Module 2 Lecture - Descriptive Statistics

Introduction to Statistical Methods

Quinton Quagliano, M.S., C.S.P

Department of Educational Psychology

Table of Contents

"I don't mind not knowing. It doesn't scare me." — Richard P. Feynman

1 Overview and Introduction

1.1 Textbook Learning Objectives

- Display data graphically and interpret graphs: stemplots, histograms, and box plots.
- Recognize, describe, and calculate the measures of location of data: quartiles and percentiles.
- Recognize, describe, and calculate the measures of the center of data: mean, median, and mode.
- Recognize, describe, and calculate the measures of the spread of data: variance, standard deviation, and range.

1.2 Instructor Learning Objectives

- Understand the complementary value of visual and numeric descriptions of data
- Be able to compare and contract the use cases for different metrics

1.3 Introduction

- Data, in it's raw form, is very ____________
  - Especially large data sets are likely to be borderline ____________

Discuss: With the following table, try to say something meaningful about this data/explain it

Student	Q1	Q2	Q3	Q4	Q5
Student 1	1	0	1	1	0
Student 2	0	1	0	1	1
Student 3	1	1	1	0	0
Student 4	0	0	1	0	1
Student 5	1	1	0	1	1

- Realistically, we need a way to ____________ our datasets in a way that capture the overarching trends in the values

- This can be done both __________ and via __________ statistics, and often times, both!
- We'll start by talking about descriptive graphs, and then later move on to statistics

- There are several common __________ of graphs that show up in research - and we should be able to readily interpret them
  - Funny enough, sometimes the graphs are the most understandable part of a research paper!

> **Important**
>
> There is usually not only one 'right' way to graph or represent data - it may be advantageous to try multiple methods and see how they compare

- The book is filled with examples and practices to get better at navigating these, please try some of them!

2 Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

2.1 Introduction

- In this first section, we'll look at some graphs used often to show __________ of certain values in the data
  - **Frequency** is how often (or how little) a value shows up in the dataset
  - Most of these initial plots are __________ primarily to noncontinuous, discrete data (or continuous data that is treated as discrete)

> **?** Review: Which of the following is NOT quantitative data?
>
> - (●) A) Hair color
> - ( ) B) Age (in years)
> - ( ) C) Height (in cms)
> - ( ) D) Total test score
>
> Explanation:

2.2 Stem-and-Leaf Plots

- **Stem-and-leaf graphs**, also known as **stemplots**, are technically a ________ method to represent data - but function to visually inspect the distribution of the data.
- Stemplots a valid choice for representing ________, quantitative data for a single variable
  - However it works ________ if the range of values is reasonably restricted, and with the same decimal structure
  - Continuous data ________ work, but discrete data can be somewhat easier
  - E.g., Test scores ranging from 0 - 100 $\longrightarrow$ ________!
  - E.g., Reaction time ranging from 0.50 secs - 420 seconds $\longrightarrow$ ________!
- Stemplots contain a **leaf** which contains the **final significant digit** (in practice this is usually just the ________ digit)
  - Then, the ________ is everything *but*, the digits of the leaf
  - Each number entry in a leaf represents a different value in the data
  - So technically, the number of digits that are leaves in the ________ of points in the data
- Practically, stemplots are ________ showing distribution, skew, and frequency of values
  - Especially common in ________
  - However, I do find they are ________ to interpret for non-statistical audiences
- Example of test scores 0 - 100: 2, 3, 5, 6, 9, 11, 14, 15, 16, 20, 21, 23, 27, 30, 32, 38, 41, 44, 46, 53, 55, 59, 60, 62, 67, 74, 79, 81, 85, 90

Stem	Leaf
0	2 3 5 6 9
1	1 4 5 6
2	0 1 3 7
3	0 2 8
4	1 4 6
5	3 5 9
6	0 2 7
7	4 9
8	1 5
9	0

