal curve, except for outliers such as Cobb, Williams and Brett. While the men batting average has been held roughly constant by rule changes and the balance between hitting and pitching, the standard deviation has dropped over time. Here are the facts:

Decade	Mean	Standard Deviation	Z-score
1910s	0.266	0.0371	
1940s	0.267	0.0326	
1970s	0.261	0.0317	

Compute the standardized batting averages (z-scores) for Cobb, Williams, and Brett to compare how far each stood above his peers.

3.11.3 Exercise 3: Class Size

The average class size this semester in the business school of a university is 38.1 students with a standard deviation of 12.9 students. The z-score for a class with 21 students is _____.

3.11.4 Exercise 4: ACT Score

Scores on the ACT college entrance exam follow a Normal distribution with a mean 18 and a standard deviation 6. Wayne's standardized score (z-score) on the ACT was -0.7. What was Wayne's actual ACT score?

3.11.5 Exercise 5: Albert Einstein

IQ scores have a mean of 100 and a standard deviation of 16. Albert Einstein reportedly had an IQ of 160.

a. What is the difference between Einstein's IQ and the mean?

b. Convert Einstein's IQ score to a z score.

c. If we consider "usual IQ scores to be those that convert z scores between -2 and 2, is Einstein's IQ usual or unusual?

3.11.6 Exercise 6: Chocolate Bars

The weight of chocolate bars from a chocolate factory have a mean of 8 ounces with standard deviation of 0.1 ounce. What is the *z*-score corresponding to a weight of 8.17 ounces?

If a chocolate bar has a z-score of 1.5, what is the corresponding weight of that bar?

3.11.7 Exercise 7: Library Books

Books in the library are found to have average length of 350 pages with standard deviation of 100 pages. What is the *z*-score corresponding to a book of length 80 pages?

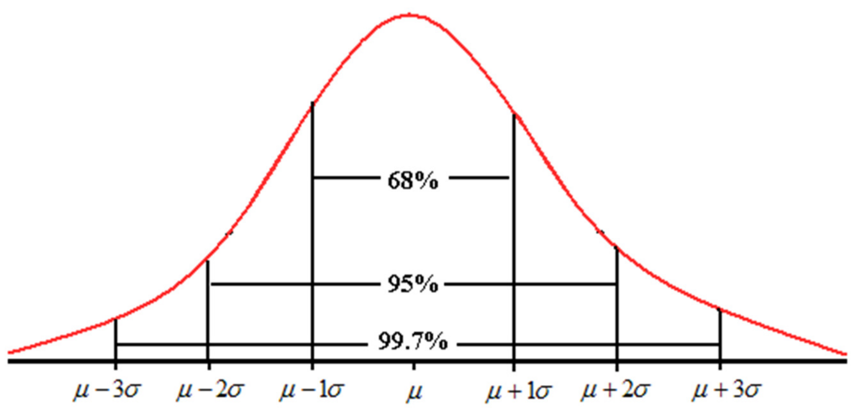
Your English teacher tells you to read a book that has a minimum number of pages, and the minimum page count corresponds to a z-score of 1.5. What is the minimum page count?

1. No outliers
   lowest z=-1.20
   highest z=1.64
2. Ty 4.15; Ted 4.26; George 4.07
3. -1.33
4. 13.8
5. 60, 3.75, unusual
6. 1.7, 8.15oz
7. -2.7, 500 pages

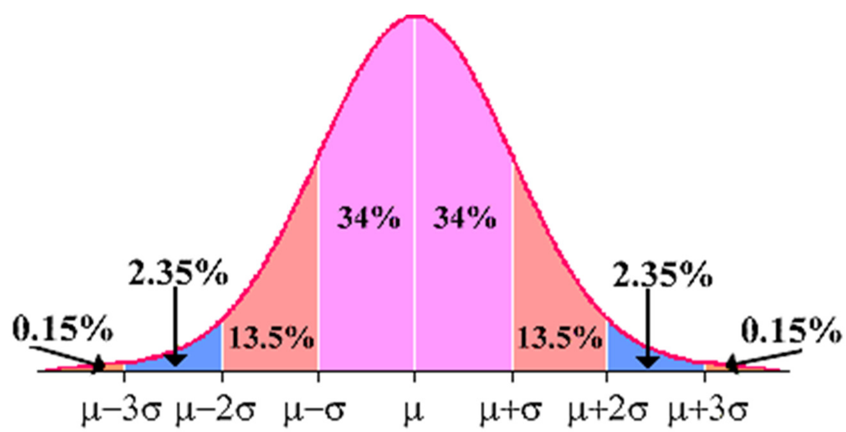
3.12 Empirical Rule

In a Normal distribution with mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$…

- Approximately **68%** of the observations fall within one standard deviation of the mean, i.e., within σ of μ.
- Approximately **95%** of the observations fall within two standard deviations of the mean, i.e., within 2σ of μ.
- Approximately **99.7%** of the observations fall within three standard deviations of the mean, i.e., within 3σ of μ.



[4]



- The above sections and their respective percentages are true for any normally distributed dataset.
- The sections can be added together or subtracted from one another to solve problems.
- The empirical rule allows you to find percentages and percentile ranks for data values and groups of data values that are within a specified number of standard deviations from the mean.

[4] http://www.stat119review.com/more-material/normal-distribution/empirical-rule (both) 9/10/17

3.13 Practice Problems: Empirical Rule

3.13.1 Exercise 1: Accounting Class Scores

In an accounting class of 200 students, the mean and standard deviation of scores was 73 and 6, respectively. Use the empirical rule to determine the percentage of students who scored between 67 and 79.

3.13.2 Exercise 2: Professors' Average Salary

Professors at a local university earn an average salary of $90,000 with a standard deviation of $4,000. The salary distribution is approximately bell-shaped.

a. What can be said about the percentage of salaries that are less than $82,000 or more than $98,000?

b. What can be said about the percentage of salaries that are at least $86,000?

c. Because of budget limitations, it has been decided that only those whose salaries are approximately in the bottom 2.5% would get a raise. What is the maximum current salary that qualifies for the raise?

3.13.3 Exercise 3: Immigration Wait Time

Suppose the wait to pass through immigration at JFK Airport in New York is thought to be bell-shaped and symmetrical with a mean of 21 minutes. It is known that 68% of travelers will spend between 14 and 28 minutes waiting to pass through immigration. The standard deviation for the wait time through immigration is _________.

3.13.4 Exercise 4: Fluffy Kittens

The weights of adorable, fluffy kittens are normally distributed with a mean of 3.6 pounds and a standard deviation of 0.4 pounds. Answer the following questions, using the Empirical Rule.

a. What percent of adorable, fluffy kittens weigh between 2.8 and 4.8 pounds?

b. What percent of adorable, fluffy kittens weigh less than 2.4 pounds?

c. What value corresponds to a 97.5th percentile of kitten weights?

3.13.5 Example 5: Heights of Men

The distribution of heights of adult American men is approximately normally distributed with mean 69 inches and standard deviation 2.5 inches. Use the Empirical Rule to answer the following questions.

a. What percent of men are taller than 74 inches?

b. Between what heights do the middle 95% of men fall?

c. What percent of men are between 64 and 66.5 inches tall?

d. A height of 71.5 inches corresponds to what percentile of adult male American heights?

3.13.6 Exercise 6: Gas Mileage

A new line of cars has gas mileage that is normally distributed with a mean of 32 mpg with a standard deviation of 4 mpg. Use the Empirical Rule to answer questions below.

a. The middle 68% of cars gets between how many mpg?

b. The middle 95% of cars gets between how many mpg?

c. 2.5% of all cars get no more than how many mpg?

d. 0.15% of all cars get no more than how many mpg?

e. Only 16% of all cars get more than how many mpg?

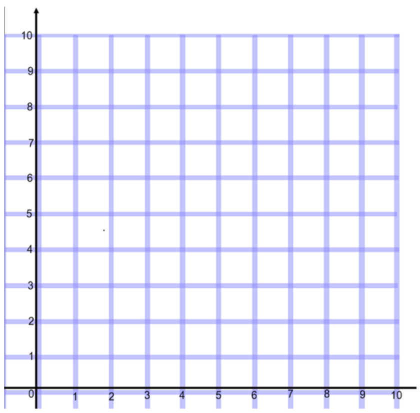
f. Only 0.15% of all cars get more than how many mpg?

1. 68%
2. 5%, 84%, 82,000
3. 7
4. 97.35%, 0.15%, 4.4 lbs
5. 2.5%, 64" & 74", 13.5%, 84th percentile
6. 28 mpg & 36 mpg, 24 mpg & 40 mpg, 24 mpg, 20 mpg, 36 mpg, 44 mpg

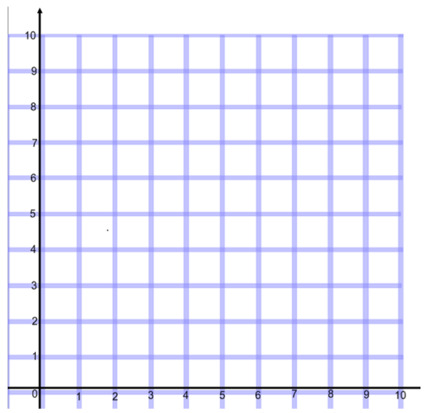
[This page intentionally left blank]

3.14 Covariance and Correlation

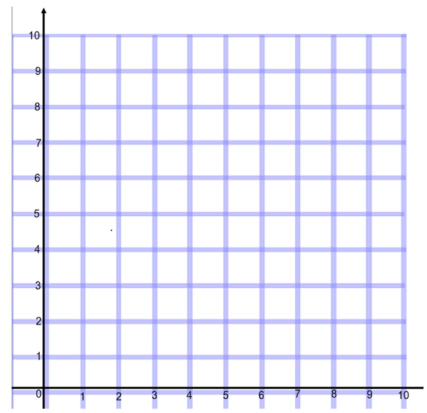
Covariance and correlation are numeric measures that describe the relationship between two quantitative variables. In general, there are three possible relationships between two quantitative variables: no relationship, linear relationship, or nonlinear relationship.



No relationship



Linear relationship



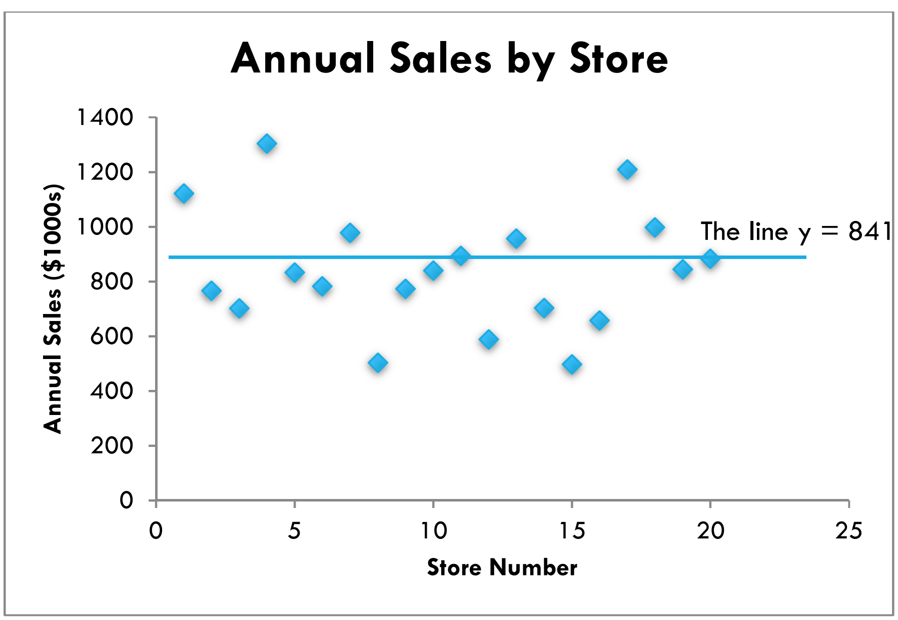
Nonlinear relationship

Create a Scatterplot in Excel ***Note:*** The variable on the left column is always plotted on the x-axis.

3.14.1 Covariance

The 7-11 example is in the Excel file:
Scatterplot and Correlation with 7-11 Stores.xlsx

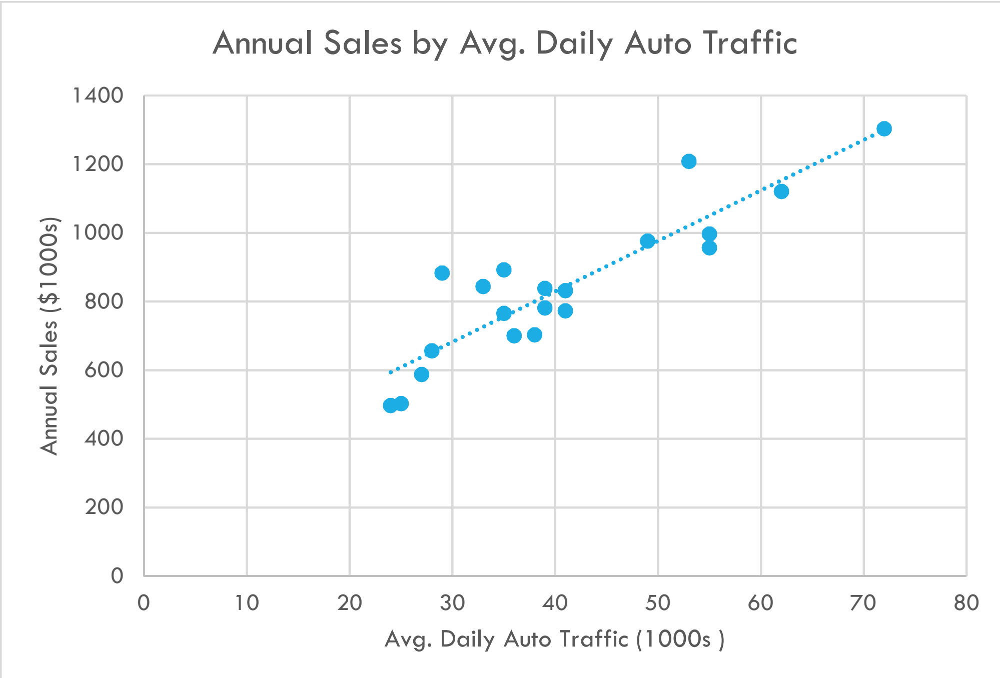
Recall that the variance of a variable y, denoted s_y^2, is a numerical measure that describes the average scatter of points about the mean of y.



The mean of y is 841.3 so the line y=841.3 represents the "center" of the scattered points.

The store number is just an ID, a way to separate the points. This shows visually where each point is with respect to the mean of y.

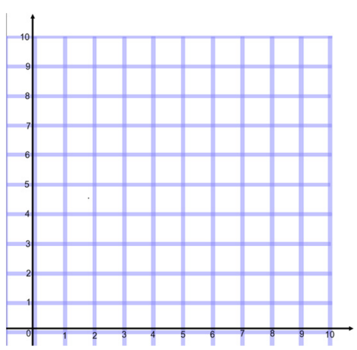
The covariance, denoted s_{xy}, is a numerical measure that describes the average scatter of points with respect to another variable when plotted together as coordinate pairs in a two-dimensional plane. **The covariance applies only to variables with a linear relationship.**



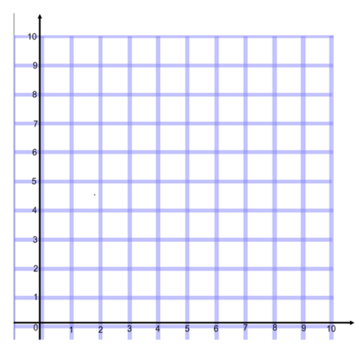
A line similar to the line y = 841.3 in that it represents the “center” of the scattered points.

The deviation of the x values around the mean of x and the deviation of the y values around the mean of y are taken together to compute the covariance. *See Excel calculations.*

The covariance can be positive or negative in sign, and this sign reveals the direction of the linear relationship between the two variables: positive or negative.



