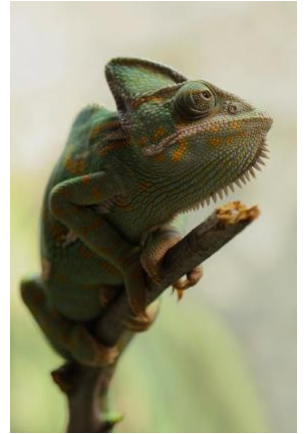


Chapter 1: Statistical Investigations

Investigation: Chameleons catch prey by extending their tongue quite some distance from their body. As a zoologist you're studying this ability among a species of chameleons in central Africa.

One question your research group has had is whether the environment that these chameleons live in might be affecting tongue extension. In particular, you'd like to compare the tongue extension abilities of chameleons in the wetter rainforest to those in the dryer savanna. **You're curious to know: Does environment affect how far these chameleons can stretch their tongue?**



Realistically, *how* would you collect some data that might give you insight toward this question? *Like really think through the process here so that someone else could replicate it!*

If Time: Other zoologists (e.g., your classmates!) were given the same question. Do you think the process they described will be the same as yours? Or do you expect there will be differences?

If Time: You and another researcher agreed on a process for data collection/analysis but completed your studies separately. Do you think the results will be the same, or could you imagine the results being different?

The Vocabulary of a Statistical Investigation

- A statistical investigation collects data from a sample, typically with the purpose of trying to make claims or draw insights about a larger population.
 - A **population** would represent everyone/everything that we would like to generalize toward.
 - A **sample** would represent a _____ of the population. Even if they are but a small fraction of the population's size, it can still be a good sample if it's representative of the population!

In the chameleon investigation...

Population:

Sample:

- We'll also need to identify who/what we're observing and what it is we're measuring from each unit.
 - A _____ would be one element or one case from that population that we are collecting data from (a person, an animal, an object, a time point, a location, a group, etc.). We might think of this as the source from which we generate one data point.
 - A **variable** can be thought of as an _____ or a _____ that might vary with each unit we observe. Variables might be numeric or categorical.
 - Be careful not to confuse **unit of observation** with **unit of measurement** for a particular variable. The first is a case that we collect data from, while the second is our scale for measuring something (e.g., centimeters, gallons, heart bpm)

In the chameleon investigation...

Unit of Observation:

Variables of interest:

- To summarize our findings, we might report one or more statistics of interest that we hope serve as good estimates for our parameter(s) of interest.
 - A **statistic** is a variable summary at the sample level.
 - A **parameter** is a variable summary at the _____ level. In a statistical investigation, we typically don't know parameters with certainty—rather, our statistics serve as our parameter estimates!
 - Variable summaries might describe just one variable from our data (like a mean), or we could calculate a summary measuring the relations between several variables (like a difference in means for two groups, or a correlation coefficient).
 -

In the chameleon investigation...

Statistic possibilities:

Acknowledging Uncertainty! Why is it difficult to offer a definitive answer to a research question from one statistical investigation?

1. Uncertainty in whether our sample results **generalize** to the wider population
 - The way that we choose or locate individuals to be in our study may result in an _____ sample of the population.
 - If our sampling process doesn't give everyone an equal chance, there could be a **systematic bias** in the group we eventually have.
 - A question to ask when considering generalizability:
2. Acknowledging that our statistics will have some **precision** error when estimating parameters.
 - Even if we did have a representative sampling process, your sample will differ from my sample. We call that _____.
 - Since every possible sample will differ, the statistics we generate will vary from one sample to another.
 - A question to ask when looking at a statistic:
3. And when asking causality questions, we might have uncertainty about the **mechanism of causality** between two variables we studied.
 - We might find a link between two variables we examined, but there could be other confounders that explain why that association exists.
 - Is environment actually affecting tongue length? Or are there other differences between these two groups of chameleons that might be causing the difference?
 - A question to ask when evaluating causality:

And in general, we need to acknowledge decisions we make, from the way we measure our variables, to the analytical measures and methods we use. These may not be right or wrong choices, but just different approaches or perspectives. A question to ask in general:

Read on your own



Identifying Different Types of Variables

- **Nominal Variables**
 - Variables whose outcomes fall into categories with no inherent ordering/scale.
 - What flu symptoms have you been experiencing? (nausea, fever, chills, etc.)
 - What fruits do you like to eat? (apples, grapes, strawberries, kiwi, etc.)
 - Does this state require photo ID to vote in elections? (yes, no)

- **Ordinal Variables**

- Variables whose outcomes fall into categories that have a meaningful ordering (but not on a true numeric scale)
 - Are you a Freshman, Sophomore, Junior, or Senior?
 - Do you strongly disapprove, somewhat disapprove, somewhat approve, or strongly approve of the President's job performance?

- **Discrete Variables**

- Variables whose outcomes fall on a numeric scale, but only takes limited values (like whole numbers). These are typically things that are *countable*.
 - What *year* of school is this for you? (1, 2, 3, 4...)
 - How many days last month did you go to the gym?
 - How many people showed up to class today?
 - What is the number of blueberries that you picked today?

- **Continuous Variables**

- Numeric and measurable (can take any value in a range)
 - What is the heaviest amount of weight that you can bench-press?
 - How much time did you spend on your exam before turning it in?
 - How many ounces of blueberries did you pick today?



- **Special cases of identifying types of variables**

- Binary variables would typically **not** be thought of as ordinal or discrete...that is because you can't have meaningful ordering with only two categories. We think of it as **nominal**.
- Just because a variable is recorded numerically does **not necessarily** mean it is discrete/continuous. **Zip Codes**, or categories that have been arbitrarily numbered, may better be thought of as nominal if the numbers are really just category names.
- **Likert-scale items** (e.g., 1 to 5, 1 to 10 ratings) are *typically* considered ordinal, even when presented as numbers. That's because the distance between a 1 and 2 may not be equal to the distance between 2 and 3. Likert scales have ordering, but the numbers are really just categories rather than numeric values.

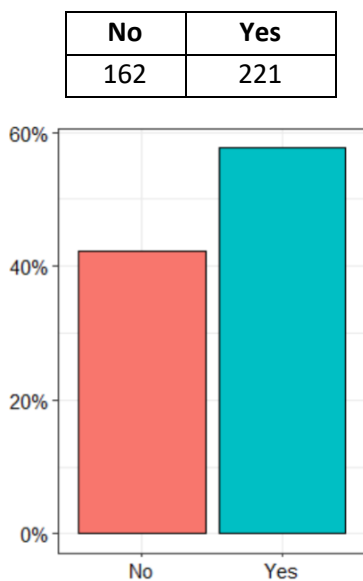
Visualizing Categorical Data

• Barplots

- Barplots are a common visualization choice for a single categorical variable.
- Since observations of categorical data fall into distinctly identifiable groups, we can represent those groups on one axis and represent the frequencies or proportions on the other axis.
- On the left is an example of a plot where the x-axis represents potential categories, and the y-axis shows the proportion of responses in each category. Likewise, the graph on the right shows categories on the x-axis, but instead is counting up number of responses on the y-axis.

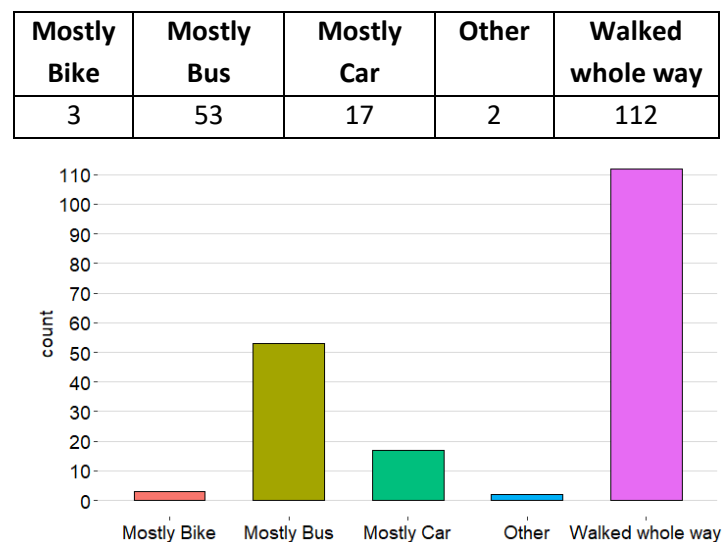
We asked students whether they have a pet at home

Table 1. Frequency table showing how many people said yes or no.



We asked students in January 2020 how they got to class that morning

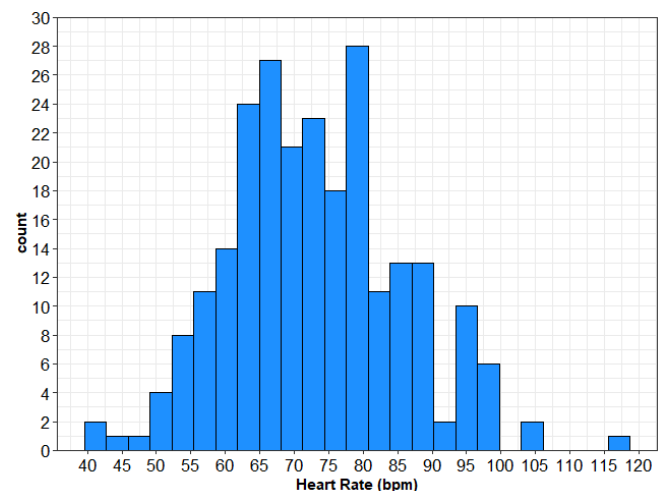
Table 2. Frequency table showing how many students gave each response



Visualizing Numeric Data

• Histograms

- Histograms are a common visualization choice for a single numeric variable.
- A single **variable** is represented in the **x axis**, while the **y axis** typically represents the **count**: # observations in each particular bin.
- The key difference is that now, our observations are not in distinguishable categories. We choose a bin size to represent how many observations are in each possible numeric range.
- In this plot, we're representing how many times students' heartbeat per minute and counting responses in each numeric range.



Reflection Questions

1.1: What is the difference between a population and a sample? Which one does a statistic describe, and which one does a parameter describe?

1.2: What is the difference between a unit of observation and a unit of measurement?

1.3: Come up with your own example of a statistical investigation! What might be the population? The unit of observation? A variable (or variables) of interest? An example of a statistic you might calculate?

1.4: Why is it difficult to provide definitive answers from a statistical investigation? What questions might we ask to help us assess our uncertainty?

1.5: How might we distinguish discrete from continuous variables? Or nominal from ordinal variables? Why might a variable like telephone area codes not be treated as “numeric” data?

1.6: When do we use a histogram versus a barplot?

Examining _____ Data

Investigation: Let's say we wanted to know the likelihood that a randomly selected University of Illinois graduate student (at or above the legal age of 21) has used a marijuana product at least once since being a student.



Population of interest:

Unit of Observation:

Variable of interest (and what type of variable?):

Let's say that in this study, we contacted 54 graduate students and assured them that their responses would remain anonymous. Of these 54 students, 19 of them answered yes.

Our sample in this investigation would be...

- Statistics for describing Categorical Data
 - A **proportion** represents the number of cases that fit a category of interest divided by the total number of cases. It ranges from _____
 - π is a _____, representing a **population** proportion.
 - \hat{p} is a _____, representing a **sample** proportion.
 - A proportion is just the mathematical form of a percentage.
 - A proportion of 0.42 is the same as 42%.
 - A proportion of 0.894 is the same as 89.4%.

Do we know \hat{p} in this investigation?

Do we know π in this investigation?

Let's fill in this table with some of the symbols we encounter!

Table 3. Symbol Representations

Statistics	Parameters

Examining _____ Data

Investigation: Let's say that you recently started taking a medication that tends to mildly suppress appetite as a side effect. You would like to better understand what your typical daily caloric intake is while on this medication. You collect data over an 8-day period, and the results are below:

1920, 1860, 2570, 1520, 1860, 2050, 1750, 2180



Variable of interest (and what type of variable?): Daily caloric intake (continuous?)

Unit of Observation (A Calorie? A Day? A Person?):

Population of interest: All days that she has team practice

- Measuring Numeric Data – **Measures of Center**
 - The _____ represents the *balancing point* of our data. It is found by adding up all data values and dividing by the sample size.
 - μ is a parameter, representing a **population** mean.
 - \bar{x} is a _____, representing a **sample** mean.
 - The _____ represents the value of the middle observation. It's the value such that approximately half the data is at or below that value and half the data is at or above.
 - Some use m to represent a sample median and M as a population median, but median is not commonly used symbolically.
 - In the case of an odd number of data points, the median is the middle data value. In the case of an even number of data points, it's the average of the middle two values.
 - Mean and median differences.
 - While the median is not responsive to outliers, the mean is responsive to every data point, and outliers can significantly change the mean!
- Visualizing Numeric Data with a Dotplot
 - While histograms work well with larger sample sizes, a simple dot plot helps us visualize all of our numeric values on a scale. Our x axis represents a numeric scale, and we are clustering dots in "bins" if they are rather close in value so we can see each observation.



Chapter 1: Statistical Investigations

Let's calculate both the median and mean of our caloric intake and then note how they relate to the dotplot on the previous page.

What is our sample median: m ?

What is our sample mean: \bar{x} ?

Let's say that the 2570 data point was mis-recorded. It was supposed to be 2070.

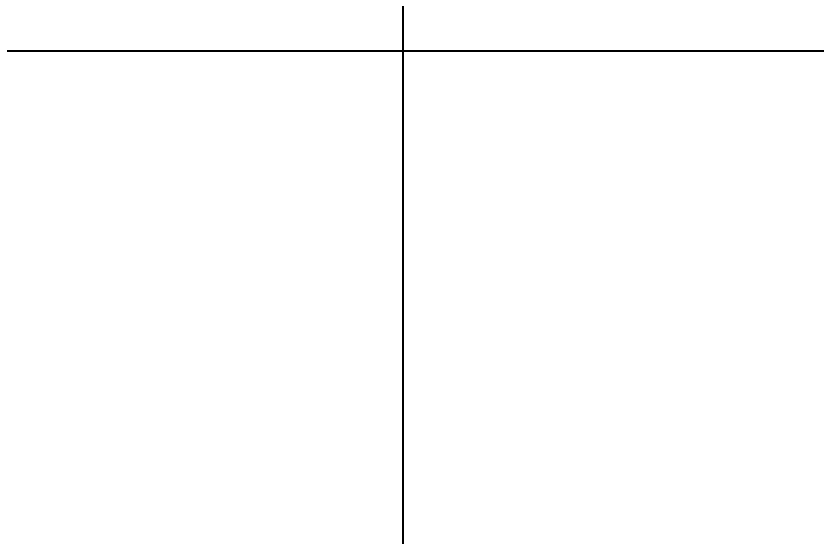
Would the median be affected by that change?

How about the mean?

Investigation Reconsidered: What if I instead wanted to know how *consistent* my daily caloric intake was from day to day. Would taking the mean or median also answer that?

- Measuring Numeric Data – **Measures of Variability**
 - The _____ is a very basic measure of variability.
 - It reports the distance between the highest and lowest value.
 - The range of our sample caloric intake data is: _____.
 - But the range is not a very “robust” measure. It's affected by outliers and doesn't tell us anything about the points in between.

The _____
(MAD) is a more robust option. We find the
average distance from the mean in our data.



- To calculate the MAD, we find each point's distance from the mean and then average the distances. *The absolute value symbols ensure that each distance is reported as a positive value.*

$$\frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \dots + |x_n - \bar{x}|}{n} = \frac{|1920 - 1964| + |1860 - 1964| + \dots + |2180 - 1964|}{8} = 227 \text{ calories}$$

- The **variance** is more commonly used in statistics as a measure of variability. It's quite similar to MAD, but now we are finding the average *squared* distance from the mean, rather than the average distance.
 - If we knew the population mean, we could calculate the **parameter** for variance like this.

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \dots + (x_N - \mu)^2}{N}$$

- More commonly though, if calculating variance as a **statistic**, we need to make some adjustments. What changes do you notice below?

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}$$

- One problem with variance is that it will no longer be in the approximate units of our variable since we squared the distances. That's why analysts often report the standard deviation.
 - σ is a **parameter**, representing the population standard deviation.

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}} = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \dots + (x_N - \mu)^2}{N}}$$

- s is a **statistic**, representing our sample standard deviation

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}}$$

- A simpler, more informal interpretation for **standard deviation** would be...

Later, we'll talk about how we can use standard deviation to discuss how **unusually far** certain values are from the mean. To get a head start (and get a more visual explanation), here is a [fun video!](#)

Digging Deeper: Why are variance and standard deviation used rather than the MAD?

- The formulas we are presenting above are data-centered descriptions, but in higher level statistics, we actually use calculus-based definitions for the measures we create
- The MAD does not work well as a calculus-based representation. That is because absolute deviations are not natural algebraic operations and are not "continuously differentiable."
- The variance in particular has very nice mathematical properties. For example, $\text{Var}(X) + \text{Var}(Y) = \text{Var}(X + Y)$, and for this reason, variance will be used commonly in statistical methods that require some measure of variability!

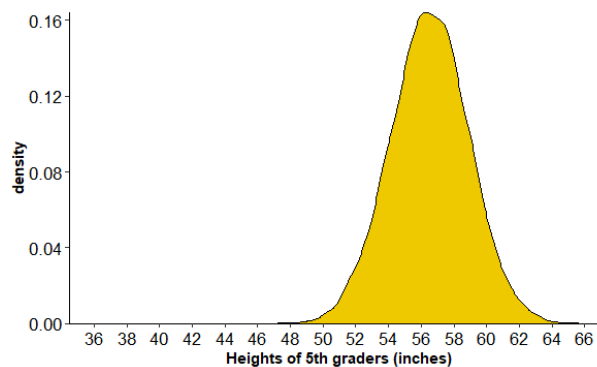
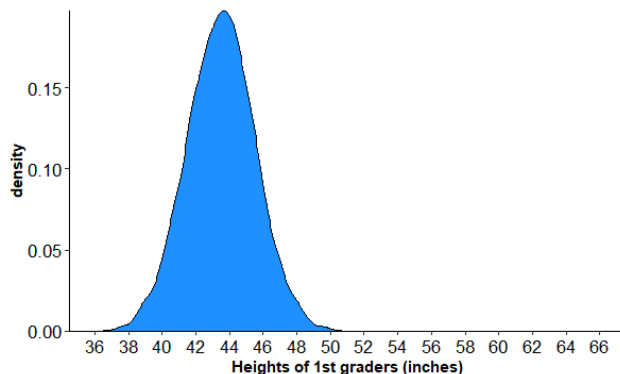
(Digging Deeper info boxes are extra informative pieces that you do not need to know for the exam!)

Read on your own

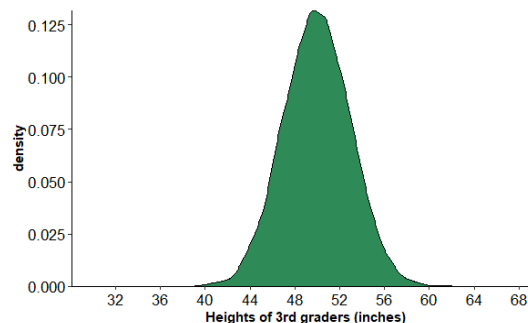
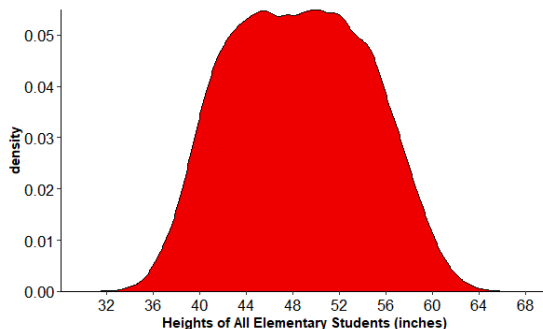


Identifying features of a distribution

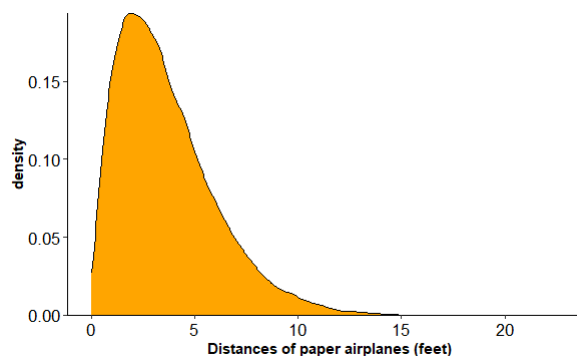
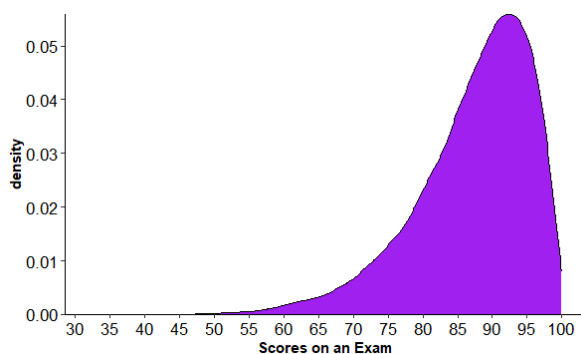
Center: Where is the “middle” of the distribution?



Variability: How far do data points typically extend from the center?



Symmetry/Skewness: Is the data symmetric, or is it skewed in one direction or another?

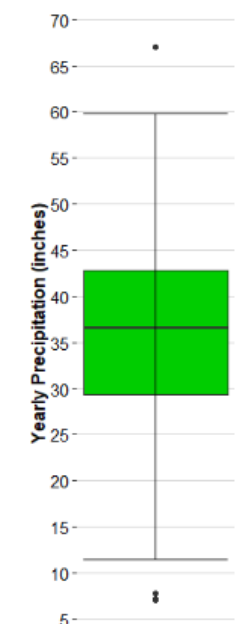
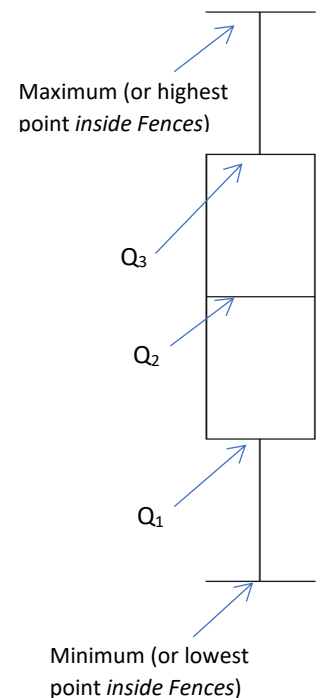
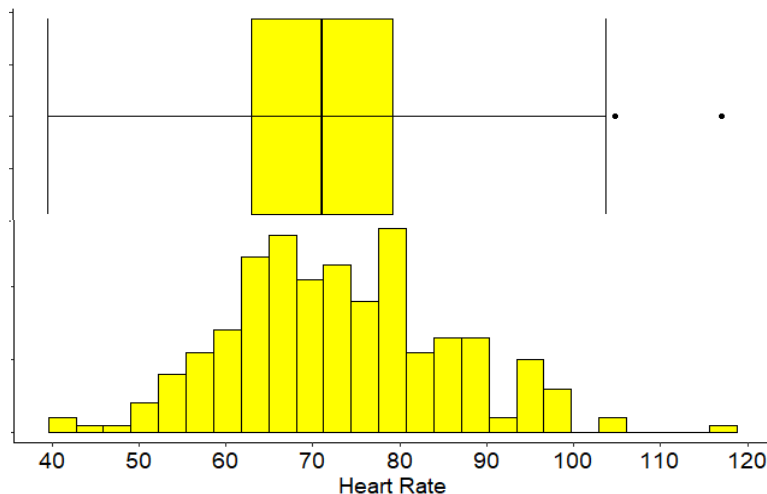


Data that stretches to the left side may be called **left skewed** or **negatively skewed**.

Data that stretches to the right side may be called **right skewed** or **positively skewed**.

Chapter 1: Statistical Investigations

- Other Measures of Position
 - **Percentiles** are used to reference values at key positions in a distribution.
 - For example, the 20th percentile would be the value such that 20% of your data is at or below that point.
 - The median may also be referred to as the 50th percentile.
 - Data descriptions may commonly reference **quartiles** as well.
 - Q_1 is the 25th percentile (the median of the *lower* half of the data).
 - Q_2 is the 50th percentile (the median of the entire set of values).
 - Q_3 is the 75th percentile (the median of the *upper* half of the data).
 - 5-Number Summary
 - **The 5-number summary** represents the boundary points of the 4 quarters of your data: (Minimum, Q_1 , Q_2 , Q_3 , Maximum).
- **Boxplots** are a graphical representation of the 5-number summary of a numeric variable.
 - The “whiskers” (outside lines) are the minimum and maximum values still inside the Upper/Lower fences.
 - Lower Fence = $Q_1 - 1.5(Q_3 - Q_1)$
 - Upper Fence = $Q_3 + 1.5(Q_3 - Q_1)$
 - Outliers are denoted by tiny dots past the first or last whisker—data values that fall outside these fences.



Practice: A meteorologist records the yearly precipitation in 70 large U.S. cities. Between what 2 precipitation amounts do the middle 50% of cities fall in?

The first and third quartiles appear to be around 29 to 43 inches, so that range is capturing the middle 50%

What proportion of cities see at least 43 inches of rain a year?

Since the third quartile is around 43, then about 25% of cities are at or above that.

Reflection Questions

1.7: How do we pronounce each of these symbols, and what do they represent? \hat{p} , π , μ , \bar{x} , s , σ

1.8: How are mean and median similar measures? How are they different?

1.9: What statistics comprise the “5-number summary” of a numeric variable? How does this set of numbers relate to the boxplot representation?

1.10: The daily temperatures of two different cities are tracked for a year. Both report the same mean temperature, but one city’s temperature has a much higher standard deviation. What does that tell us about temperatures in that city?

Multivariate Investigations

- **Univariate vs. Multivariate Investigations**
 - **Univariate Questions:** Ask about characteristics of...
 - **Multivariate Questions:** Ask about the...
- Identifying a response variable
 - A **response variable** is a variable that we have an interest in better understanding or predicting. It is the target outcome of our investigation.
 - An **explanatory variable** (or may also be called a predictor variable) is a variable that we think might help predict or explain the response variable. We *may* suspect it is the causal agent.

Example: Do students who come to class score better on the Exam than students who don't?

The response variable is...

The explanatory variable is...

Comparing Proportions

Investigation: A [Study from Science Daily](#) found that people who express a variant of the DNMT3B gene were more likely to develop a nicotine dependence and be heavy smokers. The researchers collected data from 38,600 adults across the U.S., Iceland, Finland, and the Netherlands.

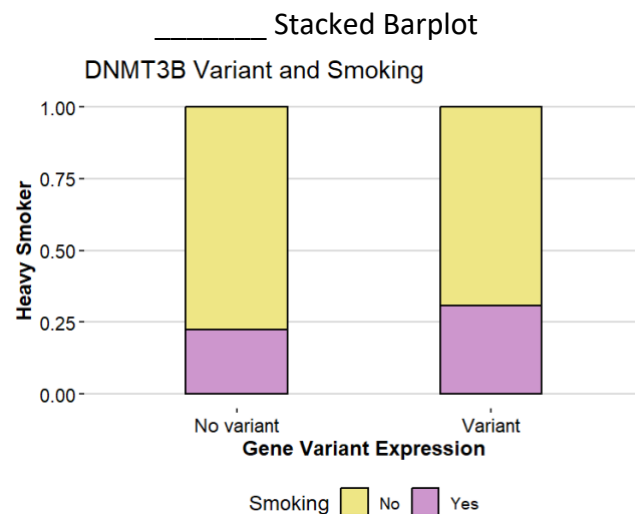
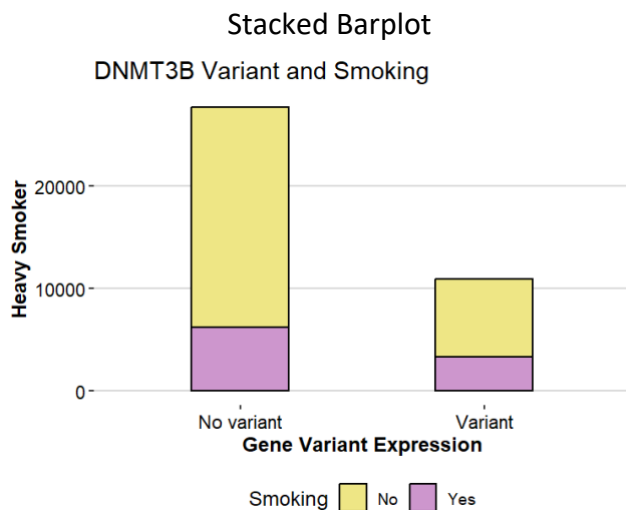
Unit of observation:

Response variable (and type):

Explanatory variable (and type):

One simple option for our parameter of interest would be...

If that's the case, then our statistic of interest used to estimate that parameter would be...



Comparing Means/Medians

Investigation. Consider a garden of iris plants. A botanist measures petal length of each iris that has blossomed. He wants to know if the petal length of the Virginica species might be higher than that of the Versicolor species.

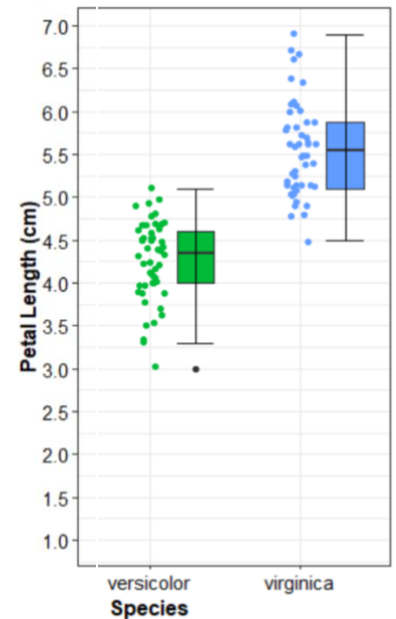
Unit of observation:

Response variable (and type):

Explanatory variable (and type):

One possible parameter of interest would be $\mu_1 - \mu_2$

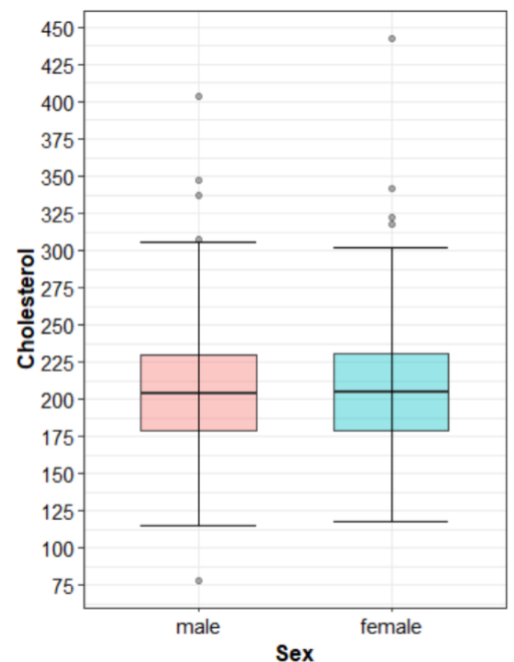
And that would make our statistic of interest be...



*Note that with a numeric variable and categorical variable, we might do **side-by-side boxplots** or a **jitter plot** to easily compare the numeric distribution of each group.*

Example. Are cholesterol levels different by biological sex? Consider the following data representing approximately 403 adults.

Just using the graph, do you think biological sex explains much variability in cholesterol levels? In other words, does knowing biological sex help us make a better prediction of someone's cholesterol level?



Does iris species seem to explain much variability in petal length?

Measuring Association between Numeric Variables

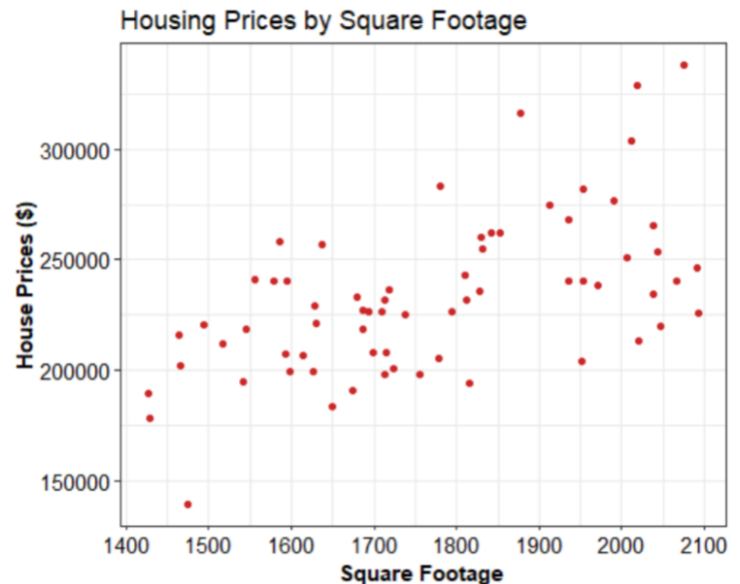
Investigation. Consider the following plot, representing 67 houses in a particular community. Does the square footage of a house help us better predict the price of the house?

Unit of observation:

Response variable (and type):

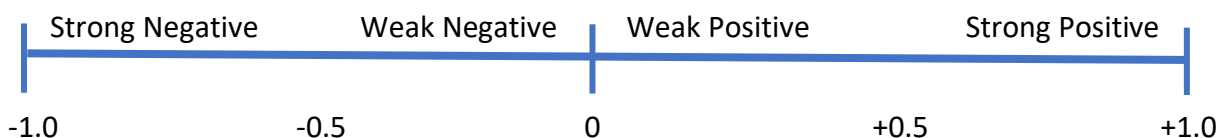
Explanatory variable (and type):

There are a lot more options for statistics and parameters in this data context, but we'll first consider a fairly simple measure of association.



With two numeric variables, it's common to create a **scatterplot** to represent the data.

- One option for comparing two numeric variables is Pearson's _____.
 - A statistic between ____ and ____ that describes the direction and strength of a **linear** association between two numeric variables.
 - Negative values imply that as one variable increases in value, the other decreases in value. (*Negative correlation*). Positive values imply that as one variable increases, the other variable increases as well (*Positive correlation*).
 - The correlation coefficient is abbreviated r (for sample statistic) or ρ (for population parameter).



- How would you calculate the correlation coefficient between two variables?**
 - In this class, you will **never** be asked to calculate r by hand from a set of data, but here is the formula!

$$\text{Formula: } r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

*Note that Pearson's correlation coefficient is not designed to measure _____ relationships.

Chapter 1: Statistical Investigations

Practice! Identify how we might frame each investigation below statistically based on the basic approaches we've learned so far. *Answers may differ, as some investigations may look different depending on how we think about the variables we are identifying or measuring!*

Investigation: A public health official is examining flu rates per county (infections per 10,000) to see the flu vaccination rates (some proportion from 0 to 1) might help explain differences in infection rates.

Unit of Observation:

Variable(s) of interest and variable type(s):

This is best visualized with:

A. Histogram B. Univariate Barplot C. Stacked Barplot D. Scatterplot E. Side-by-side Boxplots

Investigation: Are men more likely to overestimate their height on dating apps as compared to women?

Unit of Observation:

Variable(s) of interest and variable type(s):

This is best visualized with:

A. Histogram B. Univariate Barplot C. Stacked Barplot D. Scatterplot E. Side-by-side Boxplots

Investigation: How much did University of Illinois students spend on food last week?

Unit of Observation:

Variable(s) of interest and variable type(s):

This is best visualized with:

A. Histogram B. Univariate Barplot C. Stacked Barplot D. Scatterplot E. Side-by-side Boxplots

Reflection Questions

1.11: In a multivariate investigation, how would we distinguish a response variable from an explanatory variable?

1.12: What are the multivariate data visualizations we learned about in this chapter? Which types of variables might we use with each of them?

1.13: Earlier in the chapter, we compared two sets of side-by-side boxplots and asked which showed more evidence of the predictor explaining a lot of variability in a response variable. What would we expect that to look like in a scatterplot? In a 100% stacked barplot?

Chapter 1 Additional Practice (Videos available for these in Canvas Ch 1 module)

Investigation: In 2019, Gallup conducted a poll to gauge the opinions of Adult U.S. Residents about gun laws. Gallup contacted a representative sample of 1,526 people. Among several questions asked, one asked about whether or not you supported a complete ban on individual gun ownership. 29% said yes.

Our population is...

The unit of observation is...

Our variable of interest is...

The sample statistic they gathered is...

Do we know what the population parameter is?

Here is [Gallup's full report](#) from their poll to Americans on gun policy:

Practice Identify the variable studied and its data type.

20 runners run a mile as fast as they can. Their times are recorded.

Identify the variable of interest: _____

Nominal, Ordinal, Discrete, or Continuous (circle one)



50 Students are asked what their major is.

Identify the variable of interest: _____

Nominal, Ordinal, Discrete, or Continuous (circle one)

100 Married Couples are asked how many children they have.

Identify the variable of interest: _____

Nominal, Ordinal, Discrete, or Continuous (circle one)

20 runners are asked to run a mile as fast as they can. Next to each runner's name, the coach records "yes" or "no" to indicate whether or not they broke the 5-minute mark.

Identify the variable of interest: _____

Nominal, Ordinal, Discrete, or Continuous (circle one)

Judges score musicians across a number of different criteria using four choices: "superior," "excellent," "good," or "needs work."

Identify the variable of interest: _____

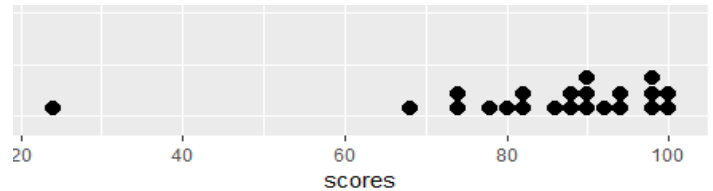
Nominal, Ordinal, Discrete, or Continuous (circle one)

Practice: Consider the following 22 scores for a recent test, where scores could be anywhere from 0 to 100.

Scores: 24, 68, 74, 74, 78, 80, 82, 82, 86, 88, 88, 90, 90, 90, 92, 94, 94, 98, 98, 98, 100, 100

The median of this distribution is...

- A. 24 B. 50 C. 60 D. 82 C. 89 D. 100



If we removed that score of 24, which value do you think would be most affected: mean or median?

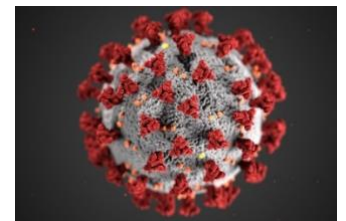
If we removed that score of 24, would that increase or decrease the variability in our data?

Does this distribution appear to be skewed? If so, in what direction?

Investigation: Early in the COVID-19 pandemic, researchers were trying to understand just how dangerous a threat it was to someone infected. Imagine you were a medical researcher. How might you collect data to estimate the mortality rate of COVID-19?

Unit of Observation:

Variable(s) of interest and variable type(s)



Statistic of interest:

This is best visualized with:

Investigation: A psychiatrist wants to see whether patients who have begun taking the antidepressant “Zoloft” are more likely to report having experienced nausea in the past two weeks compared to patients who took a placebo (non-effective) tablet.

Unit of Observation:

Variable(s) of interest and variable type(s)



Statistic of interest:

This is best visualized with:

