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 chameleons in a large natural habitat with thousands of adult chameleons.

In the last few months, a second chameleon species has migrated into this habitat (they are easy to distinguish based on their coloring). You're curious to know how the introduction of this species might potentially change the original species' tongue stretching ability in this habitat if cross-species mating happens. So, two questions we might wish to answer with data:

1) How far on average can adult chameleons in this habitat stretch their tongue to successfully catch a bug.
2) Is this average distance different between the two chameleon species in this habitat?

What data would you need to answer these questions, and *how* might you realistically plan to collect it?

Once you collected that data, what would you do with it to help you answer each question?

If other zoologists (perhaps your classmates!) were given the same questions, would you expect them to get the same answers? If not, why not?
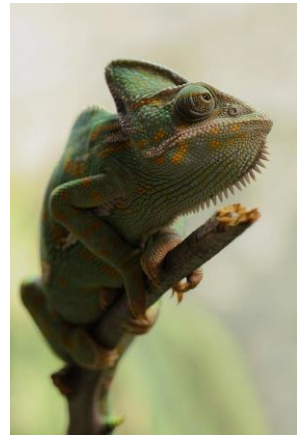
Setting up a Statistical Investigation
- A statistical investigation collects data from a sample, typically with the purpose of trying to make claims or draw insights about a larger population.
    - A **population** would represent everyone/everything that we would like to generalize toward.
    - A **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 our sampling units 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 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:

Uncertainty! Acknowledging what we don't know.

So why might two people come to different answers when tasked with the same statistical questions?

1. There will be uncertainty in the _____ of our statistic.
   o The attributes we are collecting will **vary** from unit to unit. That's why we call them *variables*!
   o Thus, every possible sample (even if collected in a representative manner) will differ, and the statistics we generate will vary from one sample to another. We call that **sampling variability**.

2. There may also be uncertainty in the _____ of our investigation's findings depending on the choices we made in designing and carrying out our investigation.
   o This might be because of problems and mistakes that introduce **bias**—a systematic imbalance or error that will steer our responses to be consistently off.
     ▪ These might be measurement biases or sampling biases.
   o But even without errors, we need to acknowledge that researchers may make different *choices* about the setting, the instrumentation, and statistical measures we use.
     ▪ These aren't right or wrong choices, but just different approaches or perspectives.
   o What might a bias look like in the context of the chameleon investigation? What might constitute a neutral choice that would affect our generalizability.

3. And specifically in the context of multivariate questions, we might have uncertainty about the mechanism of _____ between two variables we studied.
   o We might find a link between two variables we examined, but determining whether there's a causal relationship from one the other requires careful design and analysis.
   o Does something about species genetically affect tongue extension, or is it just differences in the environments or the types of bugs each species was in that won't show up in offspring?

 Read on your own

Identifying Different Types of Variables

- **Nominal Variables**
  o 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?
    - Likert-scale items that ask the extent to which you agree or approve (e.g., strongly disagree, somewhat disagree, neutral, somewhat agree, strongly agree).
- **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?

```
         Categorical                    Numeric
        /          \                   /         \
   Nominal      Ordinal           Discrete    Continuous
```

- 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.

No	Yes
162	221
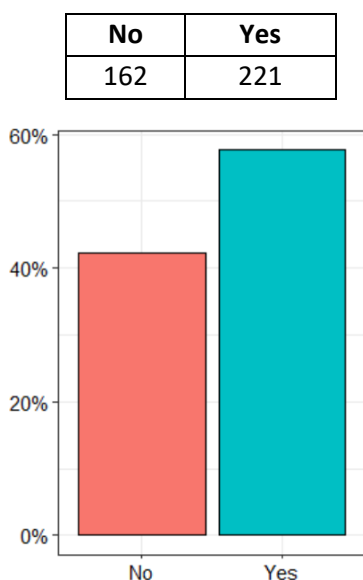
We asked students in January 2020 how they got to class that morning