Positive



Negative

Note: Unlike the standard deviation (square root of the variance), which measures the average scatter of points in the same units as the data, we do not take the square root of the covariance because it can be negative.

Note: The units of the covariance don’t make much sense because the unit of measure for x is typically very different than the unit of measure for y.

3.14.2 Correlation

The correlation, *r,* is a numerical summary measure that describes the strength and direction of a *linear* relationship between two *quantitative* variables:

The correlation is a 'standardized' covariance: To calculate the correlation, divide the covariance by the standard deviation of x and the standard deviation of y, thus canceling the units.

$$r = \frac{covariance}{s_x s_y}$$

3.14.3 Facts About Correlation

- **The correlation applies only to variables with a linear relationship.**
- The correlation, *r,* is always a number between -1 and 1. The strength of the linear relationship increases as *r* moves away from 0 toward -1 or 1. Zero indicates no relationship.

$$-1 \leq r \leq +1$$

- Correlation makes no distinction between the two variables. The statements, "x is related to y" and "y is related to x," are the same because there is only one correlation between the two variables.
- The correlation has no units so the value of r does not change if the units of measurement of x and/or y change.
- Correlation is affected by outliers. Any calculation that involves the actual data points is influenced by extreme values.
- An association/correlation does not imply causation (lurking variables); any inference about the cause of an association must be justified by a reasonable theoretical relationship.
  E.g. number of ice cream cones sold is correlated with the number of drowning victims

3.14.4 Interpret the Correlation

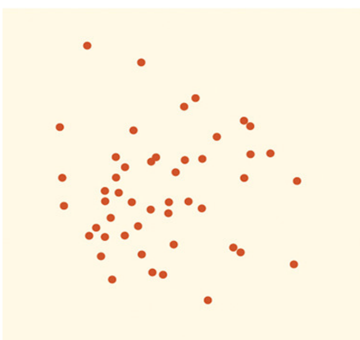
- Values of *r* near 0 indicate a very weak or no relationship. Knowing the value of one variable has no effect on knowing the value of the other. (1)

- $r > 0$ indicates a positive association; as one variable **increases** in value, the other variable also **increases** in value. (2), (3)

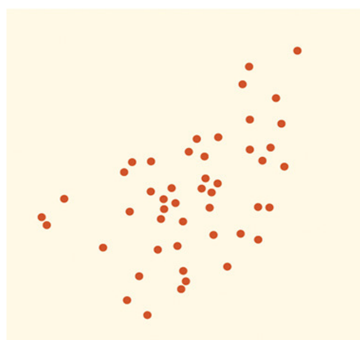
- $r = +1$ is a perfect one-to-one positive relationship

- $r < 0$ indicates a negative association; as one variable **increases** in value, the other variable **decreases** in value. (4), (5), (6)

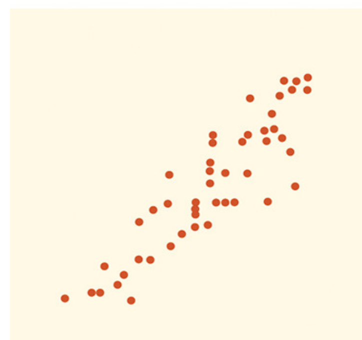
- $r = -1$ is a perfect one-to-one negative relationship



(1) r = 0 (no correlation)

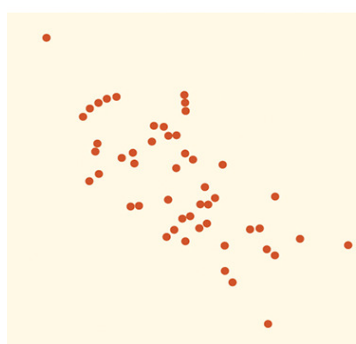


(2) r = 0.5 (moderate)

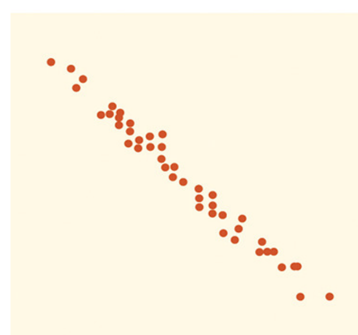


(3) r = 0.9 (strong)

(4) r = -0.3
(weak)

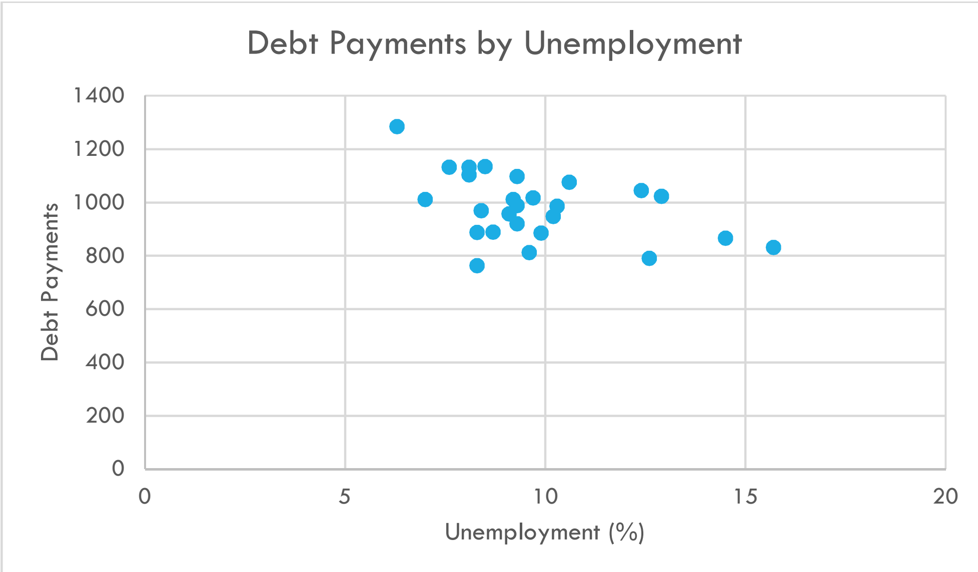


(5) r = -0.7
(moderate toward strong)



(6) r = -0.99
(very strong)

3.14.5 Example Problem: Covariance and Correlation



Covariance	-126.367
Unemployment Mean	9.77%
Unemployment Standard Deviation	2.23%
Debt Payments Mean	983.46 ($)
Debt Payments Standard Deviation	124.61 ($)

Use the information in the table above to calculate the correlation.

$$r = \frac{-126.367}{(2.23)(124.61)} = -0.45$$

3.15 Practice Problems: Covariance and Correlation

The following exercises are also in the Excel file:
Excel File: Covar and Correlation Practice Problems.xlsx

3.15.1 Exercise 1: Stocks A and B

The covariance between the returns of stock A and stock B is –0.114.The standard deviation of the rates of return is 0.23 for stock A and 0.84 for stock B. The correlation of the rates of return between A and B is the closest to _____.

3.15.2 Exercise 2: More Stocks A and B

The covariance between the returns of stock A and stock B is –135. The standard deviation of the rates of return is 18 for stock A and 12 for stock B. The correlation coefficient of the rates of return between A and B is closest to ______.

3.15.3 Exercise 3: GRE and GPA

The director of graduate admissions is analyzing the relationship between scores in the Graduate Record Examination(GRE) and student performance in graduate school, as measured by a student's GPA. The table below shows a sample of 10 students.

a. Use Excel to calculate the covariance.
b. Use Excel to calculate the correlation.
c. Describe the direction and strength of the relationship between GRE and GPA scores.

GRE	GPA
1,550	3.4
1,500	3.6
1,200	3.3
1,050	2.9
1,100	3.1
1,250	3.3
900	3
850	2.8
950	3.2
1,350	3.3

1. -0.59
2. -0.625
3. 51.33, 0.868, strong and positive

[This page intentionally left blank]

4 Introduction to Probability

4.1 Objectives

- Describe fundamental probability concepts
- Explain the two defining properties of probability
- Assign probability to events
- Apply the addition and complement rules to calculate probabilities
- Show whether events are independent or dependent by comparing conditional and marginal probability
- Construct joint and conditional probability tables
- Apply the multiplication rule to find joint probabilities (independent events)
- Apply the multiplication rule to find joint probabilities (dependent events)
- Show whether events are independent or dependent by comparing conditional and marginal probability
- Construct joint and conditional probability tables
- Apply the multiplication rule to find joint probabilities (independent events)
- Apply the multiplication rule to find joint probabilities (dependent events)

4.2 Experiment

Outcome – An outcome is a possible result of an experiment (also called a response).

Trial – A trial is a process that leads to one of several possible outcomes (e.g., rolling a die is the trial, the side that is facing up is the outcome).

Sample Space – The sample space is all possible outcomes for the trial.

Student Activity
For each of 10 die rolls, i.e., each **trial**…

a. Student 1 predicts the face of the die, i.e., the **outcome**
b. Student 2 rolls the die
c. Student 3 records the **outcome**

*What are the possible outcomes of each trial, i.e., what is the **sample space**?

S = { 1, 2, 3, 4, 5, 6 }

	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6	Trial 7	Trial 8	Trial 9	Trial 10
Prediction										
Outcome										

4.3 Randomness

Random behavior is unpredictable in the short run (e.g., one trial of rolling a die), but has a regular and predictable pattern in the long run (e.g., one-thousand trials of rolling a die).

Truly random phenomena are

- *equally likely* - no outcome has a higher likelihood of occurring than any other
- *independent* - the outcome of one trial does not affect the outcome of any other

https://www.geogebra.org/m/UsoH4eNl

The pattern that results from rolling one die thousands of times is uniform.

The pattern that results from rolling two dice thousands of times (outcome is the sum of the facing dots) is normal.

We can use the regular and predictable pattern of the long-run behavior to assign a **probability** to an **event** of interest.

4.4 Probability

Probability – Probability is the chance/likelihood of some event occurring; it is a measure of uncertainty.

Event – An event is a specific outcome or combination of outcomes that can occur during a trial of a random phenomenon.

4.5 Three Types of Probability

Subjective or Personal Probability	Empirical Probability	Classical Probability
A ***subjective probability*** is calculated by drawing on personal and subjective judgment. What is the probability that your friend will be within 5 minutes of your meeting time?	An ***empirical probability*** is calculated as a relative frequency of occurrence. Repeated trials are required to establish the empirical probability of an event. ***Recall:*** The relative frequency is how often an outcome occurs divided by all outcomes. Define event A = roll a 5 $P(A) = \frac{number\ of\ times\ an\ event\ occurs}{total\ number\ of\ trials} = \_\_\_$ If you roll a die 10,000 times and count 1650 5s, then the empirical probability is $P(A) = \frac{1650}{10{,}000} = 0.165$	A ***classical probability*** is based on logical analysis rather than on observation or personal judgment. For a fair die, each side is equally likely to be the outcome and there are 6 sides. Define event A = roll a 5 *Note:* Six equally likely sides so probability of rolling a 5 is 1 in 6. $P(A) = \frac{number\ of\ favorable\ outcomes}{total\ number\ of\ possible\ outcomes} = \frac{\mathbf{1}\ side\ is\ a\ 5}{\mathbf{6}\ sides\ on\ the\ die} = \frac{1}{6}$

4.6 Law of Large Numbers (LLN)

If events are independent, then as the number of trials increases, the long-run relative frequency of any outcome gets closer and closer to a single value, i.e., the classical probability.

4.7 Two Defining Properties of Probability

Sample Space	Probability
1	1/6 = 0.167
2	1/6 = 0.167
3	1/6 = 0.167
4	1/6 = 0.167
5	1/6 = 0.167
6	1/6 = 0.167
Total	**1**

1) Any probability is a value between zero and one, inclusive, i.e., $0 \leq p \leq 1$.

2) The sum of the probabilities of all outcomes in a given sample space is 1.

4.8 Probability as Relative Frequency

When you have frequency and total counts, finding probabilities is exactly the same as finding relative frequencies.

As an experiment, students counted plain M&Ms and recorded the frequency of each color. From 87 fun size packages of plain M&Ms, students counted 1255 total M&Ms. The frequency table follows:

Sample Space

S = {Brown, Yellow, Red, Orange, Green, Blue}

Plain M&Ms Color	Frequency	Relative Frequency
Brown	112	0.09
Yellow	105	0.08
Red	109	0.09
Orange	327	0.26
Green	314	0.25
Blue	288	0.23
Total	**1255**	**1.00**

1. What is the proportion of yellow M&Ms? [**0.08**]
2. If you selected one M&M from a jar containing all 1255, what is the probability you would select a yellow M&M? [**Define event A = draw yellow: P(A) = 0.08**]
3. What is the probability you would select a yellow **or** blue M&M? [**Define event A = draw yellow & event B = draw blue: P(A or B) = 0.08 + 0.23 = 0.31**]
4. What is the probability you would select a green **and** orange M&M? The probability is **zero** because these are **mutually exclusive**. You cannot have an M&M that is both yellow and orange. [**Define event A = draw yellow & event B = draw orange: P(A and B) = 0**]
5. What is the probability you would select any color except red? [**Define event A = draw red: P(A') = 1 – 0.09 = 0.91**]

4.9 Finding Probabilities from a Contingency Table

Students also counted peanut M&Ms and recorded the frequency of each color. From 41 fun size packages of peanut M&Ms, students counted 322 total M&Ms. The frequencies appear in the following table:

Note: Most questions in the following example set involves drawing one M&M and using the frequencies in the contingency table below to calculate the probabilities.

Sample Space

S = {Plain Brown, Peanut Brown, Plain Yellow, Peanut Yellow, Plain Red, Peanut red…}

M&M Color	Frequency of Plain	Frequency of Peanut	Total
Brown	112	36	148
Yellow	105	63	168
Red	109	40	149
Orange	327	60	387
Green	314	57	371
Blue	288	66	354
Total	**1255**	**322**	**1577**

Joint Frequencies – The total number of elements that fit into two or more categories, e.g., M&Ms that are both green and peanut

Joint Probability – The probability of two or more events occurring together in the same trial, e.g., P(green and peanut)

Marginal Frequencies – The total row or column, only one category of one variable, e.g., all red M&Ms regardless of Type or all peanut M&Ms regardless of color

Marginal Probability – The probability of one category of one variable occurring, e.g., all red M&Ms, includes both plain and peanut

4.10 Example Problem: Essential Rules and Definitions of Probability

<table>
<tr><th colspan="2">Essential Rules and Definitions of Probability</th></tr>
<tr><td>Any probability is a number between 0 and 1</td><td>All possible outcomes together must have probability 1</td></tr>
<tr><td>$0 \leq P(A) \leq 1$</td><td>$P(S) = 1$, where S = sample space</td></tr>
<tr><td>Mutually Exclusive Events</td><td>Complement Rule</td></tr>
<tr><td>Question
What is the probability of drawing a blue and green M&M?

Define events A and B as follows…
A = draw a blue M&M
B = draw a green M&M

Procedure
Recognize that you cannot draw one M&M that is both blue and green. Colors of M&Ms are mutually exclusive, which means the events A and B do not share any common outcome.

Answer
$\text{P(A and B)} = 0$</td><td>Question
What is the probability of drawing any color except red?

Define event A as follows…
A = draw a red M&M

Procedure
Find the probability of drawing a red M&M and subtract that probability from 1.

Formula
$P(A') = 1 - P(A)$

Answer
$P(A') = 1 - 0.094 = 0.906$</td></tr>
</table>

4.11 Example Problem: Marginal Probabilities with Single Events

<table>
<tr><th colspan="2">Calculate Marginal Probabilities (Total Row or Column)</th></tr>
<tr>
<td>

Question

What is the probability of drawing a blue M&M?

Define event A as follows…

A = draw a blue M&M

Procedure

Divide the number of blue M&Ms by the total number of M&Ms.

Formula

P(A) = $\frac{Number\ of\ Blue\ M\&Ms}{Total\ number\ of\ M\&Ms}$

Answer

P(A) = $\frac{354}{1577}$ = 0.22

</td>
<td>

Question

What is the probability of drawing a peanut M&M?

