BALL STATE
UNIVERSITY

Module 1 Lecture - Review of Scale of Measurement, Research Design, and Descriptive Statistics

Analysis of Variance

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

Department of Educational Psychology

Table of Contents

1 Overview and Introduction

1.1 Instructor Learning Objectives

- This material is to help refresh your knowledge of foundational statistics concepts, ideas, and descriptives, prior to more advanced topics to be covered in this course; this will be especially useful if you have not recently taken a statistics class
- Students will be able:
  - Appreciate the relationship between samples and populations
  - Correctly identify scales of measurement for variable
  - Appropriately use and calculate common descriptive statistics and plots
  - Understand the structure of hypothesis testing with statistics

1.2 Introduction and Overview

- Analysis of variance, or _____ for short, is the core focus of this class, but is a complex technique that benefits from a good understanding of foundational statistics
- Prior to us learning about ANOVA, we'll briefly recap the things you've likely already learned in a previous stats class, i.e., EDPS-641
- Though this is likely review, I still encourage you still engage and refresh yourself on these ideas now, so you don't get lost later on!
  - This is going to crunch the content of an intro to stats class into a single lecture

2 Sampling, Statistics, and Parameters

2.1 Concept of Sampling

- When we do research, the group we desire to _____ and understand is our population of interest
  - E.g., *All* high school students, *all* teachers
  - The population is usually a group we can't practically, _____ study (as they are usually too large or dispersed!)
  - In the case that we did somehow gather data about _____ members of a population, we would call this a census

Discuss: Given this definition of a census, I just provided, explain why would we call the periodic counting of all individuals in the US a 'census'

- Instead of gather data on the population as a whole, we take a _____ subset of individuals that are meant to represent the population, which we would call a sample
  - We get our sample via some method of sampling, which is exactly how we get our subset - it can be done in a "good" or "bad" way, resulting in a representative or non-representative sample, respectively

2.2 Statistics and Parameters

- The numeric values that we use to calculate and describe on our _____ is called a statistic, which, in turn, is meant to represent the parameter of the population
  - Another way to say this is that the sample statistic is an _____ of the population parameter
  - A mnemonic to remember this:
    - **P**oulation $\rightarrow$ Parameter
    - **S**ample $\rightarrow$ Statistic

Discuss: I am interested in the average numeric exam score of PSY-101 students, and I take 10 individuals out of the 300 total to represent all the PSY-101 students - what is the sample and what is the population in this example? I take the mean average of the 10 individuals - is this mean a statistics or a parameter, and why?

- But, remember our goal is often to study the _____, not just the subset we gather data for!

  - Thus, we want to have great _____ that whatever statistics we have, _____ represent their respective parameters
  - One part of ensuring this accuracy is to use a sampling method that results in a representative sample

> **Important**
>
> Don't lose sight of this - our sample is just a practical way for us to try and say something about our population, which is what we are really after.

3 Scales of Measurement and Describing Variables

3.1 Introduction

- The data we use for analysis will be organized into _____ and _____, with values for each individual in our sample
  - It's critical we can accurately describe variables, especially when it comes to applying _____ statistics and plots
- Variables are some defined measure/observation with variance in the data
  - I.e., there needs to be _____ numbers in the data, the characteristic can take different values - otherwise it is a statistical constant
  - _____ can also be either numeric or categorical, that is, they produce data of a specified type
  - E.g., Age in years $\rightarrow$ _____
  - E.g., Job title $\rightarrow$ _____

? I gather information about all the individuals enrolled at a local college, and all live in the state of Indiana. Is state of residence a constant or variable, and is it categorical or numeric?

- (•) A) Categorical Constant
- ( ) B) Categorical Variable
- ( ) C) Numeric Constant
- ( ) D) Numeric Variable

Explanation:

- Variables have some distribution
  - A description of how values of a variable are spread out and distributed across the range of possible values
  - We'll cover ways to represent these distributions well in the section: Descriptive Statistics and Plots

3.2 Quantitative vs Qualitative

- Let's dive more into the different _____ of data
  - These are closely related to the above "categorical" and "numeric" terms
- qualitative data come from a more descriptive (via words or _____)
  - E.g., Eye/hair color, more complicated: description of internal feelings
  - Because of it's nature, it is almost always _____ in nature
  - Qualitative data is often times represented in _____ of occurrences of a certain description, for the purpose of analysis - e.g., I have 10 people in my sample with brown eyes, and 5 with blue eyes
- quantitative data is something represent by an _____ or numeric measurement
  - However, within quantitative data, it can be discrete or continuous
  - Discrete data is that which is _____, or has no intervals between the integers, e.g., number of phone calls had $\rightarrow$ I can't have half a phone call
  - Continuous data does indeed have _____ between integers, e.g., Age $\rightarrow$ I can be half a year of age.
  - For better or for worse, these two types are often treated _____, even when that may not be accurate

? Consider this variable: Numbers of children a parent has. What could this be classified as?

- (●) A) Qualitative
- ( ) B) Discrete Quantitative
- ( ) C) Continuous Quantitative
- ( ) D) None of the above

Explanation:

3.3 Nominal, Ordinal, Interval, Ratio

- Nominal scale variables are qualitative and categorical, with classifying _____ and no defined "order".
  - E.g., state/country of residence, hair/eye/skin color
- Ordinal scale variables are those that have an order, but there is not a clear _____ between each interval or place in the order
  - Thus, it sort of blurs the line between categorical and _____
  - E.g., place in a foot race, class rank
- Interval scale variables do have a clear, consistent interval in-between each integer, but no absolute zero point
  - E.g., temperature, IQ scores, ranges, etc.
- Ratio scale is the same as interval scale, but does have a clear zero point as well.
  - E.g., score on an exam
- Generally, the different levels/scales of measurement are different levels of _____ for analysis, in this order: Nominal (most restrictive) to Ratio (least restrictive).
  - Of course, having mostly nominal scale data does not always spell doom, but it can involve more tricky _____
  - We'll discuss in this class how to account for more tricky data types, e.g., _____ in ANOVA analyses, during this class

Discuss: I put all the students in my class from shortest to tallest and assign them the number they are from being the shortest in the class, what scale of measurement would this data be and why?

3.4 Recap of Scale of Measurement

- The nature and scale of _____ of variable is limiting to what descriptive statistics, plots, and inferential statistics we can apply
  - There is significant overlap between these descriptors - usually Nominal, Ordinal, Interval, Ratio are sufficient ways to describe a variable for determining what analyses can be done
  - However, you'll see the other descriptors come up, so you should be aware

📢 Discuss: Think of a person's annual salary in dollar; look back at all the vocabulary in this last section and list all the different ways you can describe this data

4 Descriptive Statistics and Plots

4.1 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!

> **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

4.2 Measures of Central Tendency

- Measures of central tendency of a _____ of scores are statistics that _____ scores that are typical, central or average in the distribution.
  - The most _____ measures of central tendency in the social sciences are the mode, the median and the mean.
- The mode is the most commonly _____ value in a variable/-dataset
  - Mode is often just spelled out, and not given a special notation
  - 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)
- The median is the value at the 50th percentile, second quartile (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 _____ notation: X_m
  - Population median _____ notation: M
  - The best strategy for finding median is to order the values and then cross values off from both ends until you end up with one uncrossed value
  - To find the *location* of the median, one can do: $\frac{n+1}{2}$

- The mean is the arithmetic average of the data
  - Sample mean _____ notation: $\bar{x}$ (x-bar)
  - Population mean _____ notation: μ (mu)
  - The sample mean is calculated with:

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

Important

Something worth noting is that the median is less prone to be affected by outliers than the mean, making it useful in certain circumstances.

? I have data 1,2,2,2,3,4,5. Datapoint 2 is the most commonly occurring value, thus it is the [BLANK]

- (•) A) Mode
- ( ) B) Median
- ( ) C) Mean
- ( ) D) None of the above

Explanation:

4.3 Measures of Dispersion

- Measures of dispersion are statistics that summarize how data is _____ out across a distribution
  - The most _____ measures of dispersion we should be concerned with are standard deviation, variance, range, and interquartile range.
- 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
  - $x_i - \bar{x}$

- To calculate the standard deviation, we first will calculate the variance, which is just the _____ standard deviation
  - Sample variation statistic: s^2
  - Population variation parameter: σ^2
  - The variation sample statistic and population parameter formulas are, respectively:

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

$$\sigma^2 = \frac{\sum (x - \mu)^2}{n}$$

> **Important**
>
> The reason we must use variance to calculate standard deviation relates to the need to square in order to avoid getting 0 from summing the deviations

- By far, the most important and most used way to describe the _____ of all the data is the standard deviation
  - A measure of the average _____ away from the mean that a point is in a distribution
  - It is used instead of variance because it removes the "square" from interpretation
  - Sample standard deviation statistic: s
  - Population standard deviation parameter: σ (sigma)
  - The standard deviation sample statistic and population parameter formulas are, respectively:

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

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{n}}$$

> **Important**
>
> Realistically, we tend to only directly use the sample statistics calculations

Discuss: Look back at the previous equations; which of the measures of central tendency sees the greatest use in these formulas?

- The range is the _____ between the largest and smallest values in the data
- The interquartile range (IQR) is the 75th percentile (upper quartile, Q_3) minus the 25th percentile (lower quartile, Q_1). It is the width of the interval that contains the _____ 50% of the data.
  - In formula it could be represented as: $Q_3 - Q_1 = IQR$

4.4 Percentiles and Quartiles

- 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

4.5 Descriptions of Distributions

Discuss: Quickly list as much as you can remember about the characteristics of a 'normal distribution'

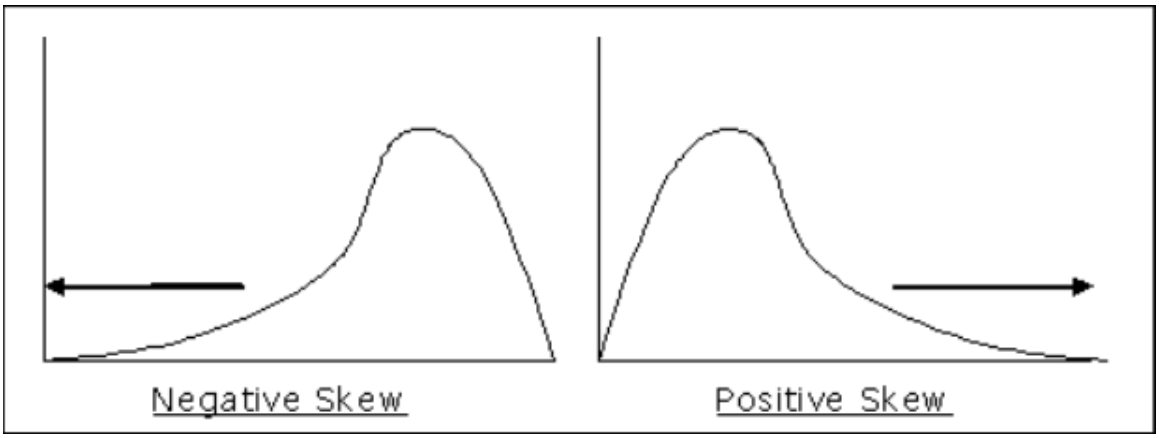
- There are several _____ that are not well described as either Measures of Dispersion or Measures of Central Tendency, but still helpful in understanding the layout of a continuous distribution
- Usually, we use The Normal Distribution as an "ideal", and describe variables' distribution in how it _____ to the normal distribution
  - This is how we will describe Skewness, Kurtosis, and Modality

4.5.1 Skewness

- Skewness is a description of how a distribution departs from _____ or has a 'tail'
  - A positive/right skew: when a distribution has most values clustered towards the _____ end, and a tail out to the right side of the frequency histogram
  - A negative/left skew: when a distribution has most values clustered towards the _____ end, and a tail out to the left side of the frequency histogram

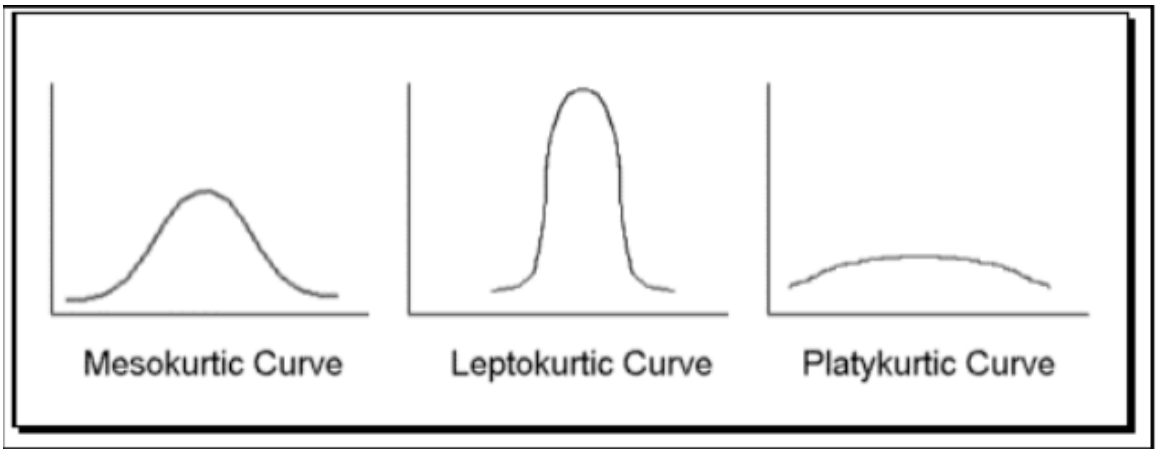
Important

As a helpful mnemonic, the direction of the skew is the direction of the 'tail', not the 'hump'



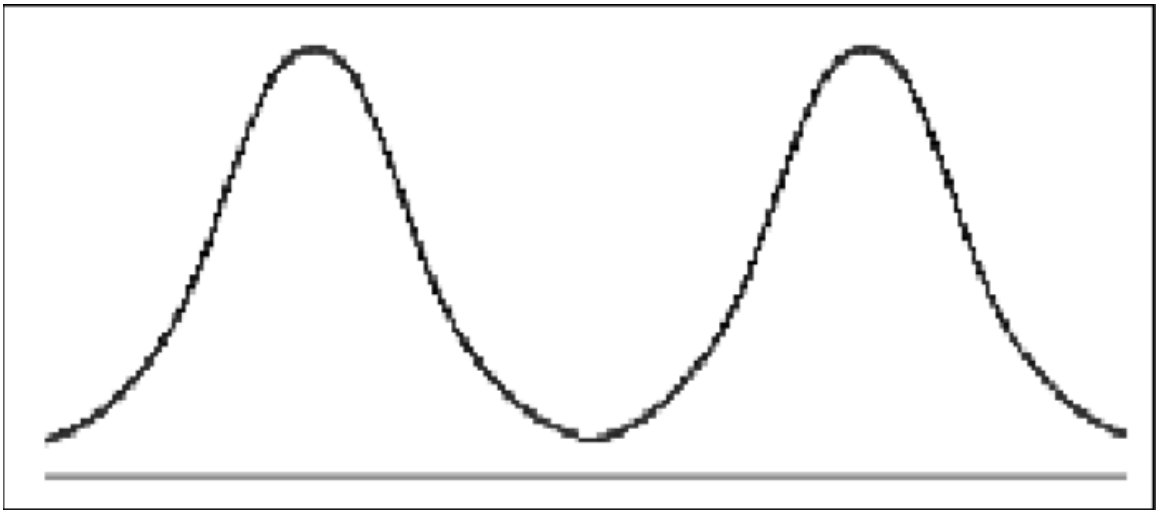
4.5.2 Kurtosis

- Kurtosis is a description of how 'peaked' a distribution is
  - A leptokurtic distribution has too many scores concentrated in the center and not _____ in the tails. (peaked)
  - A platykurtic distribution has too few scores in the center and too many in the _____. (flat)
  - The Normal Distribution is said to be _____



4.5.3 Modality

- Modality a description of how many _____ a distribution has
  - A distribution is unimodal when is has only one peak, and bimodal when it has two peaks

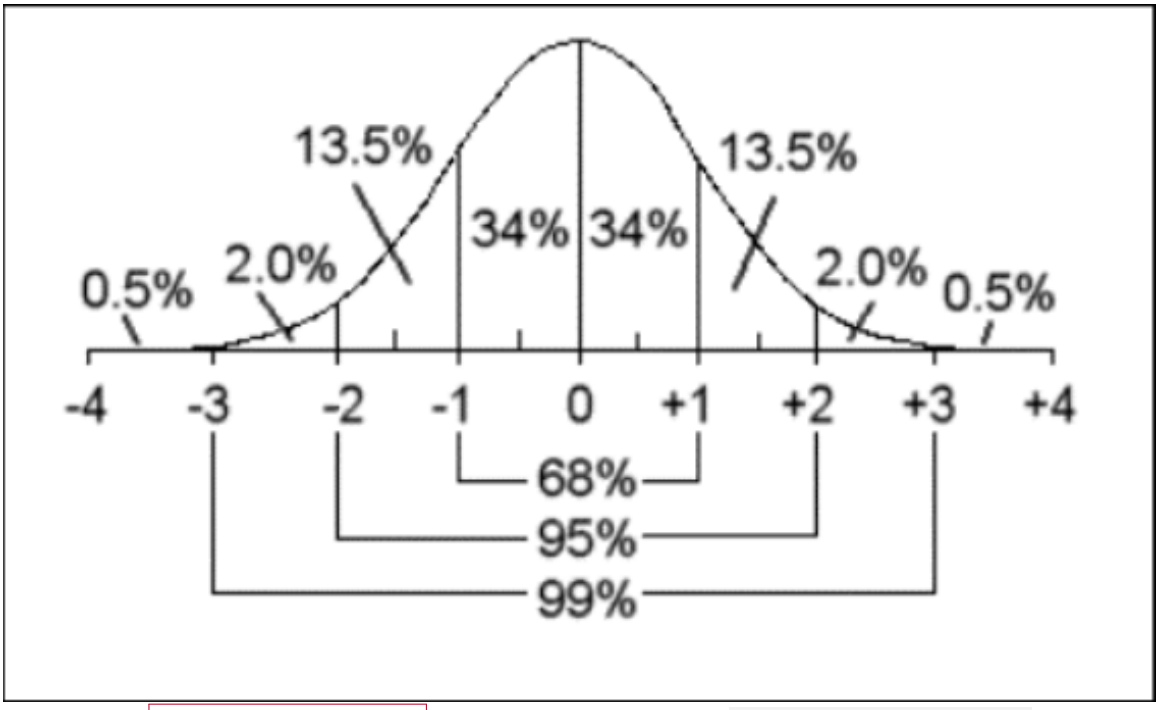


Discuss: Take a guess at what a distribution with three peaks would be called

4.5.4 The Normal Distribution

Important

The normal distribution, is an 'ideal' and used theoretically. In practice, collected raw data will never be perfectly normal. There are also other ideal distributions such as the exponential and uniform distributions, that are also useful in certain circumstances.



- The normal distribution is a distribution of a _____ variable that is unimodal, symmetrical, mesokurtic, bell-shaped, and the mean, median, and mode are equal
  - Follows the empirical rule

Discuss: Reading the distribution chart above, at what number of standard deviations are 84 percent of values beneath it?

4.6 Descriptive Plots

- There are *many* ways to _____ represent data in a coherent way - we'll cover just a few: Frequency Histograms, Bar Graphs, and Boxplots

4.6.1 Bar Graphs

- Bar graphs help show _____ of values of categorical/nominal variables

Figure 1: Frequency of Students by Class

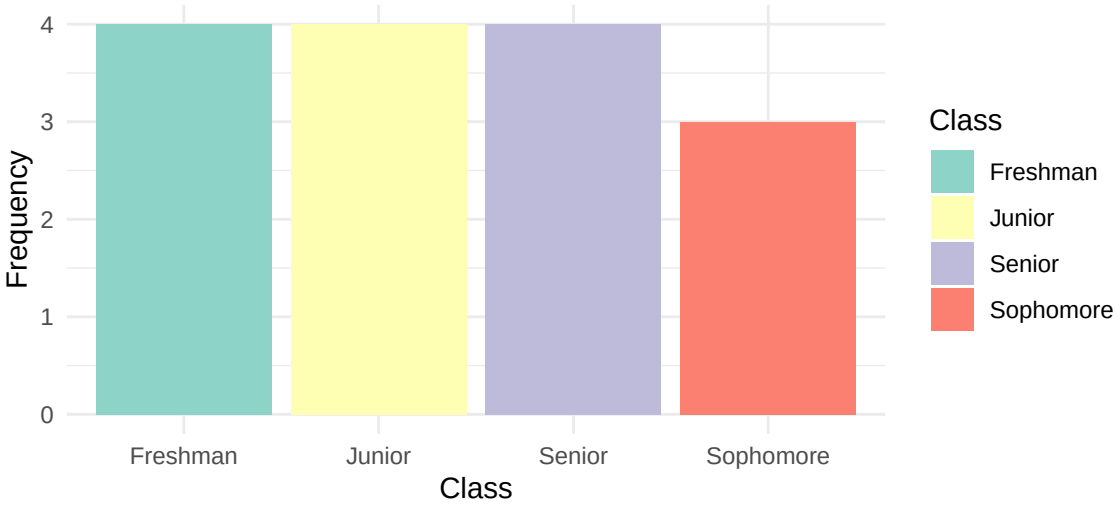
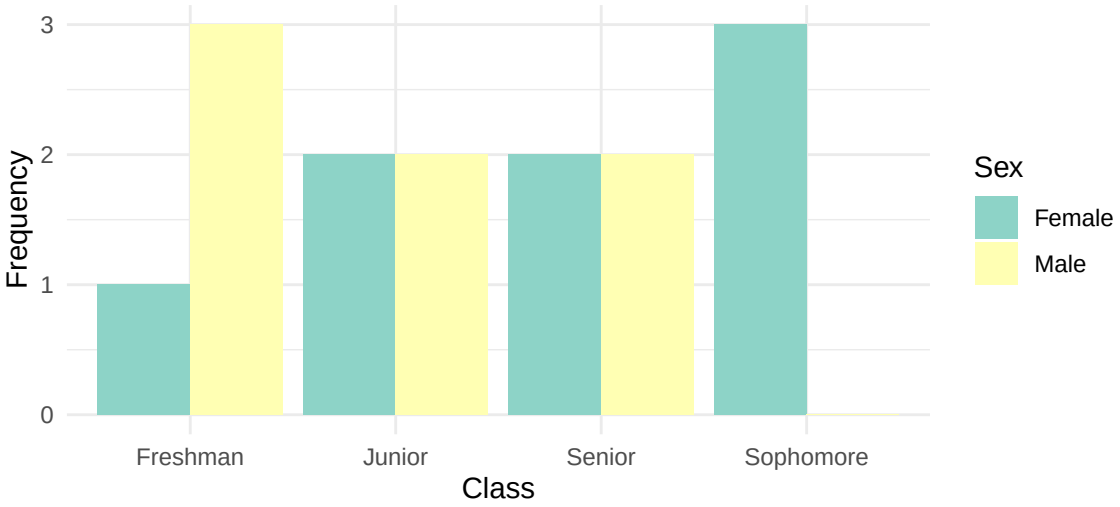