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

Mostly Bike	Mostly Bus	Mostly Car	Other	Walked whole way
3	53	17	2	112
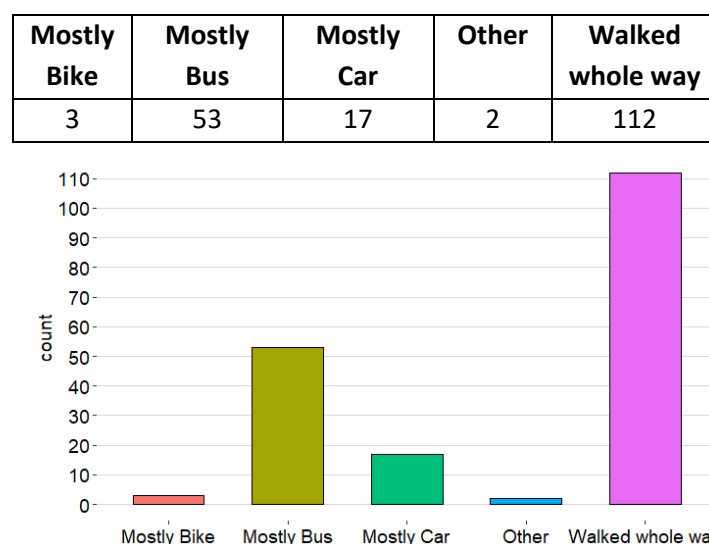
Visualizing Numeric Data

- **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 up responses in each numeric range.
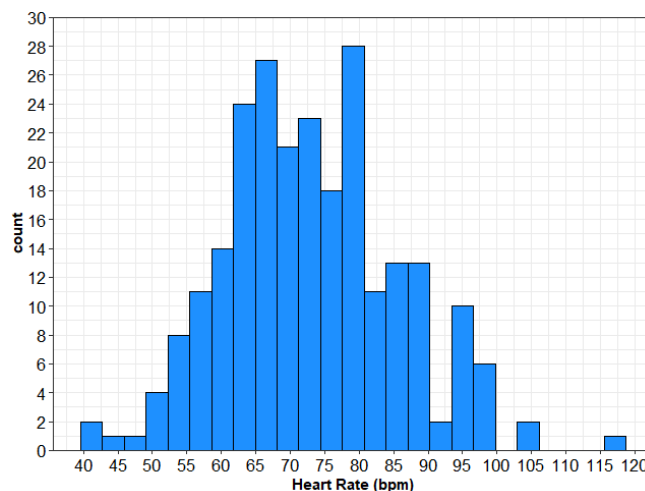
Examining _____ Data

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: I would like to better understand what my typical daily caloric consumption is. How might I complete a statistical investigation here?

Population of interest:

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

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

Let's say that we record our caloric intake for 10 days and report the following amounts:

2120, 1870, 1920, 1860, 2570, 1520, 1860, 2050, 1750, 2180

- 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.
    - With a small sample size, it's easy to again visualize our data with dots (even if we prefer histograms in general with larger samples). 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.
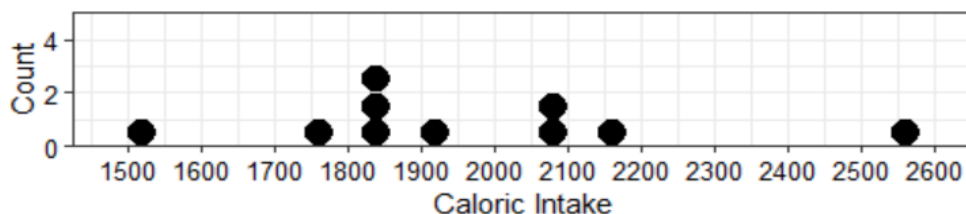
Let's calculate both the median and mean of our caloric intake and then note how they relate to the dotplot on the previous page.