Define event B as follows…

B = draw a peanut M&M

Procedure

Divide the number of peanut M&Ms by the total number of M&Ms.

Formula

P(B) = $\frac{Number\ of\ Peanut\ M\&Ms}{Total\ number\ of\ M\&Ms}$

Answer

P(B) = $\frac{322}{1577}$ = 0.20

</td>
</tr>
</table>

4.12 Example Problem: The Probability of One Event **OR** Another

Calculate Probabilities from a Contingency Table and with the Addition Rule	
Non-Overlapping Events	**Overlapping Events**
Question What is the probability of drawing a blue or a green M&M? **Define events** A and B as follows… A = draw a blue M&M B = draw a green M&M **Procedure** Add the probability of drawing a blue M&M to the probability of drawing a green M&M. **Formula** P(A or B) = P(A) + P(B) **Answer** $P(A\ or\ B) = \frac{354}{1577} + \frac{371}{1577} = 0.46$	**Question** What is the probability of drawing a blue or a peanut M&M? **Define events** A and B as follows… A = draw a blue M&M B = draw a peanut M&M **Procedure** Add the probability of drawing a blue M&M to the probability of drawing a green M&M and subtract the overlapping amount because it was counted twice, i.e., subtract the joint probability of drawing an M&M that is both blue and peanut. **Formula** P(A or B) = P(A) + P(B) – P(A and B) **Answer** $P(A\ or\ B) = \frac{354}{1577} + \frac{322}{1577} - \frac{66}{1577} = 0.37$

4.13 Example Problem: The Probability of One Event **AND** Another

Calculate Joint Probabilities from a Contingency Table	
Question What is the probability of drawing a blue and peanut M&M? **Define events** A and B as follows… A = draw a blue M&M B = draw a peanut M&M **Procedure** Divide the number of blue peanut M&Ms by the total number of M&Ms. **Formula** P(A and B) = $\frac{Number\ of\ Blue\ Peanut\ M\&Ms}{Total\ number\ of\ M\&Ms}$ **Answer** P(A and B) = $\frac{66}{1577}$ = 0.04	**Question** What is the probability of drawing a plain and red M&M? **Define events** A and B as follows… A = draw a plain M&M B = draw a red M&M **Procedure** Divide the number of red plain M&Ms by the total number of M&Ms. **Formula** P(A and B) = $\frac{Number\ of\ Red\ Plain\ M\&Ms}{Total\ number\ of\ M\&Ms}$ **Answer** P(A and B) = $\frac{109}{1577}$ = 0.07

4.14 Conditional Probability

What is the probability you draw a red M&M?
Define P(A) = draw a red M&M
P(A) = 0.094

If you draw one M&M, does the probability that you drew a red M&M change if you know the M&M is plain? Given you know the M&M is plain, what is the probability you drew a red M&M?

$$P(A \text{ and } B) = \frac{Number\ of\ (A\ and\ B)}{Total\ Number}$$

$$P(A|B) = \frac{Number\ of\ (A\ and\ B)}{Number\ of\ B}$$

If you know the type of M&M you drew, the denominator of your probability calculation changes. You now have a new "whole." The total number of both plain and peanut M&Ms is no longer relevant. You know you drew a plain M&M so the total number of M&Ms to take into consideration is the total number of plain M&Ms.

Define P(A) = draw a red M&M
Define P(B) = draw a plain M&M

$$P(A|B) = \frac{Number\ of\ plain\ red\ M\&Ms}{Total\ number\ of\ plain\ M\&Ms} = \frac{109}{1255} = 0.087$$

$$P(A|B) = \frac{Number\ of\ plain\ red\ M\&Ms}{Total\ number\ of\ plain\ M\&Ms} = \frac{P(A\ and\ B)}{P(B)}$$

Two events are **independent** if knowing that one occurs does not change the probability that the other occurs, i.e., $P(A|B) = P(A)$.

Events are **dependent** if the occurrence of one is related to the probability of the occurrence of the other, i.e., $P(A|B) \neq P(A)$.

$P(A|B) = 0.087$

$P(A) = 0.094$

M&M color and type are **dependent** events.

CONTINGENCY TABLE

	Plain M&M	Peanut M&M	Total
Brown	112	36	148
Yellow	105	63	168
Red	109	40	149
Orange	327	60	387
Green	314	57	371
Blue	288	66	354
Total	1255	322	1577

JOINT

	Plain M&M	Peanut M&M	Total
Brown	0.071	0.023	0.094
Yellow	0.067	0.040	0.107
Red	0.069	0.025	0.094
Orange	0.207	0.038	0.245
Green	0.199	0.036	0.235
Blue	0.183	0.042	0.224
Total	0.796	0.204	1

CONDITIONED ON TYPE – WE KNOW THE TYPE BEFORE CALCULATING THE PROBABILITY

	Plain M&M	Peanut M&M	Total
Brown	0.089	0.112	0.094
Yellow	0.084	0.196	0.107
Red	0.087	0.124	0.094
Orange	0.261	0.186	0.245
Green	0.250	0.177	0.235
Blue	0.229	0.205	0.224
Total	1	1	1

CONDITIONED ON COLOR – WE KNOW THE COLOR BEFORE CALCULATING THE PROBABILITY

	Plain M&M	Peanut M&M	Total
Brown	0.757	0.243	1
Yellow	0.625	0.375	1
Red	0.732	0.268	1
Orange	0.845	0.155	1
Green	0.846	0.154	1
Blue	0.814	0.186	1
Total	0.796	0.204	1

4.15 Example Problem: The Probability of One Event **GIVEN** Another

<table>
<tr><th colspan="2">Calculate Conditional Probabilities from a Contingency Table</th></tr>
<tr>
<td>

Question

What is the probability you drew a blue M&M if you know you drew a peanut M&M?

Define events A and B as follows…

A = draw a blue M&M
B = draw a peanut M&M

Procedure

Recognize that the probability you drew a blue M&M without knowing the type might be different from the probability you drew a blue M&M knowing the type. The denominator now becomes the total number of peanut M&Ms, because you know the M&M is peanut.

Formula

P(A|B) = $\frac{Number\ of\ Blue\ Peanut\ M\&Ms}{Total\ number\ of\ Peanut\ M\&Ms}$

Independent or Dependent?

Events A and B are independent if the following is true:

$\mathrm{P(A|B) = P(A)}$

$\mathrm{P(A)} = \frac{354}{1577} = 0.22 \neq \mathrm{P(A|B)} = \frac{66}{322}$ = 0.204 (Dep.)

</td>
<td>

Question

What is the probability you drew a plain M&M if you know you drew a red M&M?

Define events A and B as follows…

A = draw a red M&M
B = draw a plain M&M

Procedure

Recognize that the probability you drew a plain M&M without knowing the color might be different from the probability you drew a plain M&M knowing the color. The denominator now becomes the total number of red M&Ms, because you know the M&M is red.

Formula

P(B|A) = $\frac{Number\ of\ Red\ Plain\ M\&Ms}{Total\ number\ of\ Red\ M\&Ms}$

Independent or Dependent?

Events A and B are independent if the following is true:

$\mathrm{P(B|A) = P(B)}$

$\mathrm{P(B)} = \frac{1255}{1577} = 0.78 \neq \mathrm{P(B|A)} = \frac{109}{149}$ = 0.73 (Dep.)

</td>
</tr>
</table>

Calculate Conditional Probabilities with the Conditional Probability Formula	
Question What is the probability you drew a blue M&M if you know you drew a peanut M&M? The probability of drawing a blue peanut M&M is 0.0419 and the probability of drawing a peanut M&M is 0.2042. **Define events** A and B as follows… A = draw a blue M&M B = draw a peanut M&M **Procedure** Organize what is given in the problem… • $P(A\ and\ B) = 0.042$ • $P(B) = 0.204$ **Formula** $\mathrm{P(A\vert B)} = \frac{P(A\ and\ B)}{P(B)}$ **Answer** $\mathrm{P(A\vert B)} = \frac{0.0419}{0.2042} = 0.205$ **Contingency Table…**$\mathrm{P(A\vert B)} = \frac{66}{322} = 0.204$	**Question** What is the probability you drew a plain M&M if you know you drew a red M&M? The probability of drawing a red plain M&M is 0.069 and the probability of drawing a red M&M is 0.094. **Define events** A and B as follows… A = draw a red M&M B = draw a plain M&M **Procedure** Organize what is given in the problem… • $P(A\ and\ B) = 0.069$ • $P(A) = 0.094$ **Formula** $\mathrm{P(B\vert A)} = \frac{P(A\ and\ B)}{P(A)}$ **Answer** $\mathrm{P(B\vert A)} = \frac{0.069}{0.094} = 0.73$ **Contingency Table…** $\mathrm{P(B\vert A)} = \frac{109}{149} = 0.73$

4.16 Example Problem: The Probability of One Event **AND** Another

Calculate Joint Probabilities with the Multiplication Rule	
Independent Events	**Dependent Events**
Question If you rolled a 5 on the first roll of a die, what is the probability of rolling a 3 on the second roll?	**Question** What is the probability you drew a blue peanut M&M? The proportion of peanut M&Ms that are blue is 0.205. The proportion of M&Ms that are peanut is 0.2042.
Define events A and B as follows… A = roll a 5 on the first roll B = roll a 3 on the second roll	**Define events** A and B as follows… A = draw a blue M&M B = draw a peanut M&M
Procedure Recognize these are independent events and multiply the probability of rolling a 5 by the probability of rolling a 3. The probability of rolling a 5 on the first roll does not affect the probability of rolling a 3 on the second roll. $\mathrm{P(B\|A)} = \mathrm{P(B)} = \frac{1}{6}$.	**Procedure** Recognize the probability "proportion of peanut M&Ms" is a conditional probability conditioned on type. Also recognize the probability "proportion of M&Ms" is a marginal probability. Use the multiplication rule for dependent events: Multiply the conditional probability by the marginal probability.
Formula $\mathrm{P(A\ and\ B)} = \mathrm{P(A)} * \mathrm{P(B)}$	**Formula** $\mathrm{P(A\ and\ B)} = \mathrm{P(A\|B)} * \mathrm{P(B)}$
Answer $\mathrm{P(A\ and\ B)} = \frac{1}{6} * \frac{1}{6} = \frac{1}{36} = 0.028$	**Answer** $\mathrm{P(A\ and\ B)} = 0.2042 * 0.205 = 0.042$

4.17 Practice Problems: Introduction to Probability

4.17.1 Exercise 1: Tossing a Coin Three Times

a. Given an experiment in which a fair coin is tossed three times, the sample space is $S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$. Event A is defined as tossing one head (H). What is the event A^c and what is the probability of this event?

b. An experiment consists of tossing three fair coins. What is the probability of tossing two tails?

4.17.2 Exercise 2: Mutual Funds A and B

Alison has all her money invested in two mutual funds, A and B. She knows that there is a 45% chance that fund A will rise in price, and a 53% chance that fund B will rise in price given that fund A rises in price.

a. What is the probability that both fund A and fund B will rise in price?

b. There is also a 25% chance that fund B will rise in price. What is the probability that at least one of the funds will rise in price?

c. What is the probability that neither fund will rise in price?

4.17.3 Exercise 3: Smoking Students

Records show that 10% of all college students are foreign students who also smoke. It is also known that 45% of all foreign college students smoke. What percent of the students at this university are foreign?

4.17.4 Exercise 4: P(B|A)

Let P(A) = 0.4 and P(B) = 0.7. Suppose A and B are independent. What is the value of $P(B|A)$?

4.17.5 Exercise 5: Job Applications

Peter applied to an accounting firm and a consulting firm. He knows that 35% of similarly qualified applicants receive job offers from the accounting firm, while only 25% of similarly qualified applicants receive job offers from the consulting firm. Assume that receiving an offer from one firm is independent of receiving an offer from the other. What is the probability that both firms offer Peter a job?

4.17.6 Exercise 6: Stock Prices

The likelihood of Company A's stock price rising is 30%, and the likelihood of Company B's stock price rising is 40%. Assume that the returns of Company A and Company B stock are independent of each other. The probability that the stock price of at least one of the companies will rise is ______.

4.17.7 Exercise 7: Assembly Parts

A manufacturing firm just received a shipment of 35 assembly parts, of slightly varied sizes, from a vendor. The manager knows that there are only 28 parts in the shipment that would be suitable. He examines these parts one at a time.

a. Find the probability that the first part is suitable.

b. If the first part is suitable, find the probability that the second part is also suitable.

c. If the first part is suitable, find the probability that the second part is not suitable.

4.17.8 Exercise 8: Preferred Exercise

The contingency table below provides frequencies for the preferred type of exercise for people under the age of 35 and those 35 years of age or older.

a. Find the probability that an individual prefers running.

b. Find the probability that an individual prefers biking given that he or she is 35 years old or older.

Age Group	Preferred Form of Exercise			Total
	Running	Biking	Swimming	
Under 35 years	157	121	79	357
35 years or older	45	27	87	159
Total	202	148	166	516

4.17.9 Exercise 9: Favorite Subject

The following probability table shows probabilities concerning Favorite Subject and Gender.

a. What is the probability of selecting an individual who is a female or prefers science?
b. What is the probability of selecting an individual preferring science if she is female?

Gender	Favorite Subject			Total
	Math	English	Science	
Male	0.2	0.07	0.155	0.425
Female	0.2	0.125	0.25	0.575
Total	0.4	0.195	0.405	1

4.17.10 Exercise 10: Read a Book

Two hundred people were asked if they had read a book in the last month. The accompanying contingency table, cross-classified by age, is produced.

a. The probability that a respondent is at least 30 years old is the closest to ______.
b. The probability that a respondent read a book in the last month and is at least 30 years old is the closest to _______.

	Under 30	30+
Yes	80	61
No	25	34

4.17.11 Exercise 11: Mark Zuckerberg

Mark Zuckerberg, the founder of Facebook, announced that he will eat meat only from animals that he has killed himself (*Vanity Fair*, November 2011). Suppose 257 people were asked, "Does the idea of killing your own food appeal to you, or not?" The accompanying contingency table, cross-classified by gender, is produced from the 187 respondents.

a. The probability that a respondent to the survey is male is the closest to ____.
b. The probability that a respondent is male and feels that the idea of killing his own food is appealing is the closest to _____.
c. Given that the respondent is male, the probability that he feels that the idea of killing his own food is appealing is the closest to ____.

	Male	Female
Yes	40	15
No	51	81

4.17.12 Exercise 12: Wonderland Frozen Yogurt

2. The 150 residents of the town of Wonderland were asked their age and whether they preferred vanilla, chocolate, or swirled frozen yogurt. The results are displayed next.
   a. What is the probability that a randomly selected customer prefers vanilla?
   b. What is the probability a randomly selected customer prefers chocolate given he or she is at least 25 years old?
   c. What is the probability a randomly selected customer prefers swirled yogurt or is at least 25 years old?

	Chocolate	Vanilla	Swirl
Under 25 years old	50	20	10
At least 25 years old	20	35	15

1. 5/8, 3/8
2. 0.2385, 0.4615, 0.5385
3. 22%
4. 0.7
5. 0.0875
6. 0.58
7. 0.8, 0.794, 0.206
8. 0.39, 0.17
9. 0.73, 0.435
10. 0.475, 0.305
11. 0.487, 0.214, 0.44
12. 0.367, 0.286, 0.533

5 Discrete Probability Distributions

5.1 Objectives

- Construct a probability distribution table for a discrete random variable (probability mass function)
- Calculate the expected value, variance, and standard deviation of a discrete random variable
- Distinguish Between Discrete and Continuous Random Variables
- Describe the binomial setting
- Describe the binomial distribution and its parameters
- Calculate and interpret the binomial mean and standard deviation
- Calculate binomial probabilities

5.2 Discrete Random Variables (RV)

A **random variable** is a function that assigns numeric values to the outcomes of an experiment. A **probability distribution** is a table or an equation that links each outcome of a statistical experiment with its probability of occurrence.