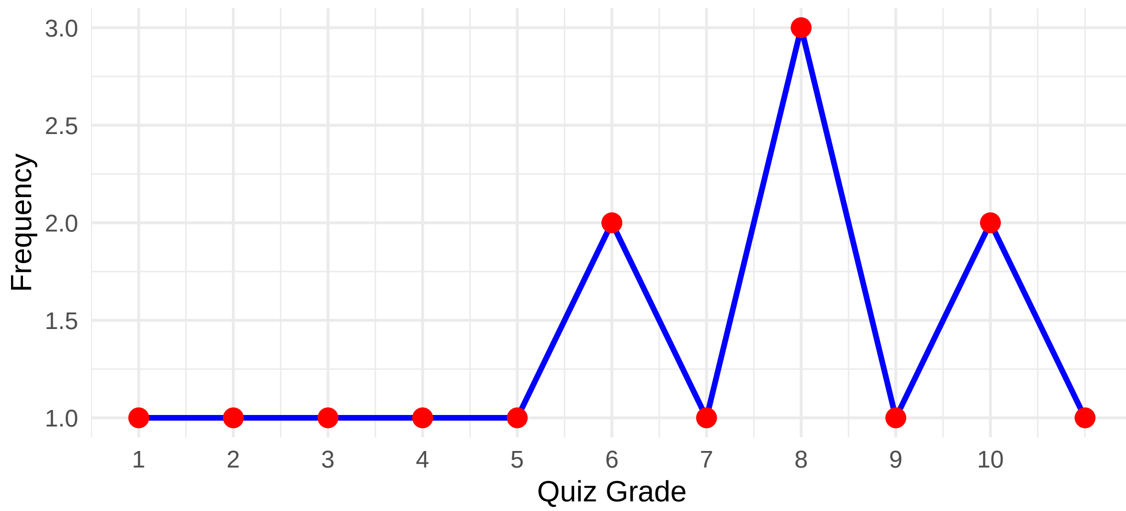
Discuss: Try to explain why a stem-and-leaf-plot would NOT work for qualitative, nominal data

- As you'll see in the following line and bar graphs, one difference is that stemplots don't actually aggregate or summaries the data at al
  - One nice thing about these plots is that they actually leave the numbers somewhat ________, so you can see the entire dataset at a glance

2.3 Line Graphs

- **Line graphs** are a broad family of plots that always have some dotted data points ________ by a continuous line
  - For our purpose, they are another way to visualize ________ of data points
- Unlike stemplots, they could technically be used for qualitative data (but I wouldn't recommend that use)
  - Otherwise, the same rules for ________ apply from the stemplots
- In the line plot:
  - The height, or **y-axis**, represents the ________ of a certain value
  - The place on the **x-axis** represents what value's frequency is ________ by the y-axis

Figure 1: Frequency of Quiz Grades (0-10)



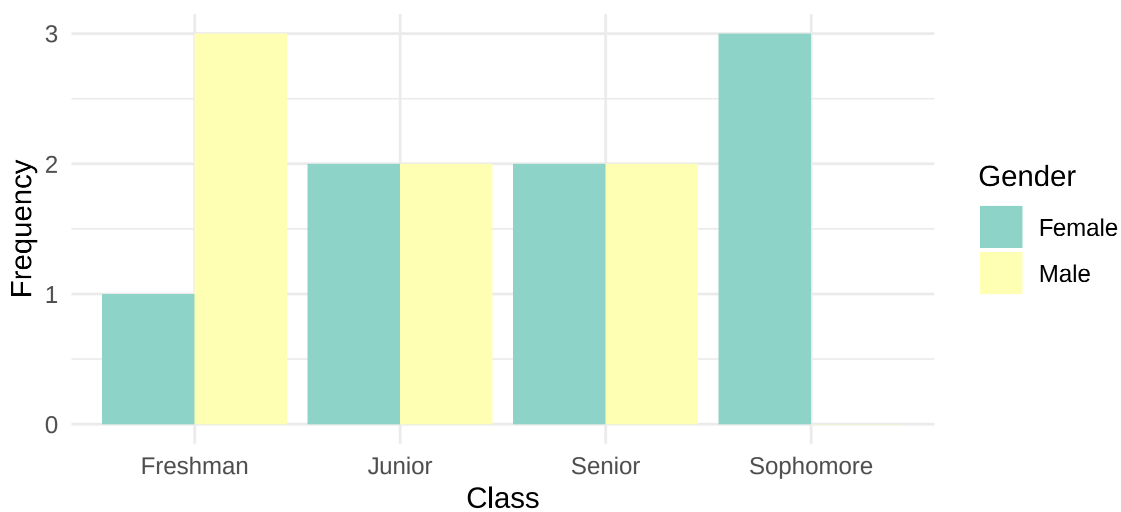
Discuss: Consider the above plot, do you feel there is anything confusing about interpreting it?

2.4 Bar Graphs

- **Bar graphs** show frequency much like stem-and-leaf and __________ graphs, with the advantage of both being easy to interpret (even for non-scientists), while also being friendly to categorical, qualitative data as well

Figure 2: Frequency of Students by Class

Figure 3: Frequency of Students by Class and Sex



Discuss: Do you feel like the above plot could be used for numeric data? Why or why not?

- For both line and bar graphs - you *should* be able to work backwards to reconstruct the __________ data, if necessary

2.5 Aside on Outliers

- An **outlier**, also known as an **extreme value**, is a data point that falls __________ from the others, somehow breaking from the pattern of the data
  - While more commonly applied to numeric data, it could sometimes be used to describe a single member of a qualitative __________

> **Important**
>
> 'Outlier' can be a frustratingly loose term in statistics, because there is not one always accepted answer for just how far a point needs to be removed to be an outlier. Make sure to read carefully to see how any one given author defines it in an individual example or study.