Figure 2: Frequency of Students by Class and Sex



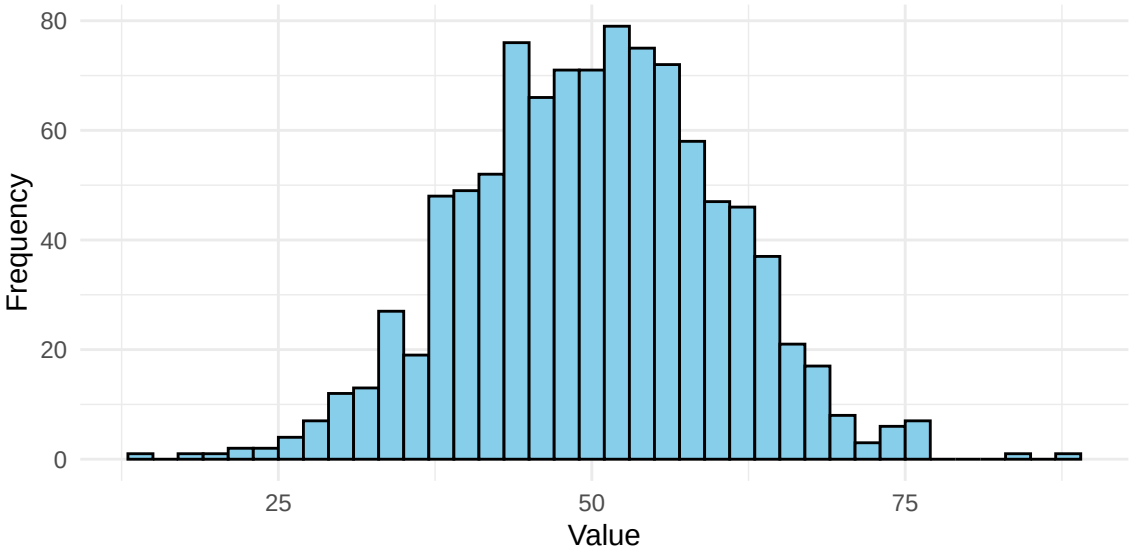
📢 Discuss: Come up with 2 additional examples of variable that could be well-represented with a bar chart

4.6.2 Frequency Histograms

- A frequency 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 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

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

Figure 3: Example of a Histogram (Bin width = 2)

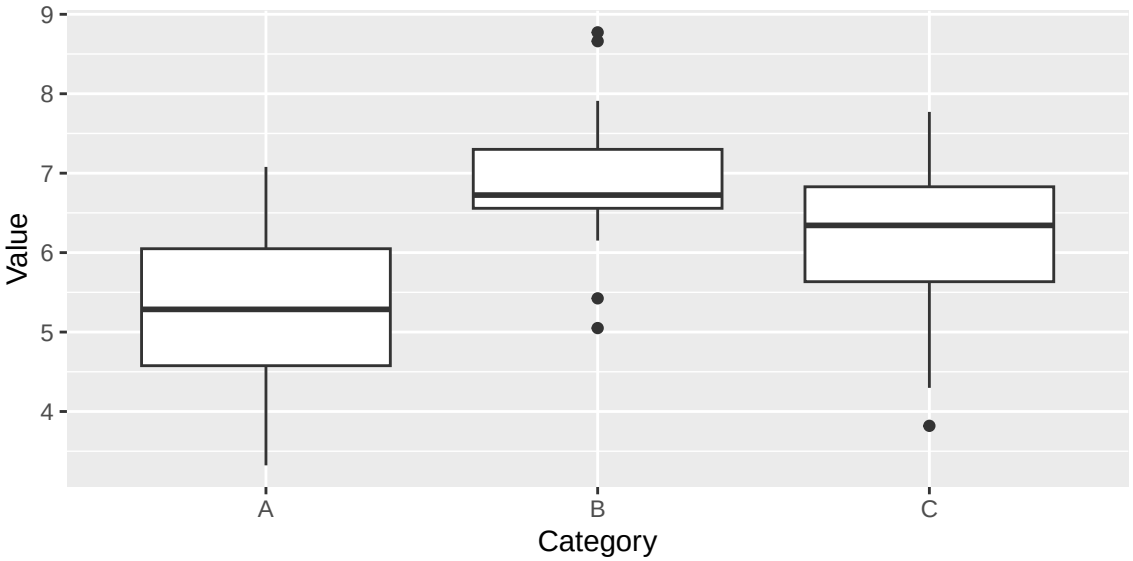


4.6.3 Boxplots

- 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 _____ of a box plot represents a median, edges of the box represent Q_1 and Q_3 (i.e., the box is the IQR), the whiskers *usually* extend to the farthest values

Figure 4: Boxplot example



4.7 Practicing Choosing Descriptives and Plots

- Not all plots and descriptive statistics provide the _____ information - try the following questions for practice

? I want to graphically examine the median and IQR of SAT scores in a local school district, how could I do this?

- (•) A) Boxplot
- ( ) B) Frequency histogram
- ( ) C) Kurtosis
- ( ) D) Arithmetic Mean

Explanation:

? I want to know the numeric value most commonly appearing in a dataset, how could I find this?

- (•) A) Mean
- ( ) B) Median
- ( ) C) Boxplot
- ( ) D) Mode

Explanation:

? I want visually compare my variable of student wellbeing ratings (continuous) against a normal distribution, how could I best examine this?

- (•) A) Boxplot
- ( ) B) Bar plot
- ( ) C) Histogram
- ( ) D) Mean

Explanation:

4.8 Descriptive Tables and Frequencies

- Simple frequency is a count of the _____ of cases – or frequency - that fall in the different _____ or that obtain certain scores.
  - E.g., - the number of girls and boys in a sample the number of students who identify themselves as African-American, Asian- American, Hispanic, or White on a questionnaire; the number of people who obtain different scores on a test
- Relative frequency is the _____ or percentage of cases for each score in the distribution
- Cumulative frequency is a count of the number of cases that fall _____

a certain score. A cumulative frequency is provided for all scores in the distribution

- Cumulative relative frequency is the _____ or percentage of cases that fall below a certain score.

Score	Frequency	Cumulative Frequency	Percent	Valid Percent	Cumulative Percent
0	1	1	2.2	2.2	2.2
2	1	2	2.2	2.2	4.3
3	3	5	6.5	6.5	10.9
4	2	7	4.3	4.3	15.2
5	4	11	8.7	8.7	23.9
6	4	15	8.7	8.7	32.6
7	6	21	13.0	13.0	45.7
8	3	24	6.5	6.5	52.2
9	7	31	15.2	15.2	67.4
10	7	38	15.2	15.2	82.6
11	3	41	6.5	6.5	89.1
12	2	43	4.3	4.3	93.5
13	2	45	4.3	4.3	97.8
15	1	46	2.2	2.2	100.0
Total	46		100.0	100.0	

Discuss: Based on your reading of the above table, what is the cumulative frequency at score '11'?

5 Hypothesis Testing

5.1 Introduction

- In statistics, hypothesis testing is the process by which we _____ whether data supports making a certain conclusion beyond a _____ doubt
- We have certain _____ to work through a hypothesis test:
  - Set up the _____ null and alternative hypotheses
  - Collect data in a _____
  - Determine an appropriate _____ test and distribution to represent our _____
  - Conclude whether we can _____ the null hypothesis or if we cannot

> **Important**
>
> I think it's important to emphasize that it is not our 'goal' to reject the null hypothesis - as empiricists, our orientation should be towards truth, not a certain outcome

5.2 Hypotheses Pairs

- A hypothesis is just a prediction or a statement of a possible _____
  - E.g., I *hypothesize (or predict)* that college students who are Gen Z have significantly worse attention than Millennial college students
  - Most scientific _____ begin with some hypothesis of what outcomes will look like

Discuss: Consider that you are a middle-school teacher being asked by administration to implement an especially permissive attendance structure, write a hypothesis of what this will do to student math scores

- When we make _____ in statistics, we do so in a _____ pair by making a null hypothesis and an alternative hypothesis
  - In _____ statistical testing, we frame our inferential tests as helping us determine whether to reject the [Null Hypothesis] - there is a whole other field of statistics called _____ statistics that has a somewhat different focus

Important

It's important to understand that, as part of the statistical testing process, we should always consider what our null and alternative hypothesis is!

5.2.1 Null Hypotheses

- A null hypothesis is a prediction or statement of _____ difference or relationship between two or more things
  - It is usually written as H_0
  - _____ of alternative hypothesis
- A null hypothesis is often written to include an _____ sign such as with:
  - $=$
  - But also, $\geq$ & $\leq$
- Example: There is *no* difference in average attention between Gen Z and Millennial college students
  - Represented in notation $H_0 : \mu_{Attn-Z} = \mu_{Attn-Millennial}$

Discuss: Write the null hypothesis for comparing mental health between women and men in the legal field, where they have equal levels of depression

5.2.2 Alternative Hypotheses

- Our alternative hypothesis suggests a difference between two or more things
  - It is most often represented as H_A or H_1
  - Opposite of Null Hypothesis
  - This is usually close to how we would phrase a _____ hypothesis
- The _____ hypothesis does *not* contain equal signs
  - This would include $\neq$, $>$, and $<$
- Example: There *is* a difference between Gen Z and Millennial college students in average attention
  - Represented in notation $H_A : \mu_{Attn-Z} \neq \mu_{Attn-Millennial}$

Discuss: Write the alternative hypothesis for comparing mental health between women and men in the legal field, where women have higher levels of depression

Discuss: Look back at the example null and alternative hypotheses - am I using notation for population parameters or sample statistics? Why do I do it this way?

5.3 Testing

- While accounting for the _____, statistics, parameters, and sampling distribution we have do help us make good choices about hypothesis testing - they don't tell us everything
  - We always have to remember that things in statistics are _____!
  - With our sampling and data gathering we may run into a rare event
- To account for the _____ of a rare event, we test the null hypothesis somehow

Discuss: Try re-explaining, in your own words, what 'probablistic' means and why it's readily applicable to the practice of statistics

5.3.1 Using the Sample to Test the Null Hypothesis

- This is where we introduce the p-value, which is the _____ that, under the null hypothesis being _____, our results will be as extreme as they are
  - Example: a test result returns a p-value of 0.07. Assuming the null hypothesis is true, this result only had a 7% chance of occurring.
- Effectively, when we _____ against the null hypothesis and determine a p-value, we are trying to gauge how likely our results were to

occur if the null hypothesis was true
  - If our results are especially _____ under the null hypothesis (i.e., a low p-value), then we may be inclined to believe that our case is somehow truly different, and thus the null hypothesis is incorrect and can be _____

> **Important**
>
> There are many misunderstandings people have about statistics, but failing to understand the meaning of p-values is probably the single most common and pervasive errors people make in interpreting statistics.

5.3.2 Decision and Conclusion

- How do we determine if the p-value is, "rare enough"?
  - Ideally, we test it against a preset/preconceived significance level, also given as α.
  - We _____ whether our p-value is $\geq$ or $<$ our α
- If you don't see further information, the most _____ significance level is $\alpha = 0.05$
  - However, this isn't a hard set rule.
- If our p is $< \alpha$ then we say we have statistical significance
  - In the context of an inferential test, like a t-test, we are looking for our test statistics to be more extreme than the critical value

5.4 Tails of a Test

- A test may be described as two-tailed, left-tailed, or right-tailed, dependent on the _____ used in the alternative hypothesis
  - $H_a : P > 0.5 \rightarrow$ right-tailed (one-tailed)
  - $H_a : P < 0.5 \rightarrow$ left-tailed (one-tailed)
  - $H_a : P \neq 0.5 \rightarrow$ two-tailed
- This needs to be set as part of the study set up, not during analysis!

Discuss: Consider the following example: Johnny predicts that Samantha has more money than Becky right now. Is this a one-tailed or two-tailed test and why?

6 Conclusion

6.1 Recap

- This was a (not so) quick recap of basics ideas in describing and examining data, and the nature of variables, samples, populations, and distributions
- We saw examples of how to appropriately apply certain descriptive procedures, and also discussed some of the limitations
- We also began talking about the framework of hypothesis testing and how that provides a way for us to determine if results are significant or not

Key Terms

A

alternative hypothesis A statement of non-equivalence between two or more things; opposite of the null hypothesis 24

B

Bar graphs Show frequencies of values of categorical/nominal variables 16
bimodal When a distribution has two peaks 14
bin width The range of each 'bar' in a frequency histogram 18
Boxplots A graphical representation of a continuous variable that shows the median, 25th percentile, and 75th percentile 18

C

D

E

F

H

I

K

The instructor-provided glossary may not include all terms worth memorizing, make sure you consider using the vocabulary list in you book and your own judgment to make sure you have all relevant terms