Define X to be the face value of a fair die tossed once

S = {1, 2, 3, 4, 5, 6}

X	1	2	3	4	5	6
P(X = x)	1/6	1/6	1/6	1/6	1/6	1/6
P(X ≤ x)	1/6	2/6	3/6	4/6	5/6	6/6

Define X to be the number of heads when tossing a fair coin three times

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

X	0	1	2	3
P(X = x)	1/8	3/8	3/8	1/8
P(X ≤ x)	1/8	4/8	7/8	8/8

5.2.1 Expected Value of a Discrete Random Variable

If you sampled repeatedly from a given distribution and recorded each sampled response, what would the average response be? The expected value of a discrete distribution of X with k possible outcomes is…

$$Recall\text{: }Arithmetic\ Average = \frac{0+1+1+1+2+2+2+3}{8} = \left(0 \times \frac{1}{8}\right) + \left(1 \times \frac{3}{8}\right) + \left(2 \times \frac{3}{8}\right) + \left(3 \times \frac{1}{8}\right)$$

$$Recall\text{: }P(x_i) = \frac{Frequency\ of\ x_i}{total\ number\ of\ possible\ outcomes}.$$

$$E(X) = x_1\left(\frac{Frequency\ of\ x_1}{Total\ Number\ of\ Values}\right) + x_2\left(\frac{Frequency\ of\ x_2}{Total\ Number\ of\ Values}\right) + \cdots + x_k\left(\frac{Frequency\ of\ x_k}{Total\ Number\ of\ Values}\right)$$

$$E(X) = \mu_x = x_1(p_{x_1}) + x_2(p_{x_2}) + x_3(p_{x_3}) + \cdots + x_k(p_{x_k})$$

$$E(X) = \mu_x = 0\left(\frac{1}{8}\right) + 1\left(\frac{3}{8}\right) + 2\left(\frac{3}{8}\right) + 3\left(\frac{1}{8}\right) = \frac{3}{8} + \frac{6}{8} + \frac{3}{8} = \frac{12}{8} = 1.5$$

5.2.2 Variance and Standard Deviation of a Discrete Random Variable

The variance, V(X), of a discrete distribution of X with k possible outcomes is…

$$V(X) = (x_1 - \mu_x)^2(p_{x_1}) + (x_2 - \mu_x)^2(p_{x_2}) + (x_3 - \mu_x)^2(p_{x_3}) + \cdots + (x_k - \mu_x)^2(p_{x_k})$$

$$V(X) = (0 - 1.5)^2\left(\frac{1}{8}\right) + (1 - 1.5)^2\left(\frac{3}{8}\right) + (2 - 1.5)^2\left(\frac{3}{8}\right) + \cdots + (3 - 1.5)^2\left(\frac{1}{8}\right) = 0.75$$

$$SD(X) = \sqrt{0.75} = 0.87\ heads$$

5.3 Practice Problems: Discrete Random Variable

The following exercises are also in the Excel file:
Random Variables Practice Problems.xlsx

Answers to the following exercises are in the Excel file:
Random Variables Practice Problems KEY.xlsx

5.3.1 Exercise 1: Homes Sold

The number of homes sold by a realtor during a month has the following probability distribution:

Number Sold	Probability
0	0.30
1	0.40
2	0.30

a. What is the probability that the realtor will sell at least one house during a month?

b. What is the probability that the realtor sells no more than one house during a month?

c. What is the expected number of homes sold by the realtor during a month?

d. What is the standard deviation of the number of homes sold by the realtor during a month?

5.3.2 Exercise 2: Cars Sold

The number of cars sold by a car salesperson during each of the last 25 weeks is the following:

Number Sold	Frequency
0	8
1	12
2	5

a. What is the probability that the salesperson will sell one car during a week?

b. What is the probability that the salesperson sells no more than one car during a week?

c. What is the expected number of cars sold by the salesperson during a week?

d. What is the standard deviation of the number of cars sold by the salesperson during a week?

5.3.3 Exercise 3: Stock's Return

An analyst believes that a stock's return depends on the state of the economy, for which she has estimated the following probabilities:

State of the Economy	Probability	Return
Good	0.10	15%
Normal	0.65	13%
Poor	0.25	7%

According to the analyst's estimates, the expected return of the stock is ____.

5.3.4 Exercise 4: Year-end Stock Price

An analyst estimates that the year-end price of a stock has the following probabilities:

Stock Price	Probability
$80	0.10
$85	0.35
$90	0.40
$95	0.15

The stock's expected price at the end of the year is _______.

5.3.5 Exercise 5: Predicted Return

An analyst has constructed the following probability distribution for firm X's predicted return for the upcoming year.

Return	Probability
–5	0.15
0	0.30
5	0.45
10	0.10

The expected value and the variance of this distribution are _____ and _____.

1. 0.7, 0.7, 1, 0.77
2. 0.48, 0.8, 0.88, 0.7111
3. 11.7%
4. $88
5. 2.5, 18.75

5.4 The Binomial Random Variable (RV)

The **Binomial random variable** counts the number of successes in a fixed number of Bernoulli trials, **n**. A **Bernoulli** trial has the following properties…

- There are only two outcomes per trial, called success and failure (the "success" is what is being counted)
- The probability of success is the same on every trial

An experiment satisfies the Binomial setting if…

- The possible values for a Binomial random variable are the whole numbers from 0 to n
- The trials are Bernoulli trials
- The trials are independent

Examples

- The number of tails when tossing a coin **n** times
- The number of no responses from **n** 'yes/no' surveys
- The number of baskets made on **n** free throws in basketball

5.4.1 The Binomial Distribution

Define the random variable: X = the number of 'successes' in **n** observations with **p** probability of success.

The expected value and variance of a Binomial random variable can be determined by knowing the number of trials, **n**, and the probability of success, **p**, so a Binomial distribution is defined by the two parameters, n and p:

$X \sim Bin(n, p)$

n = the number of observations

p = probability of success on any one sampled observation*

*The probability of failure is (1-p) or q

5.4.2 Construct the Probability Mass Function

Because trials in a Binomial setting are independent, use the multiplication formula P(A and B) = P(A)P(B).

Flip a biased coin 3 times, where p(heads) = 0.2…

X	0	1	2	3
P(X=x)	0.512	0.384	0.096	0.008

{TTT}
P(X=0) = 0.8 * 0.8 * 0.8 = 0.512

{HTT, THT, TTH}
P(X=1) = (**0.2** * 0.8 * 0.8) + (0.8 * **0.2** * 0.8) + (0.8 * 0.8 * **0.2**) = 0.384

{HHT, HTH, THH}
P(X=2) = (**0.2** * **0.2** * 0.8) + (**0.2** * 0.8 * **0.2**) + (0.8 * **0.2** * **0.2**) = 0.096

{HHH}
P(X=3) = (**0.2** * **0.2** * **0.2**) = 0.008

5.4.3 Calculate the Expected Value and Variance

$E(X) = (0)(0.512) + (1)(0.384) + (2)(0.096) + (3)(0.008) = 0.6\ heads$

$V(X) = (0 - 0.6)^2(0.512) + (1 - 0.6)^2(0.384) + (2 - 0.6)^2(0.096) + (3 - 0.6)^2(0.008) = 0.48$

$\mathrm{SD(X)} = \sqrt{0.48} = 0.6928$

5.4.4 Formulas for Calculating the Expected Value and Variance

$\mathrm{E(X)} = np$

$V(X) = np(1-p)$

$SD(X) = \sqrt{np(1-p)}$

1. Flip a biased coin 3 times, where p(heads) = 0.2…

   $E(X) = np = 3 * 0.2 = 0.6\, heads$

   $V(X) = np(1-p) = 3 * 0.2 * 0.8 = 0.48$

   $\mathrm{SD(X)} = \sqrt{0.48} = 0.6928$

2. Flip the biased coin 200 times and count the number of heads, n = 200 and p = 0.2.

   This is too large to do by hand.

   $E(X) = 200 * 0.2 = 40$ heads

   $V(X) = 200 * 0.2 * 0.8 = 32$

   $SD(X) = \sqrt{32} = 5.6\, heads$

5.4.5 Formula for Calculating Probability (The Binomial Probability Function)

$$P(X = k\ successes) = \binom{n}{k} p^k (1-p)^{n-k}$$

For k = 0, 1, 2, 3, …, n

5.4.6 Calculate Binomial Probabilities

Flip a biased coin 3 times, where p(heads) = 0.2…

X	0	1	2	3
P(X=x)	0.512	0.384	0.096	0.008

1. What is the probability of seeing 0 heads?
2. What is the probability of seeing 1 heads?
3. What is the probability of seeing 2 heads?
4. What is the probability of seeing 3 heads?
5. What is the probability of seeing less than 2 heads?
6. What is the probability of seeing 2 or more heads?
7. What is the probability of seeing at least 1 head?
8. What is the probability of seeing at most 2 heads?
9. What is the probability of seeing no more than 1 head?

5.4.7 Derive the Binomial Probability Function

Flip a biased coin 3 times, where p(heads) = 0.2…

X	0	1	2	3
P(X=x)	0.512	0.384	0.096	0.008

Recall the following probability…

P(X=1) = (**0.2** * 0.8 * 0.8) + (0.8 * **0.2** * 0.8) + (0.8 * 0.8 * **0.2**) = 0.384

By the commutative property of multiplication…

P(X=1) = $(0.2)^1(0.8)^2 + (0.2)^1(0.8)^2 + (0.2)^1(0.8)^2$

Repeated addition is multiplication…

P(X=1) = $3*(0.2)^1(0.8)^2$

There are 3 ways to arrange 1 head when flipping a coin 3 times.

How do we extend this idea to a large experiment where it is difficult or time consuming to list all the ways of arranging the outcomes?

The number of ways of arranging k successes among n observations is given by the **binomial coefficient**, denoted $\binom{n}{k}$ and pronounced "n choose k,"

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

for k = 0, 1, 2, 3, …, n.

$\binom{n}{k}$ is NOT a fraction!

$\binom{n}{k}$ can also be written ${}_nC_k$, and this is how the binomial coefficient is represented on some calculators!

5.5 Distinguish Between Discrete and Continuous Random Variables

<table>
<tr><th colspan="2">RANDOM VARIABLES</th></tr>
<tr><th>DISCRETE</th><th>CONTINUOUS</th></tr>
<tr>
<td>
A discrete random variable assumes a countable number of distinct values.

Examples
Discrete uniform, Binomial, Poisson

Probability Distribution
Probability Mass Function, $P(X = x)$
Cumulative Distribution Function, $P(X \leq x)$

Properties of the Mass Function

Describes all possible values x with the associated probabilities $P(X = x)$.
The probability of each value x is a value between 0 and 1, $0 \leq P(X = x) \leq 1$.
The sum of the probabilities equals 1 where the sum extends over all values of x.

</td>
<td>
A continuous random variable assumes uncountable values in an interval.

Examples
Continuous uniform, Normal

Probability Distribution
Probability Density Function, $f(x) = P(a < X < b)$
Cumulative Distribution Function, $F(x) = P(X \leq x)$

Properties of the Density Function

$f(x) \geq 0$ for all possible values x of X.
The area under f(x) over all values of x equals 1.
It is only meaningful to calculate the probability the random variable assumes a value within an interval…

Cannot assign nonzero probabilities to each of the uncountable values.
$P(a < X < b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b)$

$P(a < X < b)$ is the area under the curve f(x) between points a and b.

</td>
</tr>
</table>

5.6 Practice Problems: Binomial Random Variables

The following exercises are also in the Excel file:
Excel File: Binomial Practice Problems.xlsx

Answers to the following exercises are in the Excel file:
Binomial Practice Problems KEY.xlsx

5.6.1 Exercise 1: Calculators

It is known that 12% of the calculators shipped from a particular factory are defective.

a. What is the probability that exactly three of six chosen calculators are defective?

b. What is the probability that none in a random sample of six calculators is defective?

c. What is the probability that at least one in a random sample of six calculators is defective?

5.6.2 Exercise 2: CFA Candidates

Twenty-five percent of the CFA candidates have a degree in economics. A random sample of four CFA candidates is selected.

a. What is the probability that none of them has a degree in economics?

b. What is the probability that at least one of them has a degree in economics?

5.6.3 Exercise 3: Light Bulbs

On a particular production line, the likelihood that a light bulb is defective is 8%. Fifteen light bulbs are randomly selected.

a. What is the probability that two light bulbs will be defective?

b. What is the probability that none of the light bulbs will be defective?

c. What are the mean and variance of the number of defective bulbs?

5.6.4 Exercise 4: Clothing Store Coupons

For a particular clothing store, a marketing firm finds that 22% of $10-off coupons delivered by mail are redeemed. Suppose ten customers are randomly selected and are mailed $10-off coupons.

a. What is the probability that three of the customers redeem the coupon?

b. What is the probability that no more than one of the customers redeems the coupon?

c. What is the probability that at least two of the customers redeem the coupon?

d. What is the expected number of coupons that will be redeemed?

5.6.5 Exercise 5: Detroit Unemployment

According to a Department of Labor report, the city of Detroit had a 18% unemployment rate in May of 2011. Five working-age residents were chosen at random.

a. What is the probability that exactly one of the residents was unemployed?

b. What is the probability that at least two of the residents were unemployed?

c. What is the probability that exactly four residents were unemployed?

d. What was the expected number of unemployed residents when five working-age residents were randomly selected?

5.6.6 Example 6: Chauncey Billups

Chauncey Billups, a current shooting guard for the Los Angeles Clippers, has a career free-throw percentage of 89.4%. Suppose he shoots four free throws in tonight's game.

a. What is the probability that Billups makes all four free throws?

b. What is the probability that Billups makes three or more of his free throws?

c. What is the expected number of free throws that Billups will make?

d. What is the standard deviation of the number of free throws that Billups will make?

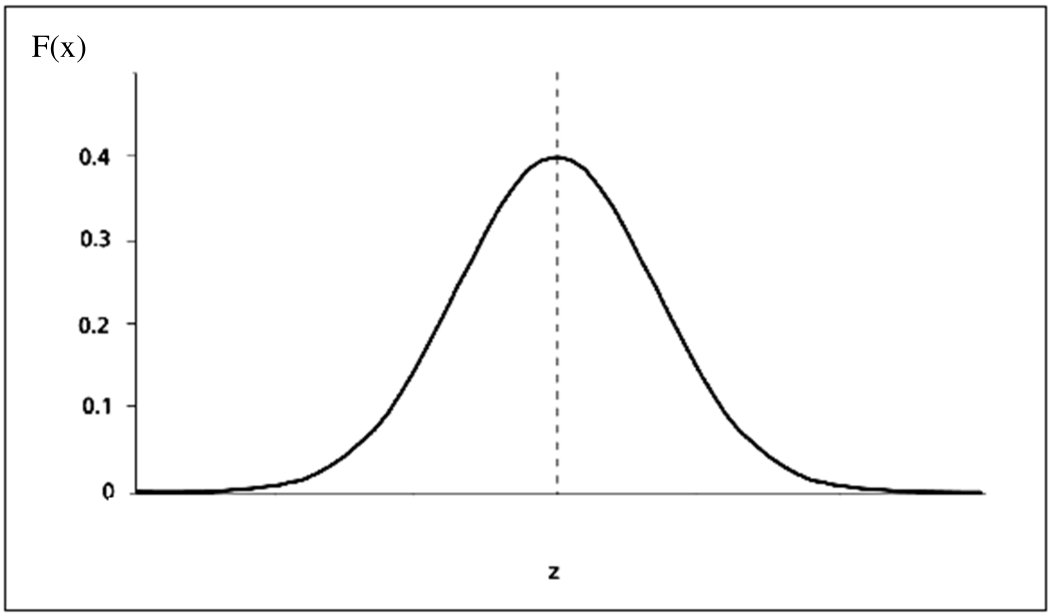
1. 0.0236, 0.4644, 0.5356
2. 0.3164, 0.6836
3. 0.2273, 0.2863, 1.2, 1.104
4. 0.2244, 0.3185, 0.6815, 2.2
5. 0.4069, 0.2224, 0.0043, 0.9
6. 0.6388, 0.9417, 3.576, 0.6157

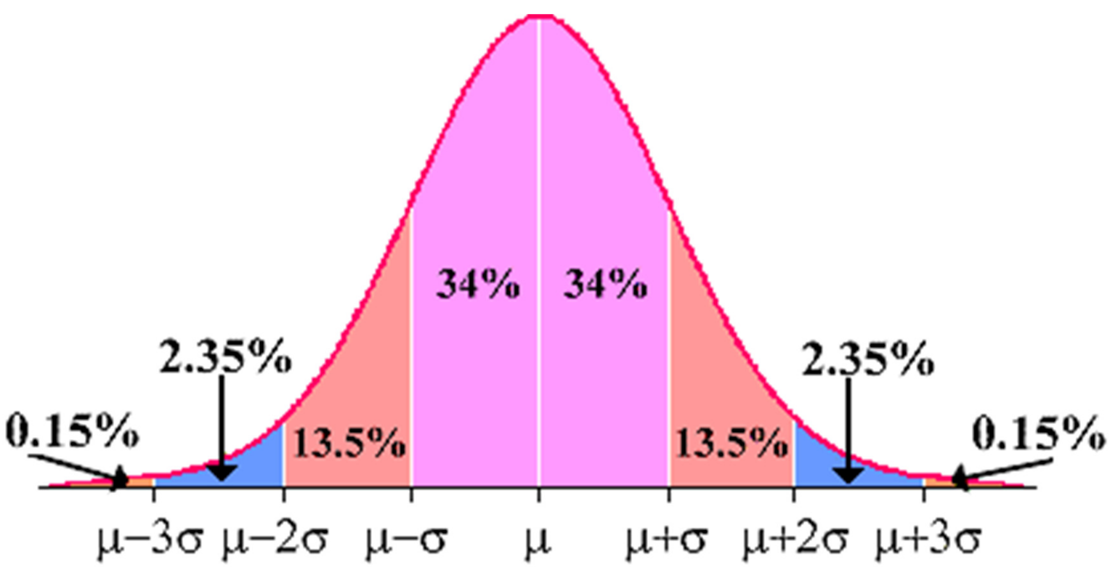
[This page intentionally left blank]

6 Continuous Probability Distributions

6.1 Objectives

- Convert data values to Z-scores and interpret the relative location.
- Convert Z-scores to data values.
- Find probabilities, given a z-score, using the standard normal z-table.
- Find z-scores, given a probability, using the standard normal z-table.





6.2 Practice Problems: Z-Table Lookup

6.2.1 Exercise 1: Find Probability

a. The probability $P(Z < -1.28)$ is closest to _______.

b. The probability $P(Z > 1.28)$ is closest to _______.

6.2.2 Exercise 2: Find Probability

a. Find the probability $P(-1.96 \leq Z \leq 0)$.

b. Find the probability $P(-1.96 \leq Z \leq 1.96)$.

c. Find the probability $P(-2 \leq Z \leq 2)$.

6.2.3 Exercise 3: Work Boots

You work in marketing for a company that produces work boots. Quality control has sent you a memo detailing the length of time before the boots wear out under heavy use. They find that the boots wear out in an average of 208 days, but the exact amount of time varies, following a normal distribution with a standard deviation of 14 days. For an upcoming ad campaign, you need to know the percent of the pairs that last longer than six months—that is, 180 days.

6.2.4 Exercise 4: Soapbox Derby

The time to complete the construction of a soapbox derby car is normally distributed with a mean of 3.5 hours and a standard deviation of 1.2 hours.

a. Find the probability that it would take more than 5 hours to construct a soapbox derby car.

b. Find the probability that it would take between 3 and 4 hours to construct a soapbox derby car.

c. Find the probability that it would take exactly 4.2 hours to construct a soapbox derby car.

6.2.5 Exercise 5: Denali National Park

You are planning a May camping trip to Denali National Park in Alaska and want to make sure your sleeping bag is warm enough. The average low temperature in the park for May follows a normal distribution with a mean of 30°F and a standard deviation of 12°F. One sleeping bag you are considering advertises that it is good for temperatures down to 25°F.

a. What is the probability that this bag will be warm enough on a randomly selected May night at the park?

b. What is the probability that the bag will *not* be warm enough?

6.2.6 Exercise 6: Alaskan Gold

Gold miners in Alaska have found, on average, 14 ounces of gold per 1,000 tons of dirt excavated with a standard deviation of 4 ounces. Assume the amount of gold found per 1,000 tons of dirt is normally distributed.

a. What is the probability the miners find more than 18 ounces of gold in the next 1,000 tons of dirt excavated?

b. What is the probability the miners find between 15 and 21 ounces of gold in the next 1,000 tons of dirt excavated?

6.2.7 Exercise 7: Laptop Battery

Suppose the life of a particular brand of laptop battery is normally distributed with a mean of 9 hours and a standard deviation of 0.7 hours. What is the probability that the battery will last more than 10 hours before running out of power?

6.2.8 Exercise 8: Baseball Player Run Average

A superstar major league baseball player just signed a new deal that pays him a record amount of money. The star has driven in an average of 115 runs over the course of his career, with a standard deviation of 28 runs. An average player at his position drives in 85 runs. What is the probability the superstar bats in fewer runs than an average player next year? Assume the number of runs batted in is normally distributed.

1. 0.1003, 0.1003
2. 0.4750, 0.95, 0.9544
3. 0.9772
4. 0.1056, 0.3256, 0
5. 0.6628, 0.3372
6. 0.1587, 0.3612
7. 0.0764
8. 0.1423

6.3 Practice Problems: Reverse Z-Table Lookup

6.3.1 Exercise 1: Find z

a. Find the z value such that $P(Z \le z) = 0.9082$.

b. Find the z value such that $P(-z \le Z \le z) = 0.95$.

6.3.2 Exercise 2: Find x

a. Let X be normally distributed with mean $\mu = 250$ and standard deviation $\sigma = 80$. Find the value x such that $P(X \le x) = 0.0606$.

b. Let X be normally distributed with mean $\mu = 250$ and standard deviation $\sigma = 80$. Find the value x such that $P(X \le x) = 0.9394$.

c. Let X be normally distributed with mean $\mu = 25$ and standard deviation $\sigma = 5$. Find the value x such that $P(X \ge x) = 0.1736$.

6.3.3 Exercise 3: Teacher Salaries

The salary of teachers in a particular school district is normally distributed with a mean of \$60,000 and a standard deviation of \$3,500. Due to budget limitations, it has been decided that the teachers who are in the top 2.5% of the salaries would not get a raise. What is the salary level that divides the teachers into one group that gets a raise and one that doesn't?

6.3.4 Exercise 4: Administrative Assistant Salary

The starting salary of an administrative assistant is normally distributed with a mean of $55,000 and a standard deviation of $2,700. We know that the probability of a randomly selected administrative assistant making a salary between μ - x and μ + x is 0.7416. Find the salary range referred to in this statement.

6.3.5 Exercise 5: Stock Price

The stock price of a particular asset has a mean and standard deviation of $62.50 and $7.25, respectively. Use the normal distribution to compute the 95th percentile of this stock price.

6.3.6 Exercise 6: Denali National Park

You are planning a May camping trip to Denali National Park in Alaska and want to make sure your sleeping bag is warm enough. The average low temperature in the park for May follows a normal distribution with a mean of 30°F and a standard deviation of 12°F. Above what temperature must the sleeping bag be suited such that the temperature will be too cold only 5% of the time?

6.3.7 Exercise 7: Alaskan Gold

Gold miners in Alaska have found, on average, 14 ounces of gold per 1,000 tons of dirt excavated with a standard deviation of 4 ounces. Assume the amount of gold found per 1,000 tons of dirt is normally distributed. If the miners excavated 1,000 tons of dirt, how little gold must they have found such that they find that amount or less only 15% of the time?

1. 1.33, 1.96
2. 126, 374, 29.7
3. $66,860
4. $51,949 to $58,051
5. $74.43
6. 10.26°
7. 9.84 ounces

7 Case Studies

7.1 Case Study 1: Tables, Graphs, and Numeric Summaries

The following exercises are also in the Excel file:
Pelican Stores.xlsx

Answers to the following exercises are in the Excel file:
Pelican Stores KEY.xlsx

Pelican Stores, a division of National Clothing, is a chain of women's apparel stores operating throughout the country. The chain recently ran a promotion in which discount coupons were sent to customers of other National Clothing stores.

The ***Proprietary card*** method of payment refers to charges made using a National Clothing charge card. Customers who made purchases using a discount coupon are referred to as ***promotional customers***. Because the promotional coupons were not sent to regular Pelican Stores customers, ***management considers the sales made to people presenting the promotional coupon as sales it would not otherwise make.*** Of course, Pelican also hopes that the promotional customers will continue to shop at its stores.

Variable	**Description**
Customer ID:	Unique Identifier
Type of customer:	Regular, Promotional (promotional customer received discount coupon)
Items:	The total number of items purchased
Net Sales:	The total amount in dollars charged to the credit card
Method of Payment:	Discover, Visa, MasterCard, American Express, Proprietary Card
Gender:	Male, Female
Marital Status:	Married, Single
Age:	Customer age in years

1. Identify the type of data (qualitative/quantitative) and the level of measurement (nominal or ordinal/interval or ratio) for the variable, ***Marital Status***.

2. What would be the appropriate type of graph to visually display the distribution of ***Marital Status***?

Pie or bar chart

3. Identify the type of data (qualitative/quantitative) and the level of measurement (nominal or ordinal/interval or ratio) for the variable, ***Age***.

4. What would be the appropriate type of graph to visually display the distribution of ***Age***?

Box plot or histogram

An example of the first 8 entries of the Pelican Stores sales transactions are shown below.

Customer	Type of Customer	Items	Net Sales	Method of Payment	Gender	Marital Status	Age
1	Regular	1	39.50	Discover	Male	Married	32
2	Promotional	1	102.40	Proprietary Card	Female	Married	36
3	Regular	1	22.50	Proprietary Card	Female	Married	32
4	Promotional	5	100.40	Proprietary Card	Female	Married	28
5	Regular	2	54.00	MasterCard	Female	Married	34
6	Regular	1	44.50	MasterCard	Female	Married	44
7	Promotional	2	78.00	Proprietary Card	Female	Married	30
8	Regular	1	22.50	Visa	Female	Married	40

The frequency, relative frequency, and cumulative relative frequency distributions for ***Net Sales*** are given below.

BinsSales	Frequency	Relative Frequency	Cumulative Relative Frequency
13 to <44	27	0.27	0.27
44 to <75	37	0.37	0.64
75 to <106	14	0.14	0.78
106 to <137	8	0.08	0.86
137 to <168	7	0.07	0.93
168 to <199	3	0.03	0.96
199 to <230	1	0.01	0.97
230 to <261	1	0.01	0.98
261 to <292	2	0.02	1.00
Total	100	1	

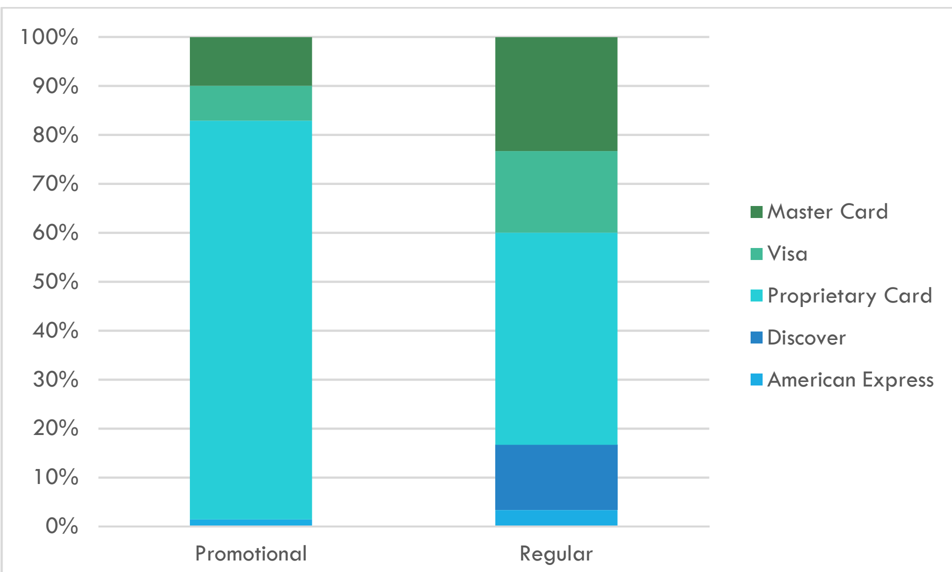
5. How many customers spent at least $168 but less than $230? 4

6. What is the relative frequency of customers who spent at least $44 but less than $75? 0.37

7. What percentage of customers spent less than $137? At least $199? 86%/4%

8. Would you describe the distribution of ***Net Sales*** as symmetric or skewed? right

9. Briefly summarize the distribution of ***Net Sales***…describe the typical customer.
Generally, the typical customer spends between $44 and $75. Or…Generally, the typical customer spends between $13 and $75.



Type of Customer	American Express	Discover	Proprietary Card	Visa	Master Card	Grand Total
Promotional	1.43%	0.00%	81.43%	7.14%	10.00%	100.00%
Regular	3.33%	13.33%	43.33%	16.67%	23.33%	100.00%
Grand Total	**2.00%**	**4.00%**	**70.00%**	**10.00%**	**14.00%**	**100.00%**

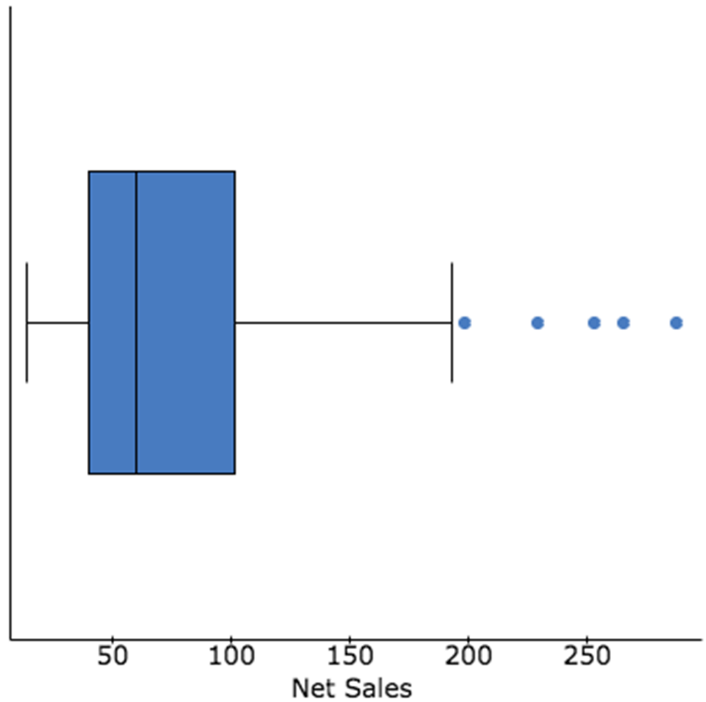
10. Using the two way pivot table and stacked bar chart, would you say that type of customer influences the method of payment. Explain. For each bar, the dark shaded area represents promotional customers and the light shaded area represents regular customers.

11. For the distribution of **Net Sales**, do you expect the mean to be equal to, less than, or greater than the median?

12. What measure of center and spread would best summarize the distribution, mean/standard deviation or median/5-number summary?

Net Sales	
Mean	77.60
Standard Deviation	55.66
Minimum	13.23
Q1	39.6
Median	59.71
Q3	101.9
Maximum	287.59

13. Sketch a box plot of the distribution of ***Net Sales***. Find outliers using the 1.5 x IQR method.



IQR = 101.9 – 39.6 = 62.3

1.5 X IQR = 93.45

Q1 – 93.45 = 39.6 – 93.45 = -53.85

Lower boundary = $-53.85

Q3 + 93.45 = 101.9 + 93.45 = 195.35

Upper boundary = $195.35

High Outliers: $198.80, 229.50, 253.00, 266.00, 287.59

7.2 Case Study 2: Normal Distribution Cases

7.2.1 BMI for 10-Year-Old-Boys

Body mass index (BMI) is a reliable indicator of body fat for most children and teens. BMI is calculated from a child's weight and height and is used as an easy-to-perform method of screening for weight categories that may lead to health problems. For children and teens, BMI is age- and sex-specific and is often referred to as BMI-for-age.

The Centers for Disease Control and Prevention (CDC) reports BMI-for-age growth charts for girls as well as boys to obtain a percentile ranking. Percentiles are the most commonly used indicator to assess the size and growth patterns of individual children in the United States.

The following table provides weight status categories and the corresponding percentiles and BMI ranges for 10-year-old boys in the United States.

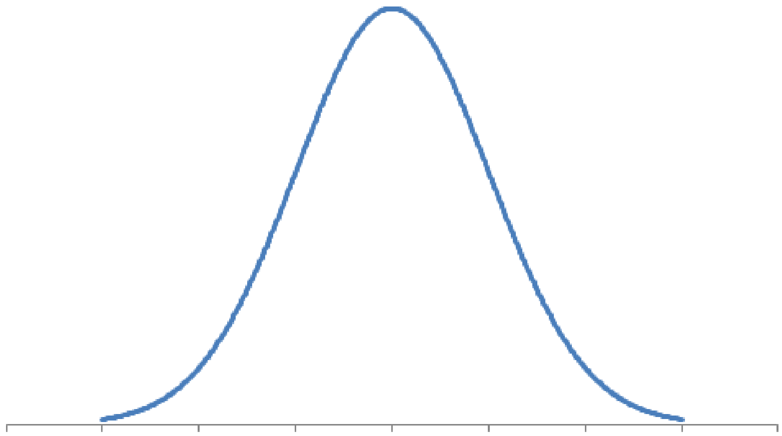
Weight Status Category	Percentile Range	BMI Range
Underweight	Less than 5^{th}	Less than 14.2
Healthy Weight	Between 5^{th} and 85^{th}	Between 14.2 and 19.4
Overweight	Between 85^{th} and 95^{th}	Between 19.4 and 22.2
Obese	More than 95^{th}	More than 22.2

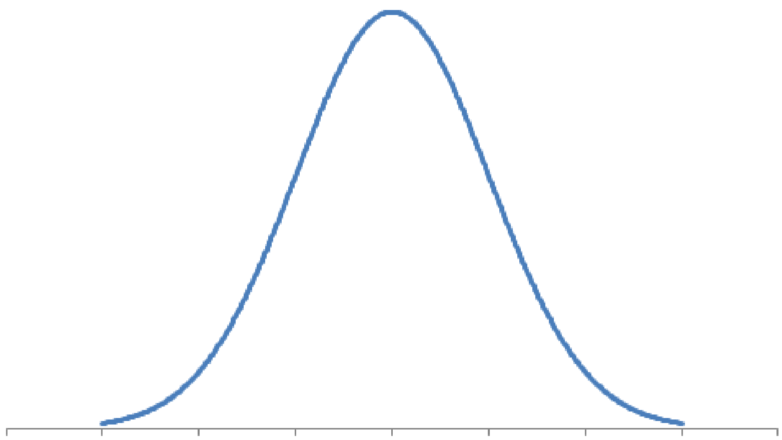
Health officials of a Midwestern town are concerned about the weight of children in their town. They believe that the BMI of their 10-year-old boys is normally distributed with mean 19.2 and standard deviation 2.6.

In a report, use the sample information to:

1. Compute the proportion of 10-year-old boys in this town that are in the various weight status categories given the BMI ranges.

Weight Status Category	Percentile Range	BMI Range
Underweight	Less than 2.74^{th}	Less than 14.2
Healthy Weight	Between 2.74^{th} and 53.19^{th}	Between 14.2 and 19.4
Overweight	Between 53.19^{th} and 87.49^{th}	Between 19.4 and 22.2
Obese	More than 87.49^{th}	More than 22.2





2. Discuss whether the concern of health officials is justified.

More than 12% of the boys in this Midwestern town are classified as Obese compared to only 5% specified by the CDC. More than 35% are classified as Overweight compared to only 10% specified by the CDC.

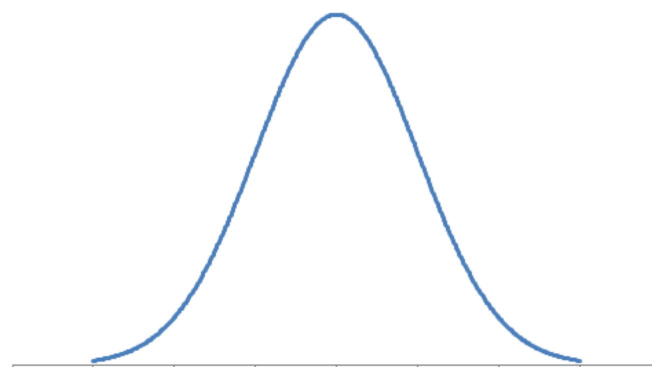
Less than 2.74^{th}, Between 2.74^{th} and 53.19^{th}, Between 53.19^{th} and 87.49^{th}, More than 87.49^{th}

7.2.2 Top Performing Mutual Funds

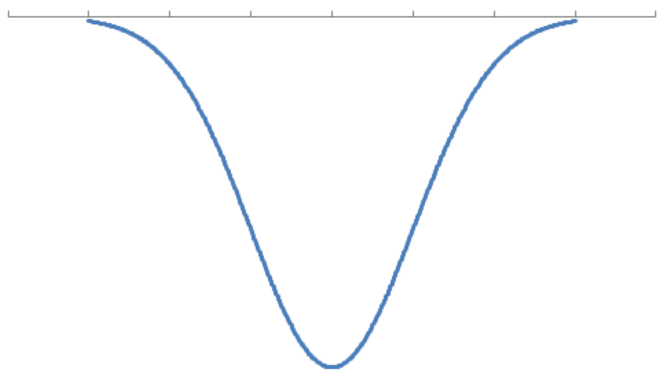
Vanguard's Precious Metals and Mining fund (Metals) and Fidelity's Strategic Income fund (Income) were two top-performing mutual funds for the years 2000 through 2009. An analysis of annual return data for these two funds provided important information for any type of investor. Over the past 10 years, the Metals fund posted a mean return of 24.65% with a standard deviation of 37.13%. On the other hand, the mean and the standard deviation of return for the Income fund were 8.51% and 11.07%, respectively. It is reasonable to assume that the returns of the Metals and the Income funds are both normally distributed, where the means and the standard deviations are derived from the 10-year sample period.

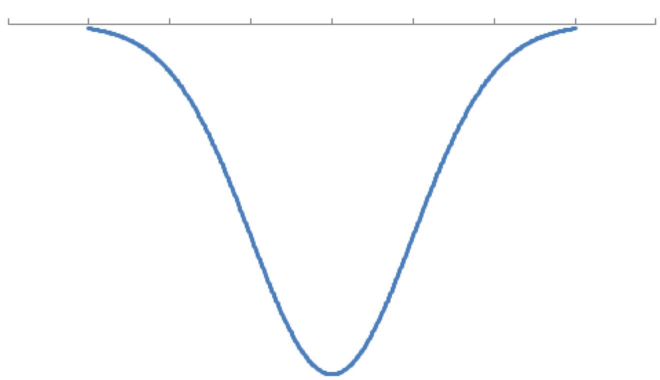
In a report, use the sample information to compare and contrast the Metals and Income funds from the perspective of an investor whose objective is to:

1. Minimize the probability of earning a negative return.
2. Maximize the probability of earning a return between 0% and 10%.
3. Maximize the probability of earning a return greater than 10%.



1. Metals: $P(x \leq 0) = P(z \leq -0.66) = 0.2546$
   Income: $P(x \leq 0) = P(z \leq -0.77) = 0.2206$
   Choose Income
2. Metals: $P(0 \leq x \leq 10) = P(-0.66 \leq z \leq -0.39) = 0.0937$
   Income: $P(0 \leq x \leq 10) = P(-0.77 \leq z \leq 0.13) = 0.3311$
   Choose Income
3. Metals: $P(x \geq 10) = P(z \geq -0.39) = 0.6517$
   Income: $P(x \geq 10) = P(z \geq 0.13) = 0.4483$
   Choose Metals





[This page intentionally left blank]

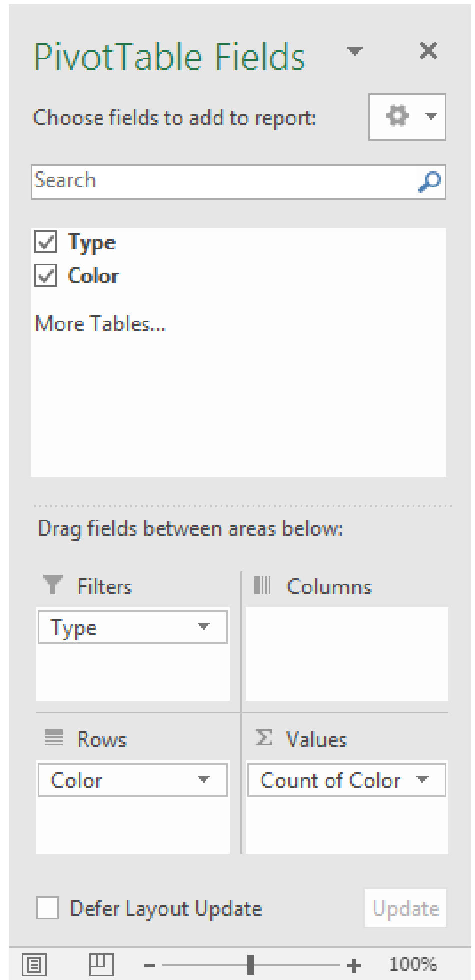
8 Excel: Construct Tables and Graphs

8.1 Qualitative Data

Excel File: M&M Tables and Graphs.xlsx

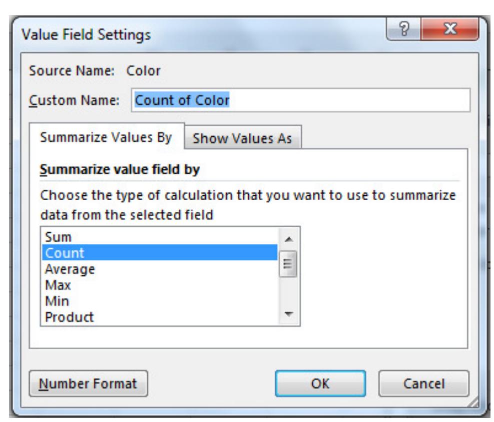
8.1.1 Construct a Frequency Distribution Table

1. Open the “M&M Color” sheet in the Excel file “M&M Tables and Graphs.xlsx.”
2. Select the entire table (click any cell within the table and press Ctrl-A).
3. Insert tab → Pivot Table → click OK (leave default settings).
4. Set the PivotTable Fields window to look like the screen shot to the right:
    a. Click and drag the variable “Type” to the Filters box and “Color” to the Rows box.
    b. Click and drag “Color” to the Values box. The default value should be the “Count of Color.”
    c. Change the Type by selecting (All), Peanut M&M, or Plain M&M.

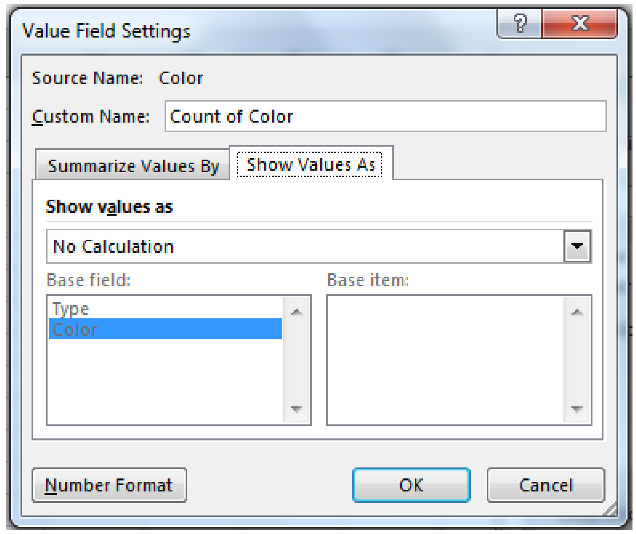


Type	Peanut M&M
Row Labels	**Count of Color**
Blue	67
Brown	36
Green	57
Orange	61
Red	42
Yellow	67
Grand Total	**330**

    d. Change the Values calculation…
        i. Click the black triangle next to the value in the Values box.
        ii. Select Value Field Settings….
        iii. Select Count in the Summarize Values By tab.



8.1.2 Construct a Relative Frequency Distribution Table



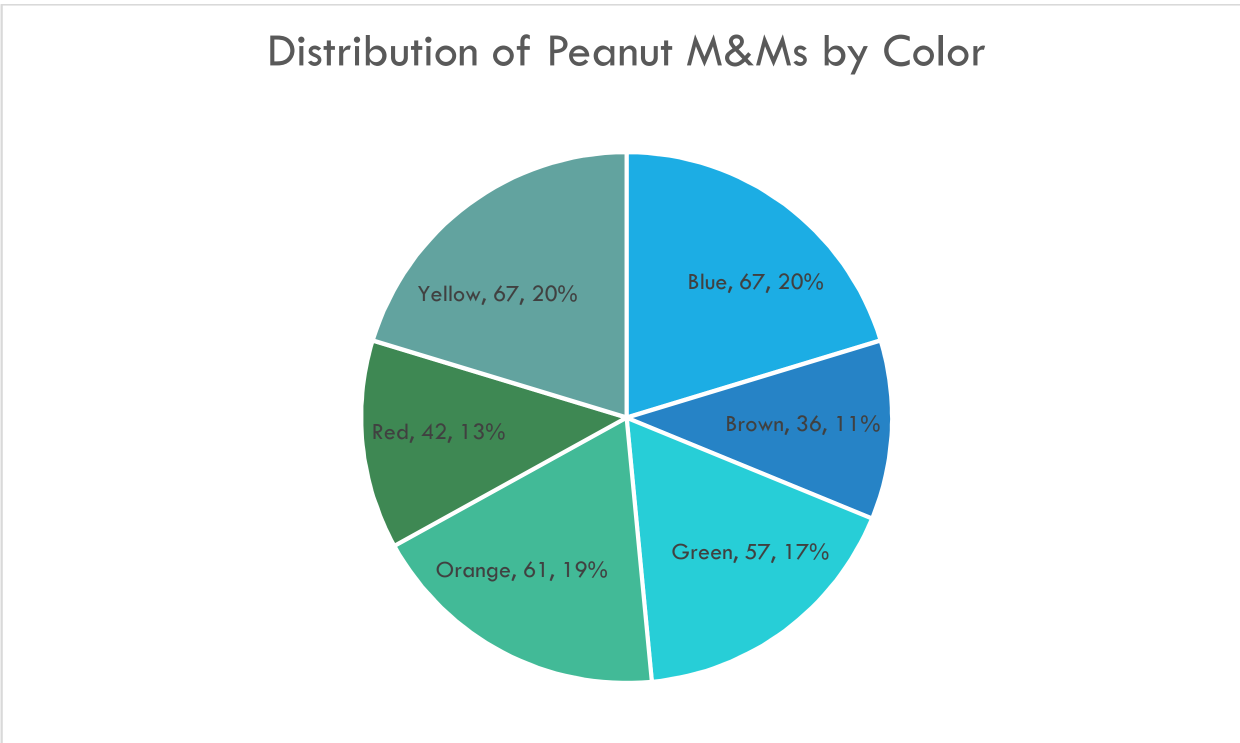
1. Open the “M&M Color” sheet in the Excel file “M&M Tables and Graphs.xlsx.”
2. Construct a frequency distribution table as described in 8.1.1 above.
3. Click any cell in the pivot table to activate the PivotTable Fields window.
4. Change the Values calculation to show relative or percent frequencies.
    a. Click the black triangle next to the value in the Values box.
    b. Select Value Field Settings….
    c. Select Show Values As tab.
    d. Open the Show values as drop down menu → select % of Column Total → OK.

Type	Peanut M&M

Row Labels	**Count of Color**
Blue	20.30%
Brown	10.91%
Green	17.27%
Orange	18.48%
Red	12.73%
Yellow	20.30%
Grand Total	**100.00%**

8.1.3 Construct a Pie Chart

1. Open the “M&M Color” sheet in the Excel file “M&M Tables and Graphs.xlsx.”
2. Construct a frequency distribution table as described in 8.1.1 above.
3. Click one of the cells in the pivot table to activate the PivotTable Fields window.
4. Insert tab → Charts → 2-D Pie
5. Right click on the pie → Add Data Labels → Add Data Labels
6. Right click on the pie → Format Data Labels → select Category Name and Percentage.
7. Select the legend and Delete.
8. Right click one of the buttons → Select Hide All Field Buttons on Chart.
9. Enlarge the graph to desired size.
10. Click Chart Title and type “Distribution of Peanut M&Ms by Color”

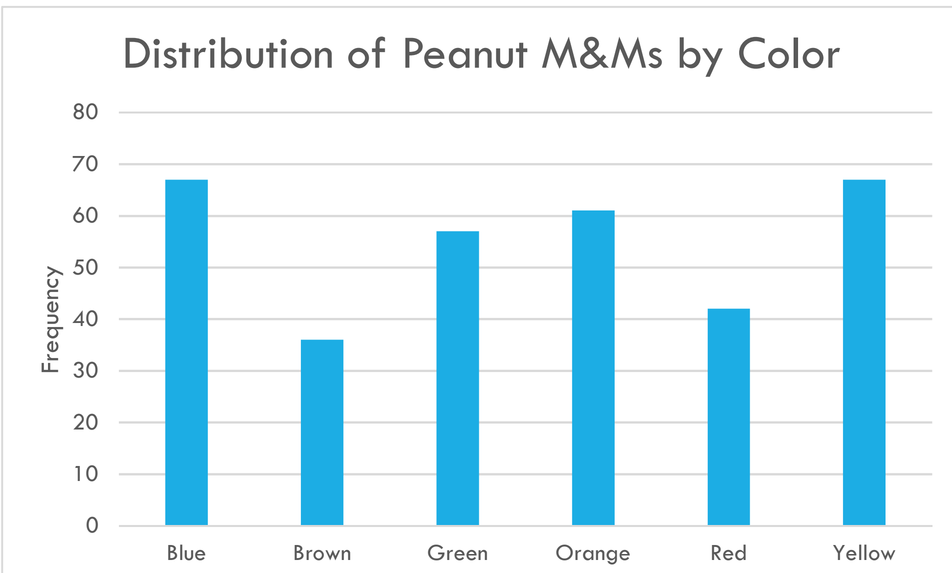


8.1.4 Construct a Bar Chart

1. Open the "M&M Color" sheet in the Excel file "M&M Tables and Graphs.xlsx."
2. Construct a frequency distribution table as described in 8.1.1 above.
3. Click one of the cells in the pivot table to activate the PivotTable Fields window.

Frequency Bar Chart

1. Insert tab → Charts → 2-D Column
2. Right click on one of the bars → Add Data Labels → Add Data Labels
3. Right click one of the buttons → Select Hide All Field Buttons on Chart.
4. Enlarge the graph to desired size.
5. Click Chart Title and type "Distribution of Peanut M&Ms by Color"
6. Go to the Design tab → select Add Chart Element → Axis Titles → select Primary Vertical → type "Frequency"



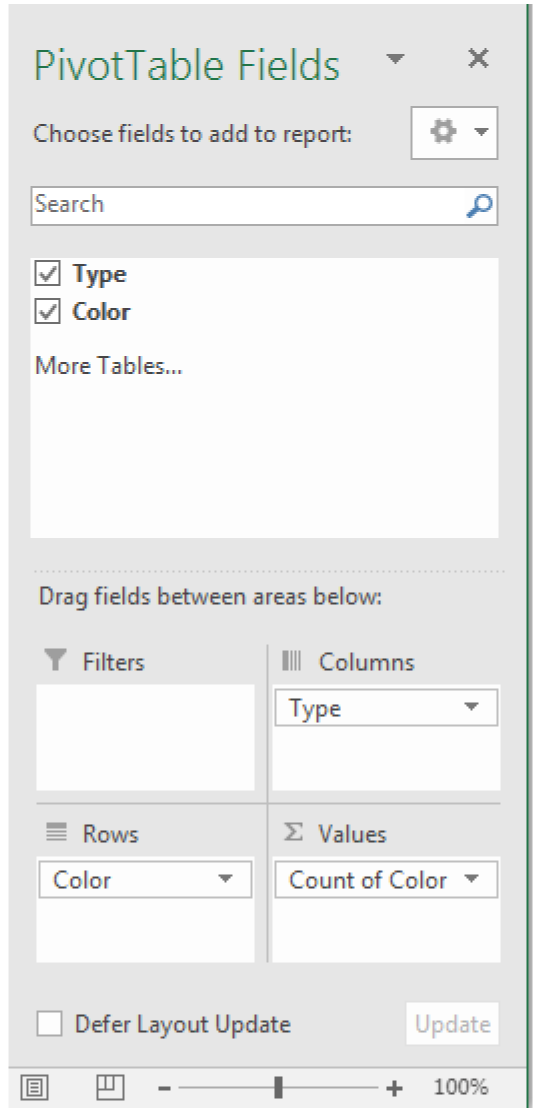
Relative Frequency Bar Chart

7. Right click on the vertical axis → select Format Axis… → expand Number → enter 0 in the Decimal places box
8. Select the vertical axis title → type "Proportion"

8.1.5 Construct a Contingency Table

1. Open the "M&M Color" sheet in the Excel file "M&M Tables and Graphs.xlsx."
2. Select the entire table (click any cell within the table and press Ctrl-A).
3. Insert tab → Pivot Table → click OK (leave default settings).
4. In the PivotTable Fields area on the right, click and drag the Color field to the Rows box → click and drag the Type field to the Columns box → drag the Color field to the Values box as shown on the right. The field should say Count of Color.

The count of observations in each combination of category pairs appears in the contingency table.



8.1.6 Construct a Percent Frequency Contingency Table

- Percent of Grand Total
- Percent of Column Totals
- Percent of Row Totals

1. Open the "M&M Color" sheet in the Excel file "M&M Tables and Graphs.xlsx."
2. Construct a frequency distribution table as described in 0 above.
3. Click to activate any cell in the table.
4. Click on the black triangle next to Count of Color in the Values box → select Value Field Settings… → select Show Values As → click to open the Show values as drop down box → select % of Grand Total, % of Column Total, or % of Row Total

Percent of Grand Total

The percent of grand total shows the proportion of each color/type category pair with respect to all the records. For example, 4.23% **of all M&Ms** are blue and peanut.

Row Labels	Peanut M&M	Plain M&M	Grand Total
Blue	4.23%	18.17%	22.40%
Brown	2.27%	7.07%	9.34%
Green	3.60%	19.81%	23.41%
Orange	3.85%	20.63%	24.48%
Red	2.65%	6.88%	9.53%
Yellow	4.23%	6.62%	10.85%
Grand Total	**20.82%**	**79.18%**	**100.00%**

Percent of Column Total

The percent of column total shows the proportion of each major with respect to the class ranks, separately. For example, 20.30% **of all peanut M&Ms** are blue.

Row Labels	Peanut M&M	Plain M&M	Grand Total
Blue	20.30%	22.95%	22.40%
Brown	10.91%	8.92%	9.34%
Green	17.27%	25.02%	23.41%
Orange	18.48%	26.06%	24.48%
Red	12.73%	8.69%	9.53%
Yellow	20.30%	8.37%	10.85%
Grand Total	**100.00%**	**100.00%**	**100.00%**

Percent of Row Total

The percent of row total shows the proportion of each category pair with respect to all of the records. For example, 18.87% **of all blue M&Ms** are peanut.

Row Labels	Peanut M&M	Plain M&M	Grand Total
Blue	18.87%	81.13%	100.00%
Brown	24.32%	75.68%	100.00%
Green	15.36%	84.64%	100.00%
Orange	15.72%	84.28%	100.00%
Red	27.81%	72.19%	100.00%
Yellow	38.95%	61.05%	100.00%
Grand Total	**20.82%**	**79.18%**	**100.00%**

8.1.7 Construct and Format Side-by-Side Bar Charts

A side-by-side bar chart is a visual representation of the distribution of two categorical variables.

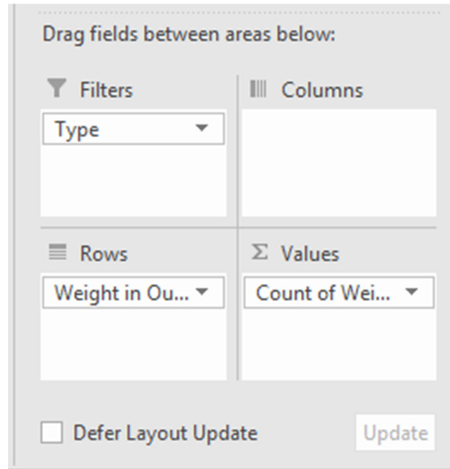
1. Open the "M&M Color" sheet in the Excel file "M&M Tables and Graphs.xlsx."
2. Construct a frequency distribution table as described in 0 above.
3. Insert tab → under Charts select 2-D Column or 2-D Bar.
4. Right click one of the buttons → Select Hide All Field Buttons on Chart.
5. Select the chart → click the Design tab → click Add Chart Element → Chart Title → select Above Chart → type your own title → press the Enter key.
6. Right click one of the columns or bars → Add Data Labels → select Add Data Labels (optional)
7. Repeat step 3 with a bar from each of the other series.
8. Go to the Design tab → select Add Chart Element → Axis Titles → select Primary Vertical → type "Frequency"

8.2 Quantitative Data

8.2.1 Construct a Grouped Frequency Distribution Table

Excel File: M&M Tables and Graphs.xlsx

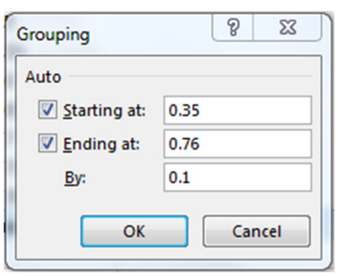
1. Open the Excel file "M&M Tables and Graphs.xlsx."
2. Open the "M&M Weights" sheet.
3. Select one of the cells that contains the data → click the Insert tab → click Pivot Table → click OK
4. In the PivotTable Fields area on the right, click and drag the Weight variable to the Rows box → click and drag the Weight variable to Values box → click and drag the Type variable to the Filter box as shown on the right



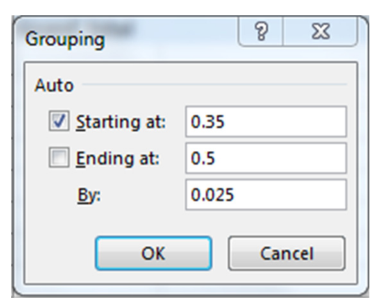
Make sure the Weight variable in the Values box says "Count of Weight."

The table now displays the frequency distribution of the Weight variable. It makes more sense to create classes so groups of similar values can be placed together within a class. Then the frequency distribution will be based on the classes and not on the individual weight values.

5. Select one of the rows in the table by clicking a row label → right click → select Group. The Grouping dialog box appears as shown below:



Excel places the minimum and maximum values of the Weight variable in the Starting at and Ending at text boxes. Excel also places a class size suggestion in the By text box. Excel uses the Starting at and Ending at values to label the smallest and largest classes. Change these values so the upper class extends beyond the range of the data. This is necessary for at least the upper class. The reason for this is because Excel is inconsistent in how it handles the upper class limits. For every class except the last class, the upper value listed in the label is not included in the class. If the Ending at value is larger than the largest data value, the classes are forced to be identical as far as what the upper class value means, i.e., it will not include the upper value. Change the Grouping settings to the following and click OK:



To make the labeling more user friendly without having to manually change each class, replace the '-' with ' up to ' so each class will read 'lower up to upper.' For example, the first class will read '0.35 up to 0.4.'

6. Select any cell outside the table.
7. Press Ctrl-F. The Find and Replace dialog box appears.
8. In the Find what text box, type '-'.
9. Select the Replace tab.
10. In the Replace with text box, type ' up to ' (Don't forget leading and trailing spaces).
11. Click Replace All.

8.2.2 Construct and Format Histograms

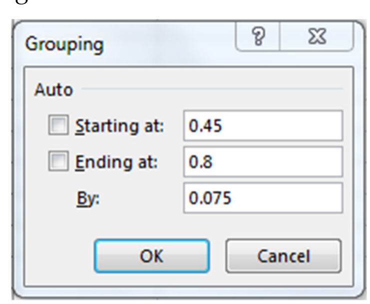
A histogram is a visual representation of the distribution of one quantitative variable.

1. Select one of the cells in the pivot table previously created → click the Insert tab → under Charts, click the black triangle next to the column chart icon → select 2-D Column.
2. Right click one of the columns → select Format Data Series… → move the Gap Width bar to 0%.
3. In the Format Data Series window, click the bucket icon → select the solid line radio button → with the Color option, select a color that stands out from the bar color (dark blue or black) → close the Format Data Series window.
4. In the chart, click the title "Total" to select it → type your own title → press the Enter key.
5. In the Design tab, click Add Chart Element → click Axis Titles → Primary Horizontal. Type "Weight (ounces)" → press the Enter key.
6. In the Design tab, click Add Chart Element → click Axis Titles → Primary Vertical. Type "Frequency" or "Relative Frequency" → press the Enter key.
7. Right click one of the columns → hover over Add Data Labels → select Add Data Labels.
8. Click on the legend to select it → press the Delete key.
9. In the VALUES square, click on the small black triangle next to Count of Weight → select Hide All Field Buttons on Chart. (Chart must be selected)
10. Right click on the vertical axis → select Format Axis → expand Number → type 0 in the Decimal places text box

Change the histogram to display relative frequencies…

11. Click on the black triangle next to Count of Weight in the Values square → select value field settings → select Show Values As → click to open the Show values as drop down box → select % of Column Total
12. Click on the vertical axis title → type "Relative Frequency"

Change the histogram to display the distribution of Peanut M&M weights…



13. Select the filter icon above the frequency distribution table → select Peanut M&M
14. Select one of the rows in the table by clicking a row label → right click and select Group
15. Change the Grouping settings to those displayed in the Grouping dialog box to the right

9 Summary of Statistical Definitions and Formulas with Excel Functions

9.1 Descriptive Statistics

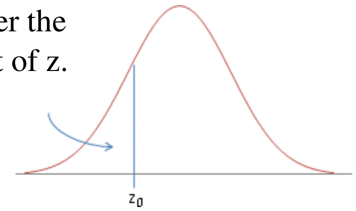
Measure	Excel Formula
Count	=COUNT(<data range>)
Sum	=SUM(<data range>)
Count if <criteria> is true	=COUNTIF(<data range>, “*criteria*”)
Count if <criteria range> is true	=COUNTIFS(<data range>, “>=*criteria*”, <data range>, “<*criteria*”)
Mean	=MEAN(<data range>)
Sample Variance	=VAR.S(<data range>)
Sample Standard Deviation	=STDEV.*S*(<data range>)
Population Variance	=VAR.P(<data range>)
Population Standard Deviation	=STDEV.P(<data range>)
Median	=MEDIAN(<data range>)
Minimum	=MIN(<data range>)
First quartile (Q1, 25th percentile)	=QUARTILE.EXC(<data range>, 1)
Second Quartile (Q2, 50th percentile, median)	=QUARTILE.EXC(<data range>, 2)
Third quartile (Q3, 75th percentile)	=QUARTILE.EXC(<data range>, 3)
Maximum	=MAX(<data range>)
Range	=MAX(<data range>) - MIN(<data range>)
Interquartile Range	=QUARTILE.EXC(<data range>, 3) – QUARTILE.EXC(<data range>, 1)

Measure	Excel Formula
Percentile	=PERCENTILE.EXC(<data range>, *percentile*)
Percent Rank	=PERCENTRANK.EXC(<data range>, *data value*)
Standardize	=STANDARDIZE(x, mean, standard deviation)
Covariance	=COVARIANCE.S(<data range 1>, <data range 2>)
Correlation	=CORREL(<data range 1>, <data range 2>)

9.2 Binomial and Normal Probabilities

Measure	Formula	Excel Formula
Binomial Probability $X \sim Bin(n, p)$	$P(X = k) = nCk \times p^k \times (1 - p)^{n-k}$	=BINOM.DIST(k, n, p, 1) → cumulative =BINOM.DIST(k, n, p, 0) → exact
Standard Normal $X \sim N(0, 1)$	$z = \frac{x - \mu}{\sigma}$	=NORM.S.DIST(z, 1) =NORM.S.INV(probability) → Reverse Lookup
Normal $X \sim N(\mu, \sigma^2)$	$x = z * \sigma + \mu$	=NORM.DIST(x, mean, standard deviation, 1) =NORM.INV(probability, mean, standard deviation) → Reverse Lookup

Table entry for z is the area under the standard normal curve to the left of z.



10 Z Table

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
-3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.6	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

11 Syllabus

11.1 Course Details

Instructor: Lisa Over

Office: 929 Rockwell Hall / 432 College Hall

Email: overl@duq.edu

Office Hours: TR 9:15 – 10:15 in room 929 Rockwell Hall
TR 12:45 – 2:15 in room 432 College Hall

Class: QSIS 285 Business Statistics 26537-03

Class Time: TR 8:00-9:15
Rockwell Hall

Text: *Business Statistics Communicating with Numbers*, 2nd Edition, by Jaggia/Kelly

Full loose leaf book + Connect + Excel Now

ISBN: 9781260309393

Price: $110

Materials: **Calculator*, laptop with Excel**

Bring your calculator to class every day! We will use the laptops often. It will be beneficial for you to have it in class but not required.

*Calculator capable of adding, subtracting, multiplying, dividing, taking a square root, and other algebraic and statistical functions such as raising numbers to specified powers ($x^{\square}$) and combinations (nCr). The Casio fx-300ES PLUS is about $10 and easy to use.

Website: https://duquesne.blackboard.com**

**Blackboard treats un-submitted assignments as if the assignment was not assigned, i.e., Blackboard does not include it in the total grade calculation. I will periodically enter zeros for any assignments not completed. You will notice that this will negatively affect your grade so keep it in mind throughout the semester.

Prerequisites: Algebra

11.2 Course Description

Welcome to QSIS 285: Business Statistics! As a business student, you are both a consumer of statistics and a researcher. Statistics generally involves the collection, description, analysis, and interpretation of data. In business statistics, you select and apply appropriate statistical tools and methods to answer questions, solve problems, and make informed, evidence-based decisions. Course topics include descriptive statistics (visual descriptions and summaries of the data using graphs and tables), inferential statistics (hypothesis testing using the normal, t, F, and chi-square distributions), and regression analysis (simple linear and multiple linear regression). By the end of the course, you will be able to…

- Identify and describe statistical tools, theories, and methods used in business
- Discuss and evaluate the use of statistical tools, theories, and methods to generate ideas, recognize opportunities, solve problems, and make decisions
- Select and implement appropriate statistical tools, theories, and methods given a specific business problem or goal
- Analyze statistical results and use them to support claims, develop business plans, evaluate courses of action, and make recommendations

I'm looking forward to working with you this semester. Please don't hesitate to take advantage of my office hours.

11.3 Course Grades

The grade you earn in this course will be a reflection of the knowledge and skills you choose to master and will be based on the following criteria:

Class Participation (10%): Class participation includes reading assignments with comprehension questions (LearnSmart) and out-of-class online exercises. In class, we will explore examples and explanations beyond what is covered in the textbook. I will periodically assign in-class exercises to prepare you for similar exercises on exams. I will not collect these, but I will post solutions on BB for you to self-assess your understanding. Attendance will not be taken but truly mastering the course requires attending and participating in class regularly, reading the text, reviewing solutions to in-class exercises, and seeking help from a tutor or myself. Collaboration is encouraged in class, however, it is to your advantage to attempt exercises on your own before working with others.

Homework (10%): Homework includes approximately 15 problem sets. Learning statistics involves doing statistics! Truly mastering the course requires conscientiously completing the homework on your own, actively reviewing solutions, and seeking help from a tutor or myself to correct misconceptions. Collaboration is allowed, but in the end, you need to be able to do the problems on your own on assessments. All homework will be assigned from the textbook and must be completed using Connect. You will have access to Connect through Blackboard. Homework is graded for accuracy and due on the specified due date. Work turned in late will incur a small reduction each day after the due date. Twenty percent of your lowest scoring homework assignments (3 out of 15) will be dropped and will not count toward the final grade. *Note:* I will not drop assignments until the end of the semester.

Quizzes (10%): One or two quizzes per chapter will be given throughout the semester. These quizzes will be administered either in class or outside of class. Twenty percent of your lowest scoring quizzes will be dropped and will not count toward the final grade. These quizzes are designed to help you prepare for exams. Truly mastering the course requires conscientiously completing the quizzes on your own, actively reviewing the feedback and solutions, and seeking help from a tutor or myself to correct misconceptions. *Note:* I will not drop assignments until the end of the semester.

Exams (40%): Two closed-book, in-class exams will be given. The exams will not simply be an exercise in memorization; test items will require an ability to apply concepts. Eighty percent of the material for exams will be derived directly from the homework and quizzes. The remaining 20% will be a combination of new problem scenarios and problems derived from examples presented in class. If you have a firm understanding of the material presented in class, you will have little trouble. You may have one 3 x 5 note card or a half sheet of paper for notes to reference during the exam. No make-up exams are permitted without legitimate, written documentation.

Final Exam (30%): The final exam is a closed-book, non-comprehensive exam given in room [Location TBA] on [Date TBA] from [Time TBA] in accordance with the Duquesne University course catalog. Eighty percent of the material for the final will be derived directly from the homework and quizzes. The remaining 20% will be a combination of new problem scenarios and problems derived from examples presented in class. You may have one 3 x 5 note card or a half sheet of paper for notes to reference during the exam. No exceptions will be made regarding the final exam schedule.

Extra Credit. Extra credit may be offered to the class as a whole on homework, quizzes, and/or exams. There will be no individual opportunities for extra credit to compensate for poor performance. Dropped scores are designed to accommodate you for any technical issues or personal circumstances that interfere with your attention to this class.

Grading Scale. End of semester grades will be administered according to the following percentage breakdown.

Final Grade	*Percentage*
A	92 to 100
A-	90 to 91
B+	88 to 89
B	82 to 87
B-	80 to 81
C+	78 to 79
C	70 to 77
D	62 to 69
F	< 62

11.4 Course Policies

Academic Integrity. Although I encourage you to work with your peers on assignments and other course matters, you are still required to submit individual solution sets to Connect. Copying another student’s assignment or having another student complete your work for you is considered cheating and will result in a 0 for that assignment. Any student found talking, regardless of the topic, during an examination will receive an F on the examination. Any student found cheating or assisting others during an examination will receive an F for this course and will be subject to further sanctions. More information regarding the University's Academic Integrity Policy can be found at: http://www.duq.edu/student-conduct/code-of-conduct/academic-integrity.cfm.

Information for Students with Special Needs. Duquesne University is committed to providing all students with equal access to learning. If you have a disability requiring accommodations, you must register with the Office of Freshman Development and Special Student Services in 309 Duquesne Union (412-396-6657) in order to receive reasonable accommodations in this course. Once a disability is officially documented by this office, and with your permission, instructors will receive letters outlining the reasonable accommodations they are required to make. Once I have received this letter, you and I should meet to coordinate the implementation of these accommodations. More information can be found at http://www.duq.edu/special-students/policies.cfm.

If your accommodations include a quiet testing environment, you are responsible for contacting me in advance to make arrangements for you to complete the assessment. If your accommodations include extra time for assessments, I will add extensions to your assessments but you must contact me in advance to make arrangements to start early, to stay late, or to establish a separate time and place for you to complete the assessment.

Calculators. Calculators may be used for homework, quizzes, and exams. Any basic calculator capable of adding, subtracting, multiplying, dividing, and taking a square root is the minimum requirement. A calculator with additional algebraic and statistical functions such as raising numbers to specified powers ($x^{\square}$) and combinations (nCr) will be helpful. You may not use the calculators on your cell phones or PED's (see Academic Integrity Policies above) nor may you share calculators with other students during the course of a quiz or exam.

Cell Phone Policy. As a courtesy to the instructor and other students, all personal electronic devices must be silenced throughout class meetings.

A student who is found using a personal electronic device during class can expect the following sanctions:

First offense - warning.

Second offense - dismissal from class for the remainder of the class period.

The instructor will decide whether or not the student will be allowed to make up any graded work performed during the remainder of the period. If the course has an attendance policy, the class will be counted as an absence.

Beyond the second offense - suspension from class attendance for a length of time to be determined by the instructor, up to and including the remainder of the semester. The instructor will decide the extent to which the student will be allowed to make up any graded work performed during the missed classes. All suspended classes will be counted as absences if the course has an attendance policy.

A student who is found using a personal electronic device during a quiz, exam, or other graded event in class can expect the following sanctions:

First offense a 0 grade on the work in question is the minimum sanction, but anything up to and including failure in the course is possible. The specific sanction will be chosen by the instructor based on factors such as the percentage of the course grade based on the graded event, whether the student has had other Academic Integrity violations, and the like. Per College Academic Integrity policy, the Department Chair and/or Graduate Program Director (in the case of graduate courses) must be consulted before failing a student.

Second offense failure in the course, again after consulting with the Chair or Graduate Director. Any associated reduction in grade will be reported to various entities within the University as required by the College and University Academic Integrity policies.

[This page intentionally left blank]

11.5 Course Schedule

The schedule outlined below is the tentative schedule and may be adjusted as needed as we progress through the course.

Assignments Due by 8:00 a.m. (by class time) on Date Indicated Below						
Date	Day	Chapter/ Sections	Objectives	LearnSmart Class Participation	Connect Homework	Connect Quizzes
11-Jan	R	1.1 - 1.3	1) Syllabus and Introduction: Describe the field of statistics using the terms population, sample, parameter, statistic, sampling, descriptive statistics, and inferential statistics. 2) Distinguish between qualitative and quantitative data, cross-sectional and time series data			
16-Jan	T	2.1 - 2.2	1) Construct and interpret frequency distributions, pie charts, and bar graphs to summarize qualitative data 2) Construct and interpret grouped frequency distributions, cumulative frequency distributions, and histograms to summarize quantitative data			
18-Jan	R	3.1, 3.2	Compute measures of location: mean, median, mode, quartiles, and percentiles	LearnSmart Ch 1, 2, 3		
23-Jan	T	3.4	Compute measures of variability: range, interquartile range, variance, and standard deviation			
25-Jan	R	3.2	Calculate outliers and construct and interpret a boxplot			
30-Jan	T	3.6	1) Convert data values to Z-scores and interpret the relative location 2) Apply the empirical rule to determine the percentage of the data within a specified number of standard deviations from the mean			

Assignments Due by 8:00 a.m. (by class time) on Date Indicated Below						
Date	Day	Chapter/ Sections	Objectives	LearnSmart Class Participation	Connect Homework	Connect Quizzes
1-Feb	R	2.4, 3.8	1) Construct scatterplots 2) Construct and interpret the correlation between two quantitative variables			
6-Feb	T	4.1, 4.2, 4.3	Calculate probabilities: addition rule, complement rule, multiplication rule, and conditional probability	LearnSmart Ch 4		
8-Feb	R	4.5, 5.4	Compute binomial probabilities	LearnSmart Ch 5	Connect Ch 1-3	Quizzes Ch 1-3
13-Feb	T	5.1, 6.1, 6.2	1) Distinguish between discrete and continuous probabilities 2) Compute standard normal and cumulative probabilities	LearnSmart Ch 6		
15-Feb	R	7.1	1) Justify the importance of sampling and how results from samples are used to estimate the unknown population parameters 2) Describe the characteristics of a simple random sample and demonstrate how simple random samples are selected	LearnSmart Ch 7		
20-Feb	T	7.2, 7.3	1) Explain characteristics of the sampling distribution of the sample mean and sample proportion 2) Explain and apply the Central Limit Theorem		Connect Ch 4-6	Quizzes Ch 4-6
22-Feb	R		Exam 1 Chapters 1 - 6			
27-Feb	T	8.1, 8.2	Construct and interpret confidence intervals for the estimate of a population mean	LearnSmart Ch 8		
1-Mar	R	8.3	Construct and interpret confidence intervals for the estimate of a population proportion		Connect Ch 7	Quiz Ch 7
6-Mar	T		SPRING BREAK - NO CLASS			

Assignments Due by 8:00 a.m. (by class time) on Date Indicated Below						
Date	Day	Chapter/ Sections	Objectives	LearnSmart Class Participation	Connect Homework	Connect Quizzes
8-Mar	R		SPRING BREAK - NO CLASS			
13-Mar	T	9.1	Introduction to hypothesis testing: define null and alternative hypotheses and explain the big idea of the p-value and critical value approaches to hypothesis testing	LearnSmart Ch 9		
15-Mar	R	9.2, 9.3	Conduct one- and two-tailed hypothesis tests about a population mean: compute and interpret p-values, calculate critical values and compare with test statistic			
20-Mar	T	9.4	Conduct one- and two-tailed hypothesis tests about a population proportion			
22-Mar	R	10.1	Conduct one- and two-tailed hypothesis tests about the difference between the means of two populations	LearnSmart Ch 10		
27-Mar	T	10.3	Conduct one- and two-tailed hypothesis tests about the difference between the proportions of two populations		Connect Ch 8-9	Quizzes Ch 8-9
29-Mar	R		EASTER BREAK - NO CLASS			
3-Apr	T		Exam 2 Chapters 7 - 9			
5-Apr	R	12.2	Conduct and interpret chi-square tests of independence	LearnSmart Ch 12		

Assignments Due by 8:00 a.m. (by class time) on Date Indicated Below						
Date	Day	Chapter/ Sections	Objectives	LearnSmart Class Participation	Connect Homework	Connect Quizzes
10-Apr	T	14.1 - 14.2	1) Estimate a simple regression equation for a sample using the least squares method 2) Verbally interpret the estimated regression coefficients 3) Make predictions using the estimated regression coefficients	LearnSmart Ch 14		
12-Apr	R	14.3	1) Estimate a multiple regression equation for a sample using the least squares method 2) Verbally interpret the estimated regression coefficients		Connect Ch 10-12	Quizzes Ch 10-12
17-Apr	T	14.4	Determine the goodness of fit of the estimated regression model			
19-Apr	R	15.1	Conduct hypothesis tests for the significance of the slopes: one at a time using t-tests	LearnSmart Ch 15		
24-Apr	T	15.1	Conduct hypothesis tests for the significance of the slopes: jointly using F-tests			
26-Apr	R	17.1	Dummy Variables in Regression	LearnSmart Ch 17		
1-May	T		Reading Day - Extended office hours for review			
2-May	R		Exam Week Starts		Connect Ch 14-17	Quizzes Ch 14-17