- Many graphical methods naturally __________ outliers in the data, but we'll talk more about statistical/numeric methods for outliers later
  - In plots, you are mainly just looking for visual __________, or breaks in the visual pattern

3 Histograms, Frequency Polygons, and Time Series Graphs

3.1 Introduction

- In talking about frequency, we inherently include talk about __________ of data, or rather, how it is spread out
- While the above plots are fine for more __________ data, there are several plot types the lend especially well to laying out continuous data

> Discuss: Consider the case of an exam that allows for half-points, like 80.5, how will this type of data complicate a stem-and-leaf or line plot?

3.2 Histograms

- A **Histogram**, at first glance, looks much like a __________ plot, as described prior.
- However, rather than use individual discrete points or labels, histograms will __________ values by a defined **interval/class/bin width**, and count the frequencies of values within that bin
  - All the intervals will be the same __________, and we choose that width somewhat arbitrarily
  - But, its worth noting that the bin interval can have a __________ impact on the overarching interpretation
  - See two examples of the same data represented below

> **Important**
>
> Smaller bin size is not always better! Especially in data that is more spread out.

Figure 4: Example of Bad Histogram (Bin width = 15)

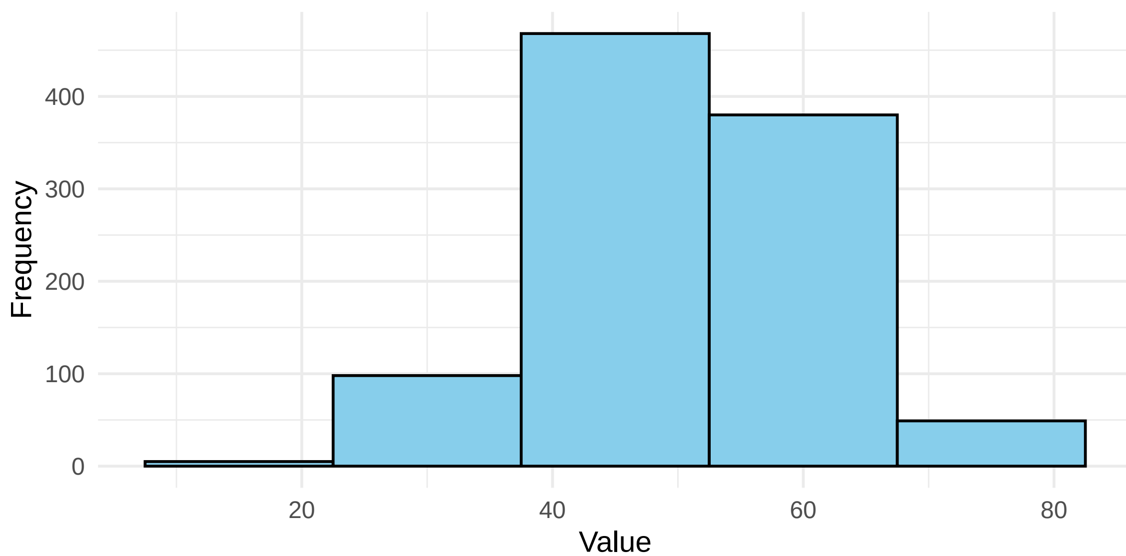
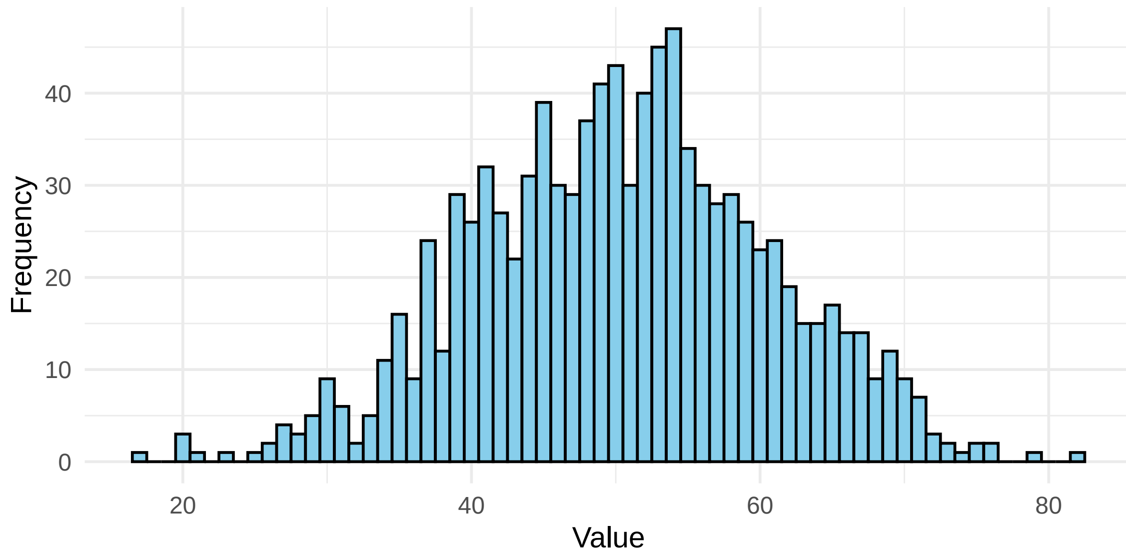


Figure 5: Example of Good Histogram (Bin width = 1)

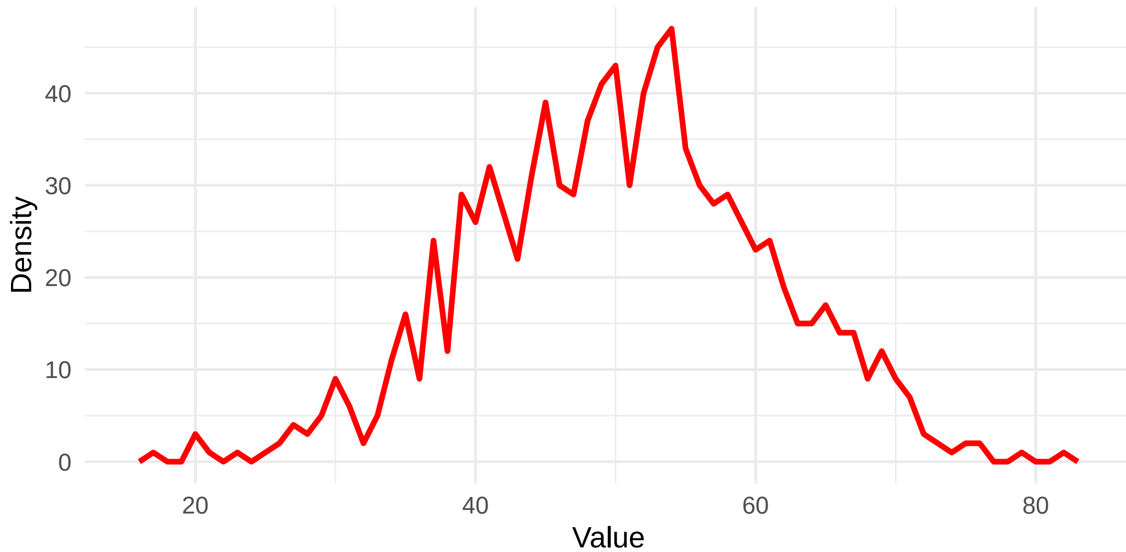


3.3 Frequency Polygons

> **Important**
>
> Frequency polygons look remarkably similar to line graphs, but start from a different perspective on understanding frequency.

- **Frequency polygons** can be though of as a line plot that that travels through the ________________ of histogram bars
  - So frequency polygons still use those same ________________ widths as with histograms, but don't show that on the graph itself (unless overlaid)
  - This makes them more appropriate for dealing with continuous data

Figure 6: Frequency polygon



3.4 Aside about Lying in Statistics

- With all of these different methods (any many more!) for ________ representing data, we run into an issue: accidental or purposeful misleading with graphs
  - *"There are three kinds of lies: lies, damned lies, and statistics"* - Mark Twain (kind of, its complicated who exactly came up with this)
  - While often times taken out of context, this statement does ring true sometimes - we owe it to our readers to be careful

> **Important**
>
> One of the best ways to ensure that data is represented fairly is to try to represent it via multiple different methods - these different methods will help us see slightly different dimensions and information about the distribution.

- Aim for *consistency and clarity in graphing*, avoiding ________ axis points, confusingly similar colors, and making sure enough information is included to understand the scale of the plot.
  - Easy tip: see if you can show the plot to someone not knowledgeable in the work and see if it makes sense to them

4 Measures of the Location of the Data

4.1 Introduction

- Now that we've covered some of the ______ methods to describe data, let us shift focus to the statistic and numerical methods
  - Remember: We don't only use one or the other, most ______ analyses will include both plots and numbers to help readers understand the data

> **Important**
>
> This is where we start introducing some math notation that gets scary - don't panic! Take your time working through the formulas we introduce, don't just mindlessly apply them.

- **Quartiles** are used to cut numerical data into 4 equal-sized ______ when the data is ordered smallest to largest
  - Q_1 (first quartile) is above 1/4 or 25% of the data
  - Q_2 (second quartile) is above 1/2 or 50% of the data
  - Q_3 (third quartile) is above 3/4 or 75% of the data
- **Percentiles** function the same as quartiles, but ______ the data into 100 equal-sized sections
  - For example, the at the 90th percentile, 90% of scores are lower
  - Percentiles are especially commonly used to describe roughly where ______ data points fall - this comes up often in standardized testing
  - The 25th percentile is equivalent to Q_1, 50th percentile is equivalent to Q_2, and the 75th percentile is equivalent to Q_3
- The **median** is the value at the 50th percentile, Q_2, or, put simply, the value that ______ the data into two equal halves when ordered from smallest to largest
  - In the case that there is no actual exact middle value when ordered smallest to largest, we'd ______ the two middle-most values
  - Sample median statistic notation: X_m
  - Population median parameter notoion: M
- The **interquartile range** is the ______ 50% or middle half of the data, given as:

$$Q_3 - Q_1$$

4.2 A Formula for Finding the Kth Percentile

- Often times, we want to know what ______ corresponds to a certain percentile in a dataset
  - We can use the following formula to calculate this

$$i = \frac{k}{100} * (n + 1)$$

- Where:
  - i is the index or rank of the value when ordered smallest to largest
  - k is the k^{th} percentile
  - n is the total number of data points
- If i ends up being between two integers, i.e., a decimal, then we ______ up and down to the nearest ranks and take the average of their integers
  - It's worth mentioning this is just one of ______ procedures for finding the k^{th} percentile…
- For a formula like this, it is easiest to approach it algebraically, just inserting the chosen values over the letters in the formula

> Discuss: Try writing out this same equation, replacing k with 20 and n with 12 - could you solve it from here?

4.2.1 Example

Dataset: 5, 6, 7, 8, 9 (ordered smallest to largest)

Prompt: What value corresponds to the 70th percentile?

$$i = \frac{70}{100} * (5 + 1)$$

$$i = 0.70 * 6$$

$$i = 4.2$$

We round i up and down to the 4th and 5th ranks, which correspond to values 8 and 9 in the dataset, their average is 8.5 $\rightarrow$ 70th percentile

? If I wanted to find what value corresponds to Q_2 what k would I use in this formula?

- (●) A) 10
- ( ) B) 20
- ( ) C) 25
- ( ) D) 50

Explanation:

4.3 A Formula for Finding the Percentile of a Value in a Data Set

- We can go the ______________ direction to figure out what a specific datapoint's percentile is

$$\frac{x + 0.5y}{n} * 100$$

- Where:
  - x is the number of data points up to and NOT including the point of interest
  - y the number of occurrences of the values of interest
  - n is the total number of data points
- Like the prior formula, this is just *one* procedure, you may run into others in the wild

4.3.1 Example

Dataset: 5, 6, 7, 8, 9 (ordered smallest to largest)

Prompt: What percentile is value '8' at?

$$\frac{3 + 0.5(1)}{5} * 100$$

$$\frac{3.5}{5} * 100$$

$$70 \rightarrow percentile$$

- **WAIT!** Didn't we just say that 8.5 is the 70th percentile in the last example?
  - Yes, that is because these formulas are only ________________, and work poorly in small datasets
  - There are several slight variations on calculating these, but the formulas above is what we will use for this class

4.4 Interpreting Percentiles, Quartiles, and Median

- As a simple check, always ensure your values are ________________ smallest to largest when calculating percentiles, quartiles, and the median
- Smaller percentiles/quartiles $\rightarrow$ ________________ values in the data set
- Q_2 = 50th percentile = median

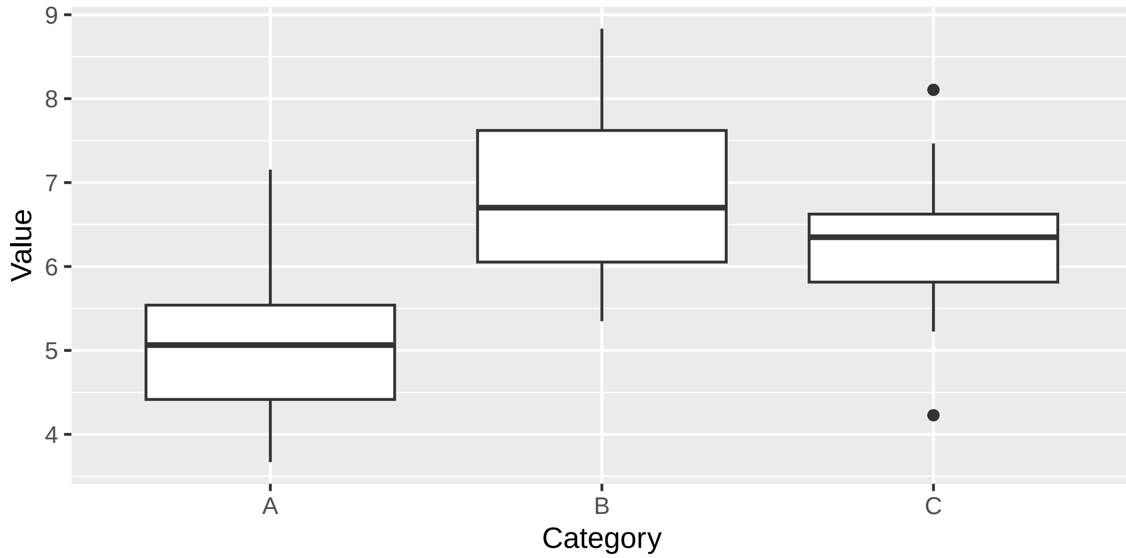
Discuss: If quartiles split the data into 4 equal sections, how many sections are in quantiles

5 Box Plots

5.1 Introduction

- Going to take a quick trip back to ________________
  - **Boxplots** are useful for representing information about the quartiles, percentiles, and ________________ all in a single plot
  - Also called **box-and-whisker** plots (may be the preferred name for the cat-lovers like myself)
- The centerline of a box plot represents a median, edges of the box represent Q_2 and Q_3 (i.e., the box is the IQR), the whiskers *usually* extend to the farthest values

Figure 7: Boxplot example



6 Measures of the Center of the Data

6.1 Introduction

- There are several ways for us to represent the ________________ of the dataset, often collectively referred to as **measures of central tendency**.
  - We already discussed the median, but also have **mean** and the **mode** of the data
  - ________________ is often what we refer to when we say **average**, and is taken by the sum of all the numbers in the data divided by the number of data points
  - ________________ is the most frequently occurring value in the data

> **Important**
>
> We are going to introduce several different notations for statistics and parameters - if you forgot the difference between those two terms, go review module 1!

- Means
  - Our sample mean statistic will be represented as $\bar{x}$ (pronounced x bar)
  - Our population mean parameter will be represented as μ (pronounced mew)
  - Recall that $\bar{x}$ is meant as an estimate of μ!
  - This same notion can be applied to any sample statistic and population parameter.
  - The sample mean is calculated with:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$$

- Modes
  - There is not a very useful formula for finding the mode, but realistically all you need to do is a bar plot and find the ________ bar (for the highest frequency)

6.2 The Law of Large Numbers and the Mean

- As a general rule, as the sample size (n) grows, $\bar{x}$ and μ converge - this is true of most statistic - parameter relationships
  - This is related to the central limit theorem, to be discussed slightly later
  - However, this also explains why bigger samples are often treated as ________ in a lot of scientific research - because the sample statistics better approximate the parameters

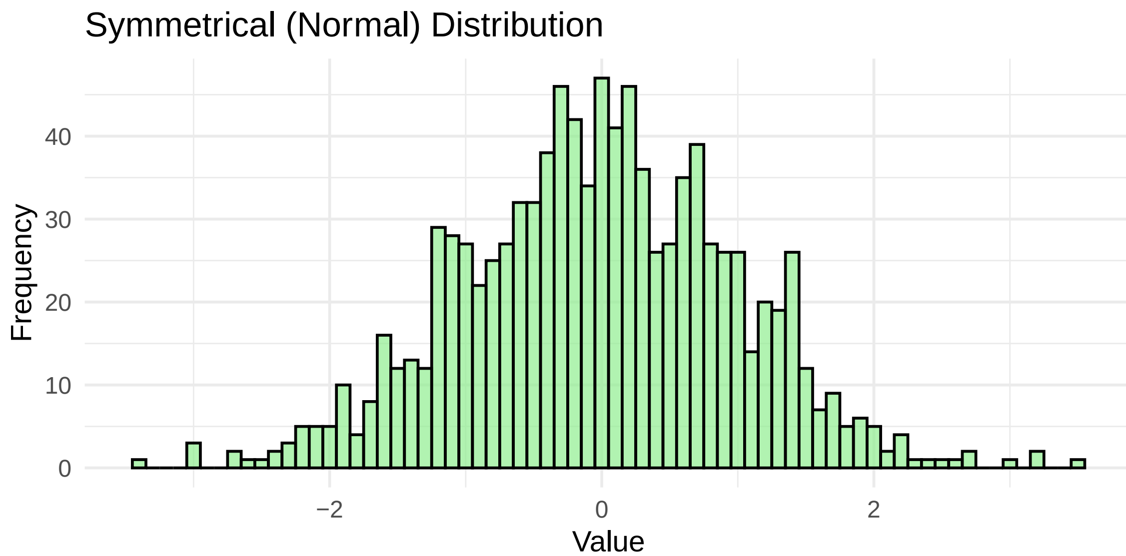
Discuss: Review: Try explaining, from Module 1, why a more representative statistic is a good thing for our research?

7 Skewness and the Mean, Median, and Mode

7.1 Introduction

- In a perfectly ________ distribution, the mean, median, and mode will all be equal or close to one another

Figure 8: Symmetrical histogram



- But, often our data will be ____________ in one direction or another
  - **Left skew** is when most of our values are ____________ to the right, with a tail extending to the left
  - **Right skew** is when most of our values are grouped to the left, with a tail extending to the right

> **Important**
>
> It is easy to get tripped up on describing the direction of skew; they key thing to remember is that the direction of the skew is the direction of the tail!

Figure 9: Left-skewed histogram

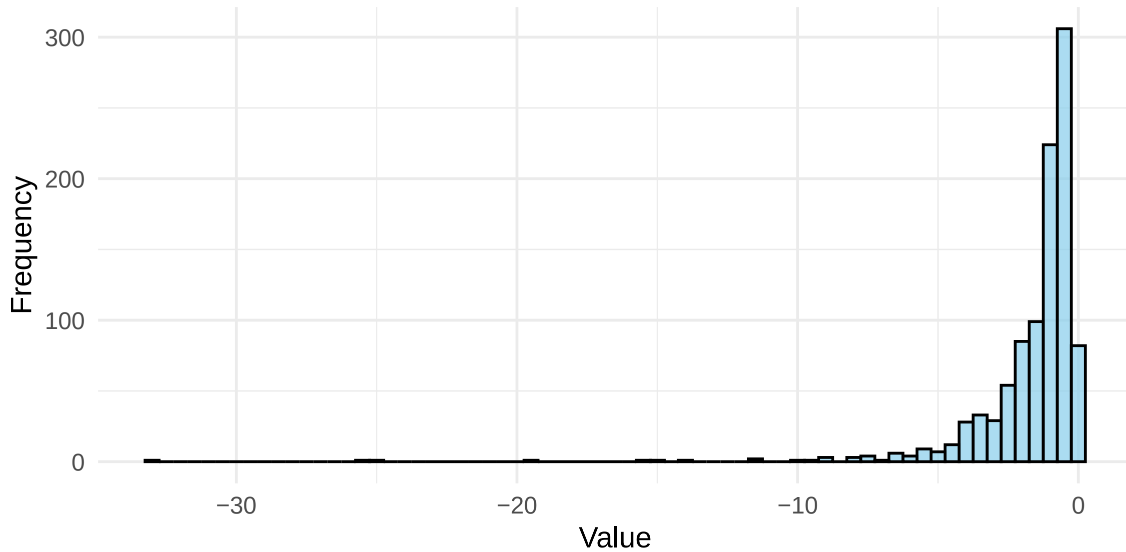
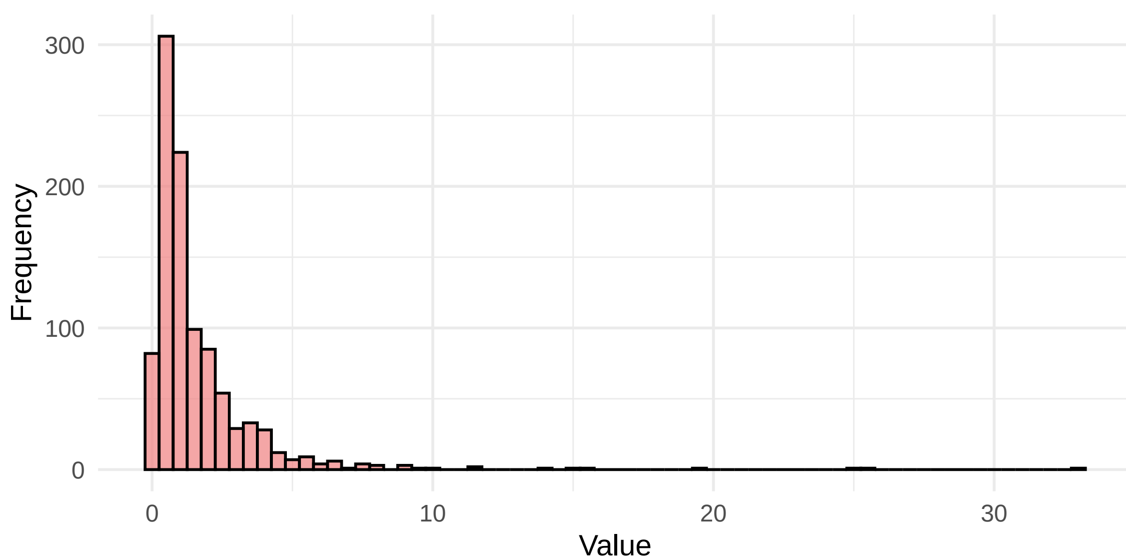


Figure 10: Right-skewed histogram



> **Important**
>
> We will revisit skew later when talking about assumptions for inferential tests, as it can cause issues there

8 Measures of the Spread of the Data

8.1 Introduction

- As a compliment to the measures of ______ tendency above, we have several **measures of dispersion**, or numbers to describe how the data is spread out
  - These are ______ important as the mean, median, and mode, and help further describe the data thoroughly

> **Important**
>
> There are more measures of dispersion than what we will discuss in the class - for various reasons they are not as popular as standard deviation and less often used in research and equations

8.2 The Standard Deviation

- Each number in a dataset has a **deviation**, or how ______ away it is from the mean
  - This is calculated simply as the ______ between the value and the sample mean of the data
  - $x - \bar{x}$ or
  - $x_i - \bar{x}$
- By far, the most important and most used way to describe the ______ of all the data is the **standard deviation**
  - Sample standard deviation statistic: s
  - Population standard deviation parameter: σ
- To calculate the standard deviation, we first will calculate the **variation**, which is just the ______ standard deviation
  - Sample variation statistic: s^2
  - Population variation parameter: σ^2
  - The variance can be described as the **average of the squares of deviations** (what a mouthful!)
  - The variation (for sample) formula is:

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

- Where:
  - x is any given single value from the sample data
  - $\bar{x}$ is the mean of the sample
  - n is the number of data points in the sample data

  - $\sum$ is the summation sign, saying we will add together whatever is in the parentheses
- Let's break it down:
  - $(x - \bar{x}) \rightarrow$ The deviations
  - $(...)^2 \rightarrow$ The squares of the deviations
  - $\frac{\sum ...}{n-1}$ The average of the squares of deviations $\rightarrow$ or variance!
- The standard deviation (for sample) formula then is:

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

- Only difference here is that we are taking the square root of the whole thing, really we can just leave it as:

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

Discuss: Here, I only show the sample-version of the formulas, whereas the book also shows the population-versions, why set our focus this way?

8.3 Calculating Standard Deviation by Hand

While not necessary with computers and calculators, it can be useful to work out statistics "by hand" for learning how they work. For example:

Dataset: 1, 2, 3, 4, 5

Size of sample: $n = 5$

Sample Mean:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$$

$$3 = \frac{1 + 2 + 3 + 4 + 5}{5}$$

Sample Variation:

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

Hint: whenever you see a $\sum$ sign, follow this tabular procedure

x	x - xbar	(x-xbar)^2
1	1 - 3	-2^2
2	2 - 3	-1^2
3	3 - 3	0^2
4	4 - 3	1^2
5	5 - 3	2^2
	Sum	10

$$2.5 = \frac{10}{5 - 1}$$

Sample Standard Deviation:

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

$$1.58 = \sqrt{2.5}$$

8.4 Describing Points and Values with Standard Deviations

- Much like describing a specific value with a percentile, we may want to describe how far away from the ________________ a certain point is, and we can do so using the standard deviation
- We can do this with what are called z-scores with the following formula for sample:

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

- Where: x is a individual data value of interest s is the standard deviation

By hand (pulling from last example):

$$-0.63 = \frac{2 - 3}{1.58}$$

We could then say that the value of 2 is -0.63 standard deviations away from the mean of 3

Discuss: Try working out the z-score of 5 given that same dataset

- Z-scores are often used in conjunction with ______ a topic we'll revisit when we describe the normal distribution in module 6

8.5 Sampling Variability of Statistic

- As mentioned prior, different ______ from the same population of interest will not be exactly the same
  - Thus, their data, means, standard deviations will all be somewhat different
  - If we were to take many ______ samples, we would see slight variation across all of them
  - Put together, they are all just different ______ of their parameters
- We may want to know just how much (on average) our statistics ______ from sample to sample, which is calculated as **standard error**
  - We'll revisit this later, but just an good time to introduce the idea now

9 Conclusion

9.1 Recap

- Understanding our data first starts with describing it, and we can accomplish that both through informative graphs and statistics
- The various graphs show slightly different information, and multiple options may be used simultaneously to more thoroughly show the characteristics of the data
- Measures of central tendency and dispersion can succinctly describe the center of data, and how spread out it is, respectively
- The statistics we calculate on our sample are meant to accurately estimate the parameters of the population distribution, but that only works if our sample is representative and our data unbiased!

- Outliers and discussion on skew will return later when we are talking about their impact on inferential statistics, but don't worry too much about that yet!

9.2 Lecture Check-in

- Make sure to complete and submit the lecture check-in