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**
  - o 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 another option. The idea here is to
find the average distance from the mean in
our data.
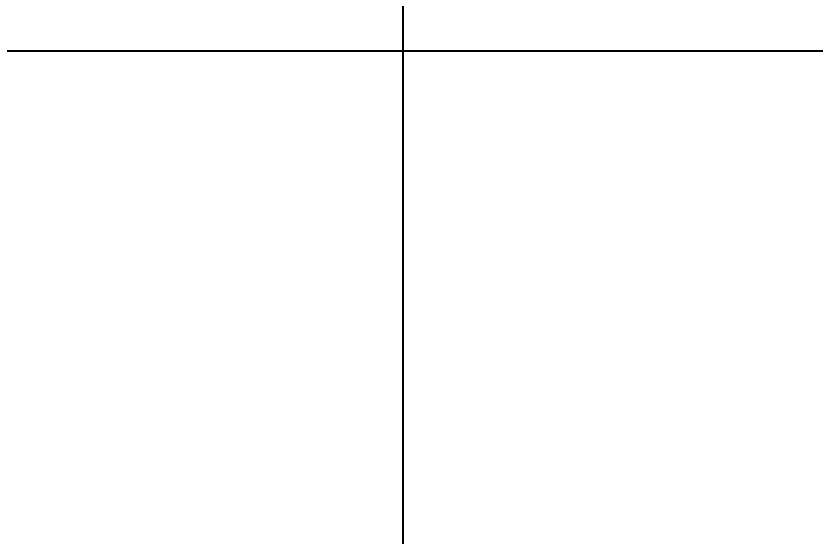
To calculate the MAD, we could find each point's distance from the mean, and then average the distances. *The absolute value symbols ensure that each distance is reported as a positive value.*

$$\frac{|x_1-\bar{x}|+|x_2-\bar{x}|+ ...+|x_n-\bar{x}|}{n} = \frac{|2120-1970|+|1870-1970|+ ...+|2180-1970|}{10} = 208 \text{ calories}$$

But while the MAD works well in some ways, but there's a reason it isn't commonly used as a measure of variability in most statistical analyses. *You don't need to remember these for the exam!*

- Mathematically, the MAD is not what we call "continuously differentiable" or "continuously integratable." In other words, absolute value operations don't work well with calculus.
- Also, there is another measure of variability that has nice mathematical properties you would learn about in a mathematical statistics course.

For these reasons, you will more commonly see data analysts use the variance or the standard deviation!

- The **variance** is quite similar to MAD. The main difference being that we are finding the average *squared* distance from the mean, rather than the average absolute distance.
    - If we knew the population mean, we could calculate the *parameter* for variance like this.

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

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

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

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

    - σ is a *parameter*, representing the population standard deviation.

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

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

$$s = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}} = \sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+ ...+(x_n-\bar{x})^2}{n-1}}$$
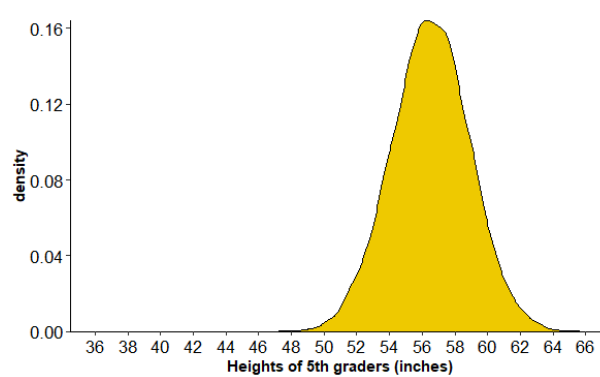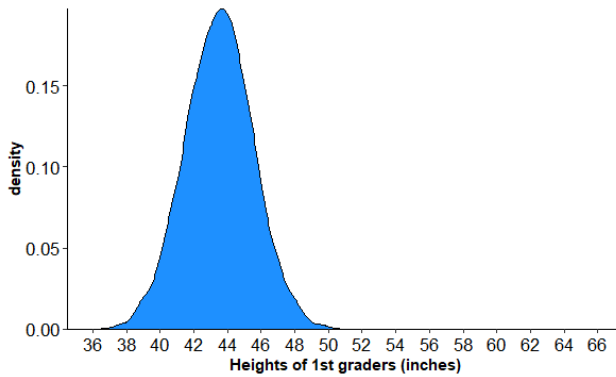
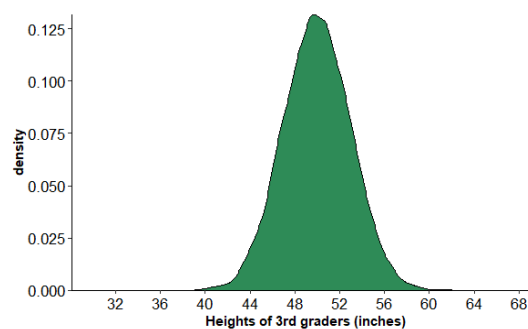With the caloric intake data, we'd find that the typical deviation from the mean is **282.96.**
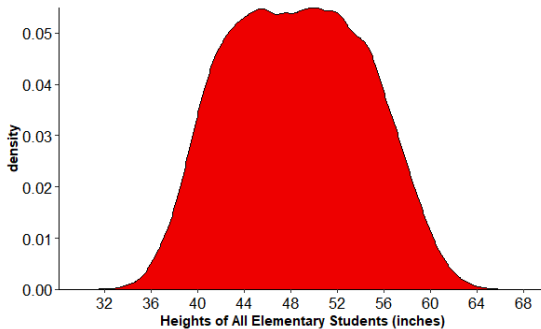
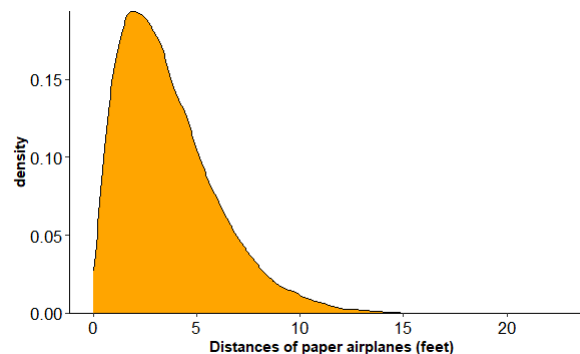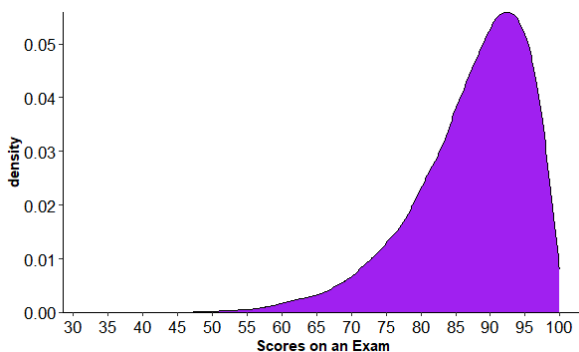## Identifying features of a distribution

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.**

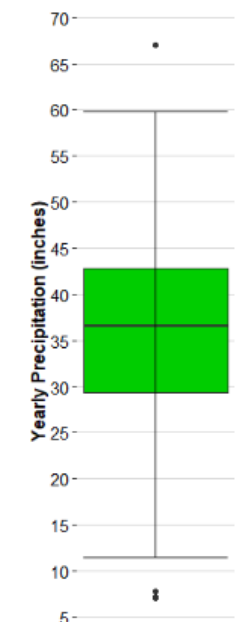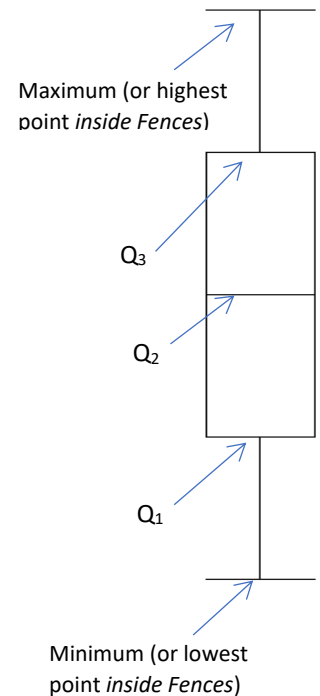Data that stretches to the right side may be called **right skewed** or **positively skewed.**
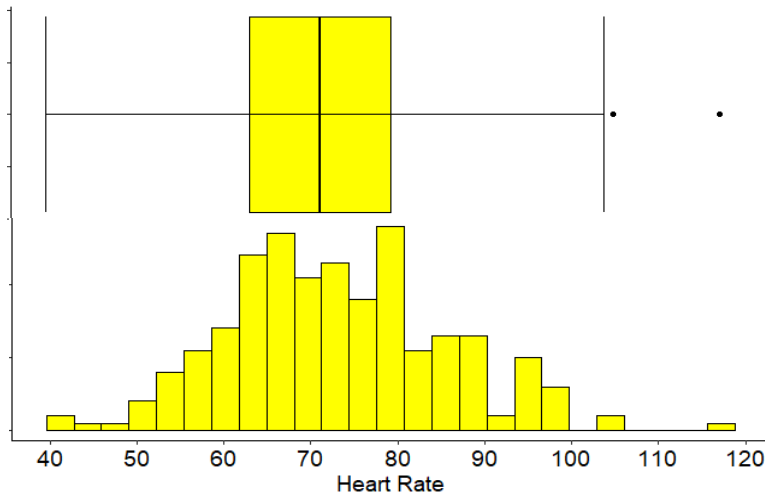
- Other Measures of Position
  - **Percentiles** are used to reference values at key positions in a distribution.
    - For example, the 20th percentile would be the value such that 20% of your data is at or below that point.
    - The median may also referred to as the 50th percentile.
  - Data descriptions 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.

Maximum (or highest point *inside Fences*)

Q_3

Q_2

Q_1

Minimum (or lowest point *inside Fences*)

Heart Rate

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.

Yearly Precipitation (inches)

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.
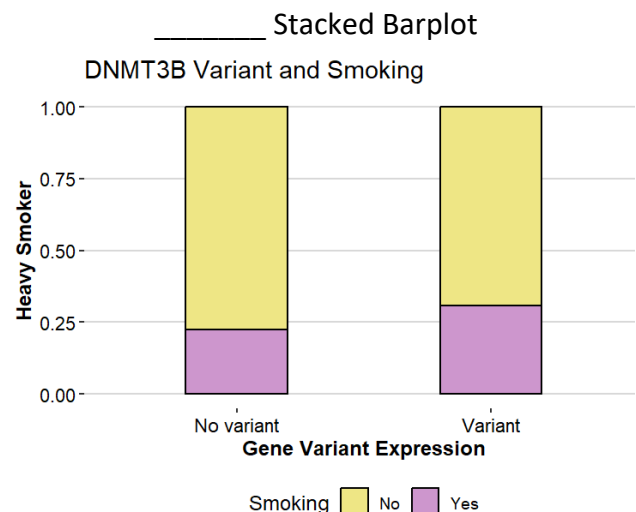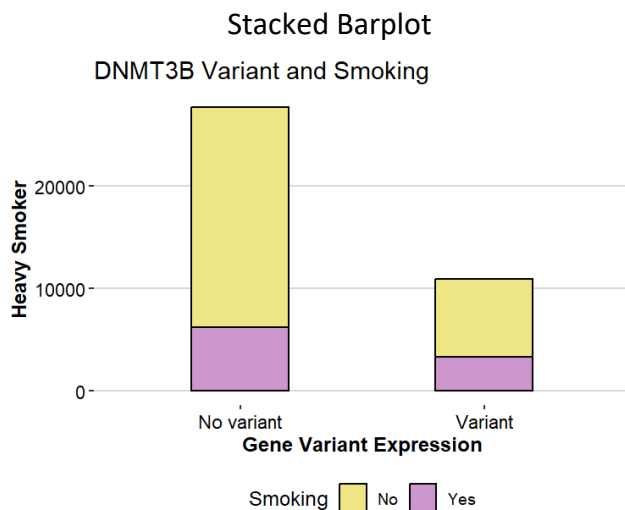
Unit of observation:

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…

Stacked Barplot
DNMT3B Variant and Smoking

_____ Stacked Barplot
DNMT3B Variant and Smoking

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.
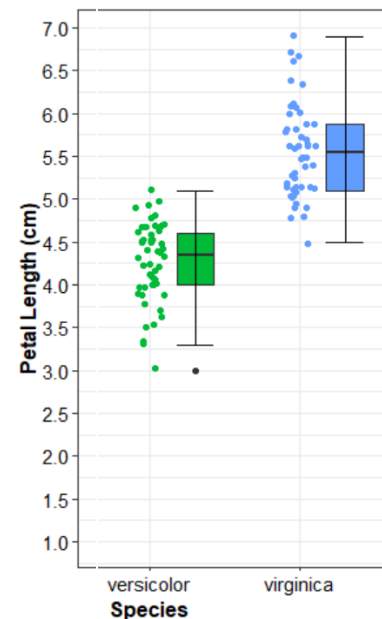
Unit of observation:

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…

Just using the graph, do you think the predictor in this investigation is explaining a lot of variability in the response variable?

*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 the predictor in this investigation is explaining a lot of variability in the response variable?
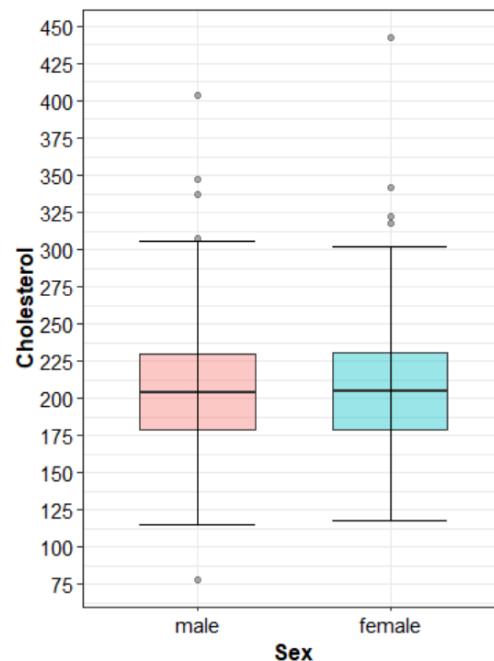
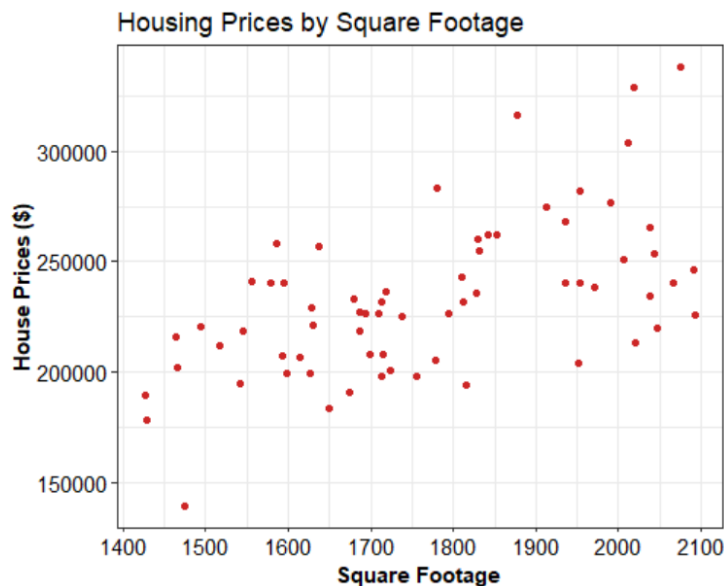**Measuring Association between Numeric Variables**

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 very simple measure of association.

Housing Prices by Square Footage

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).

Strong Negative	Weak Negative	Weak Positive	Strong Positive

-1.0   -0.5   0   +0.5   +1.0
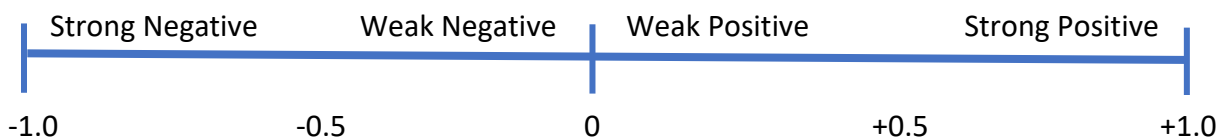
- **How would you calculate the correlation coefficient between two variables?**
  - In this class, you will **never** be asked to calculate r by hand from a set of data, but here is the formula!

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

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

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: How much are University of Illinois students spending on food each month?

Unit of Observation:

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

Statistic of interest:

This is best visualized with:

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

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):

Statistic of interest:

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):

Statistic of interest:

This is best visualized with:

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

Chapter 1 Reflection Questions

What is the difference between a population, a sample, and a unit of observation?

What is the difference between a statistic and a parameter? How are statistics and parameters related to (but still different from) a variable?

What are three different types of uncertainty we discussed that affect our ability to make population-level claims from a statistical investigation?

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?

What does a left skewed distribution look like?

In univariate investigations, what kinds of variables might we represent with a histogram? With a barplot?

In what data contexts might we use a proportion as our statistic of interest rather than a mean or median?

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

How are mean and median similar measures? How are they different, or in what situations would we expect them to be different?

What statistics comprise the "5-number summary" of a numeric variable?

Two different people track their caloric intake for 10 days. They both have the same mean caloric intake, but one person has a higher standard deviation. What does that tell us about their caloric intake from day-to-day compared to the first person?

What would be a more informal way of describing what the "standard deviation" measures?

In a multivariate investigation, how might we distinguish a response variable from an explanatory variable?

How might we symbolize a difference in two sample means? A difference in two sample proportions? With what variable pairings might we calculate each of these?

In a multivariate investigation, what kinds of variable pairings would I represent with a scatterplot? A stacked barplot? Side by side jitterplots or boxplots?

Chapter 1 Additional Practice (if you need it!)

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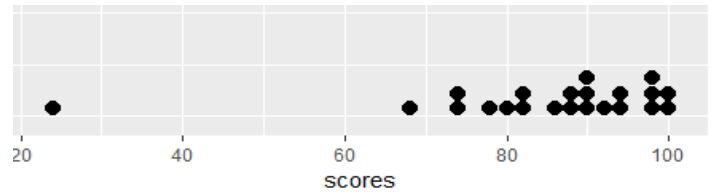Nominal, Ordinal, Discrete, or Continuous (circle one)

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)
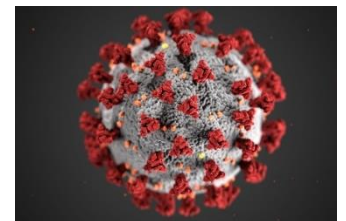
Statistic of interest:

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: