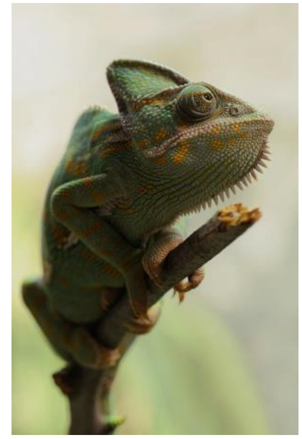
Chapter 1: Statistical Investigations



Investigation: Chameleons catch prey by extending their tongue quite some distance from their body. As a zoologist, perhaps you wish to study how far the typical adult chameleon can stretch their tongue to successfully catch a bug.
What data would you collect and *how* would you collect it?

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

Setting up a Statistical Investigation

- A statistical investigation hinges on the use of a sample of data to try to make a claim or draw insights about a larger population.
- **Identifying populations and units of observation**
    - A **population** would represent everyone/everything that we would ideally like to generalize toward.
    - A **unit of observation** would be one element or one case from that population. It might be a person, an animal, an object, a time point, a location, a group, etc.
    - In the investigation above…

Population: <u>All adult chameleons (a particular species, or all species?)</u>

Unit of Observation: <u>One chameleon</u>

- **Identifying a variable of interest and calculating an appropriate statistic**
    - A variable can be thought of as an <u>attribute</u> or a <u>feature</u> of the population that might vary with each observation
        - What information or measurements do we want to collect from each chameleon?
    - A statistic is some calculation we might make from our collected data.
        - Measures may simply be from one variable (a mean, a proportion)
        - Or they may be a function of multiple variable (a correlation coefficient, a ratio, the difference in means between two groups)
    - Let's think about how we might answer this in the chameleon example:

Variable of interest: <u>Distance it can extend its tongue (max of several trials? Tongue length of one? In what environment?)</u>

Statistic possibilities: <u>Mean? Median?</u>

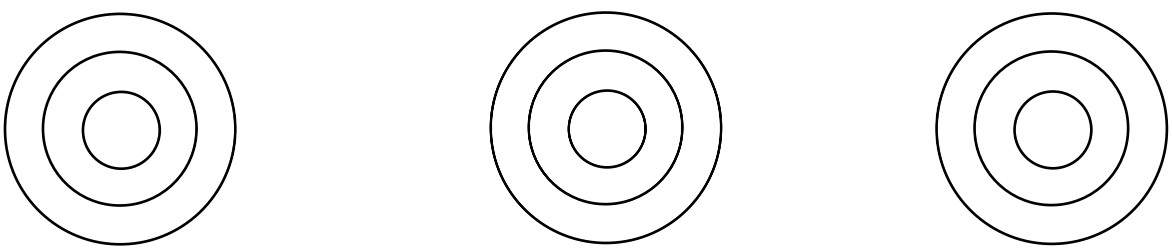
The Uncertainties Encompassing Statistical Investigations

- In a statistical investigation, we typically generate a statistic in hopes that it closely estimates a parameter, or at least helps us make a judgment about that parameter.
  - A parameter is some characteristic about the population.
  - Unless we're doing a census, we won't ever know population parameters.

- There will be uncertainty in the precision of our statistic
  - The attributes we are collecting will vary from each unit of the population (that's why we call them "variables"!).
    - Tongue extension varies from chameleon to chameleon
  - If we gather only a sample of data from the larger population, then our statistic will also vary with every possible sample we could take!

- There may also be uncertainty in the accuracy of our statistic
  - Bias could be introduced for several reasons, but a common reason might be a sampling bias. *Does my sample represent the population?*

- It's also good to broadly recognize the choices we make around instrumentation and statistical measures. Are the choices we made suitable to this investigation?
  - Researchers make choices about how they collect data and set up a study. *Would everyone have set up their chameleon study the same way?*
  - The statistics we generate may also be a matter of choice. *Is this the right statistic to answer this question?*

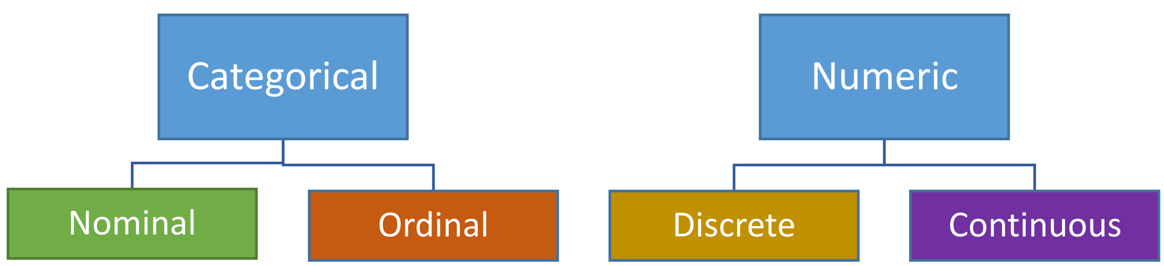


Read on your own

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 consistent 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?
    - 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?
    - ***Debatably Discrete:*** Rating customer service from 1 to 5 *(see note below though!)*
- **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?



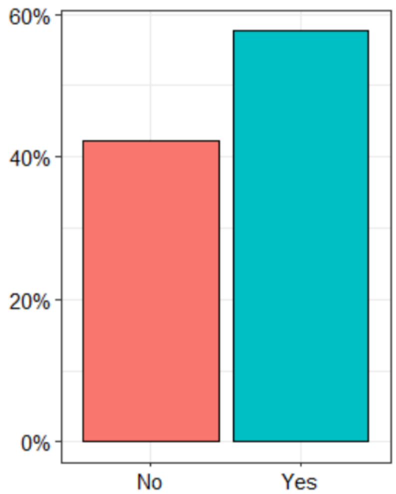
- Special cases of identifying types of variables
  - Binary variables would typically **not** be thought of as ordinal or discrete…that is because you can't have meaningful "ordering" with only two categories. We think of it as **nominal.**
  - Just because a variable is recorded numerically does ***not necessarily*** mean it is discrete/continuous. **Zip Codes**, or categories that have been arbitrarily numbered, may better be thought of as nominal if there is no inherent ordering to the numeric scale.
  - **Likert-scale items** (e.g., 1 to 5, 1 to 10 ratings) are sometimes considered ordinal. The reason being that the distance between a 1 and 2 may not be equal to the distance between 2 and 3. **In our course, we will call it discrete**, but analytical approaches vary on this one.

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

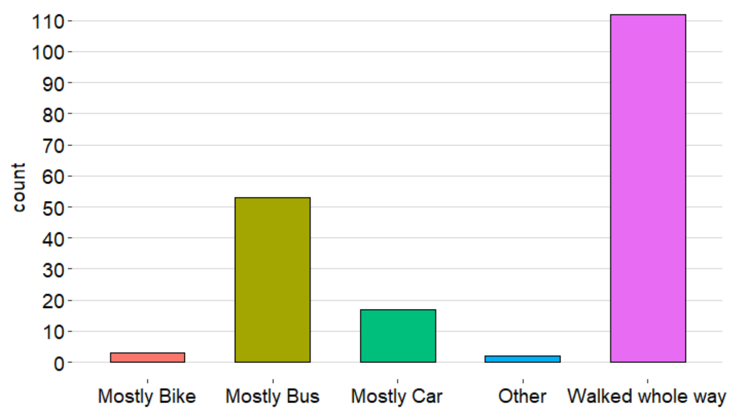
We asked students whether they have a pet at home

No	Yes
162	221



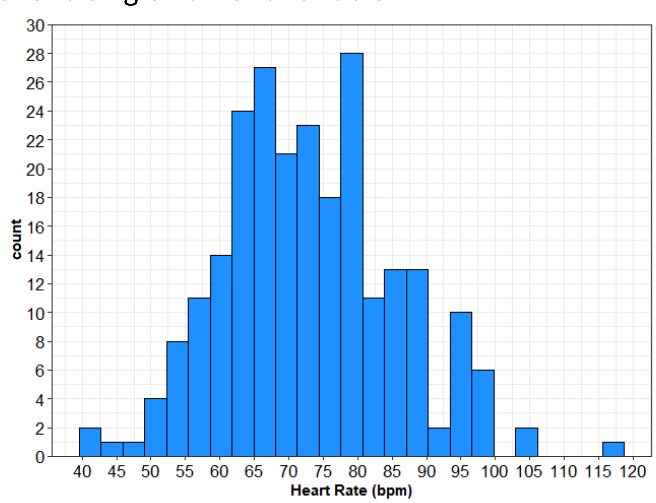
We asked students in January 2020 how they got to class that morning

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 <u>Categorical</u> 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: <u>All Illinois grad students</u>

Unit of Observation: <u>One Illinois grad student</u>

Variable of interest (and what type of variable?): <u>Whether or not tried marijuana product (categorical)</u>

Let's say that in this study, we contacted 54 graduate students. After being assured that their responses would remain anonymous, 22 of them answered yes.

- Measuring 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 <u>0 to 1.</u>
        - π is a <u>parameter</u>, representing a **population** proportion
        - $\hat{p}$ is a <u>statistic</u>, representing a **sample** proportion
    - A proportion is just the mathematical form of a percentage.
        - A proportion of 0.42 is the same as <u>42%</u>
        - A proportion of 0.894 is the same as <u>89.4%</u>

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

Do we know π in this investigation?

Draw a barplot to represent our findings!

Examining <u>Numeric</u> Data

Investigation: How might one complete a statistical investigation to better understand how many calories they typically consume per day?

Population of interest: <u>All days</u>

Unit of Observation: <u>One day</u>

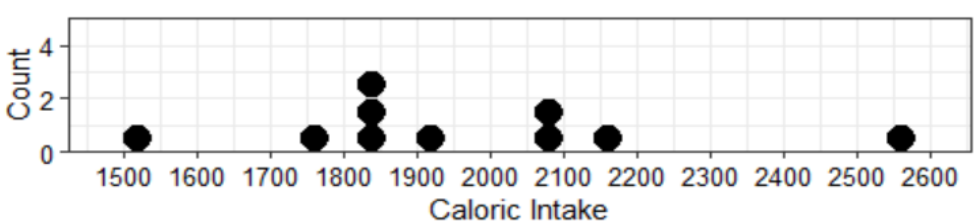
Variable of interest (and what type of variable?): <u>Daily caloric intake (numeric)</u>

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 <u>mean</u> 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 <u>parameter</u>, representing a **population** mean
    - $\bar{x}$ is a <u>statistic</u>, representing a **sample** mean
  - The <u>median</u> 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 our calorie counter wanted to ask a slightly different question? Perhaps they would like to know how consistent their caloric intake is—how much does their caloric intake vary from day to day?

- Measuring Numeric Data – **Measures of Variability**
  - The range is a very basic measure of variability. It simply measures the distance between the highest and lowest value.
    - The range of our sample caloric intake data is: 2570 – 1520 = 1050
  - The standard deviation is a more common and "robust" measure. Think of it as measuring the *typical* deviation from the mean.
    - σ is a parameter, representing a population standard deviation
    - s is a statistic, representing a sample standard deviation

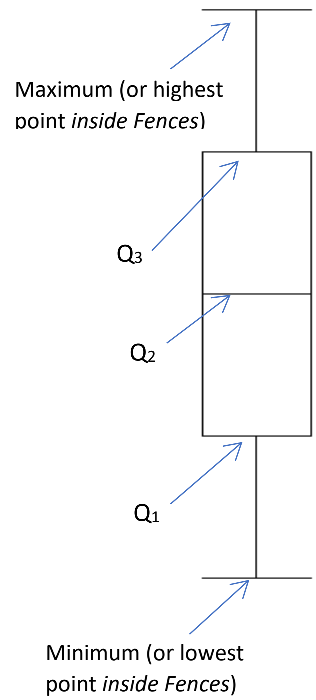
$$\sigma = \sqrt{\frac{\sum(x_i-\mu)^2}{N}} = \sqrt{\frac{(x_1-\mu)^2+(x_2-\mu)^2+\ldots+(x_N-\mu)^2}{N}}$$

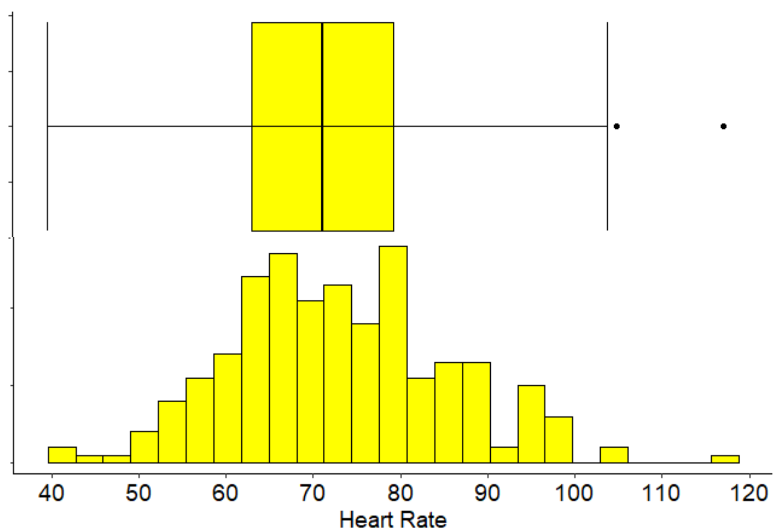
$$s = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}} = \sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n-1}}$$

In the caloric intake example, our standard deviation is 282.96. This tells us that

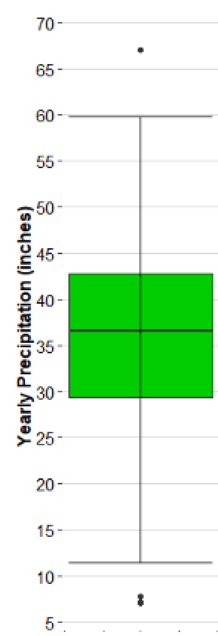
The **typical** deviation from the mean, per day, is about **282.96 calories**.

- Measuring Numeric Data – **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 <u>50th</u> 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)
- <u>Boxplots</u> 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 a 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?

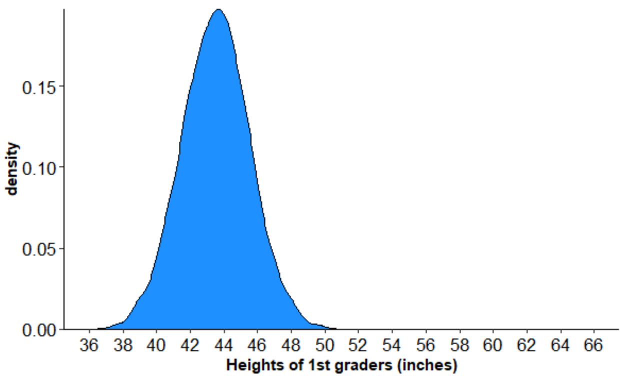


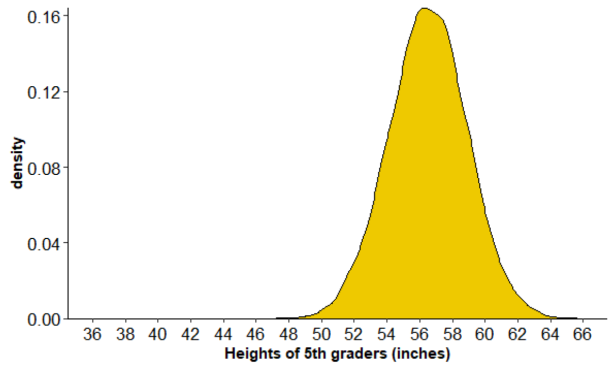
What proportion of cities see at least 43 inches of rain a year?

Read on your own

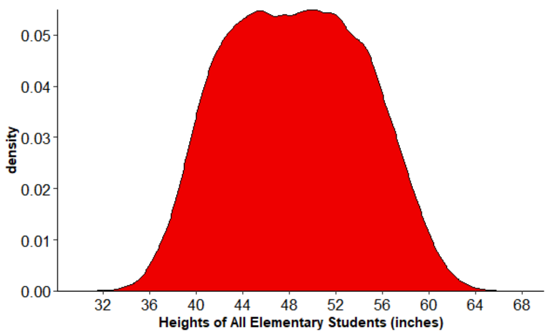
Identifying features of a distribution

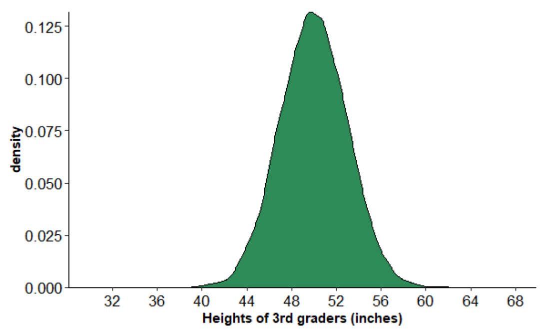
<u>**Center**</u>**:** Where is the "middle" of the distribution?



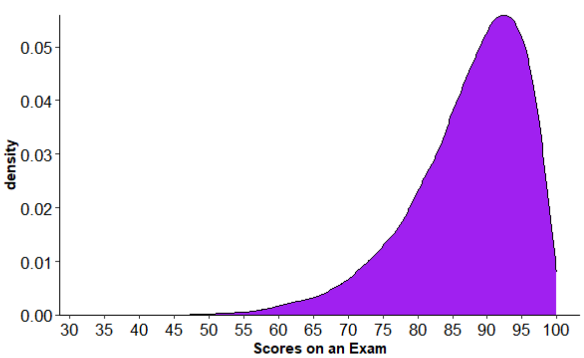


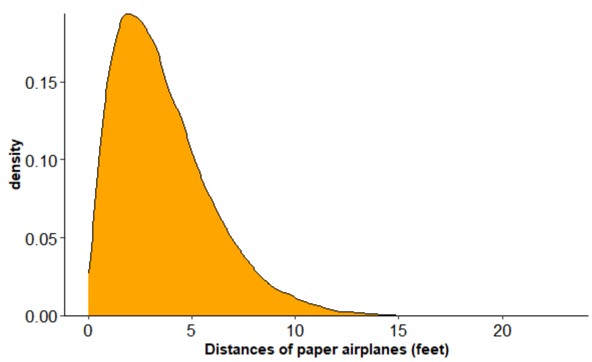
<u>**Variability**</u>: How far do data points typically extend from the center?





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



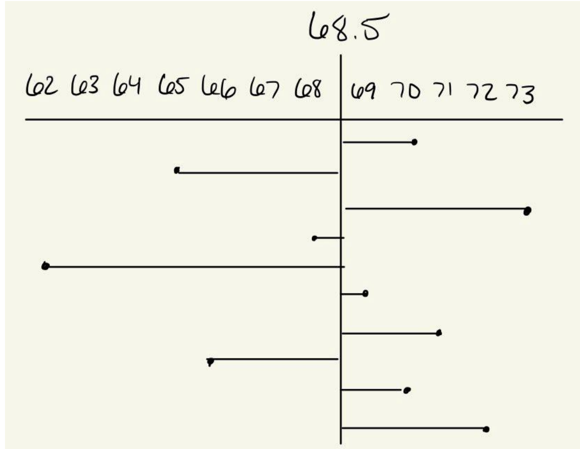


- We refer to the direction of the skew as the direction in which the distribution is thinly stretching out.
  - Data that stretches to the left side may be called **left skewed** or **negatively skewed.**
  - Data that stretches to the right side may be called **right skewed** or **positively skewed.**

Unpacking the standard deviation formula

Consider the following dataset representing the heights in inches of 10 high-school boys. Their mean height is 68.5 inches, but how far do their heights vary on average? Let's first visualize each boy's distance from the mean by showing how far their height value is from 68.5.

Rodrigo	70
Stan	65
Jeremy	73
Justin	68
David	62
Nick	69
Mickey	71
Anay	66
J. T.	70
Morgan	72



To find average deviation from the mean, we could take the average distance! But we'll need to use absolute value signs to ensure all differences are positive.

$$\frac{|x_1-\mu|+|x_2-\mu|+\ldots+|x_N-\mu|}{N} = \frac{|70-68.5|+|65-68.5|+\ldots+|72-68.5|}{10} = 2.7 \text{ inches}$$

This measure is called "mean absolute deviation" or MAD for short. It works quite well as a measure of variability! But sometimes absolute values can pose problems in higher mathematics.

As a result, the field of statistics developed around a very similar measure known as "Variance." Instead of taking the mean absolute deviation, we take the mean squared deviation.

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

Squared deviations keep the differences positive, but change the scaling now. But if we take the **square root of the variance**, we've essentially converted our measure back into the units of variable!

The variance should come out to be 10.45, then the **standard deviation** will be $\sqrt{\mathbf{10.45}}$ = 3.23 inches!

Above and Beyond: *Why divide by n-1 with "s" formula?*

We call n – 1 the "degrees of freedom" for this measure. If we're using our sample data to also estimate μ with $\bar{x}$, then we "lose" a degree of freedom. We only have n – 1 unique pieces of information to estimate σ with s!

If my sample size was 1, and I need to estimate μ with $\bar{x}$, then I have 0 pieces of information left to estimate the *variability* in my data. **One point can't vary from itself!** With 2 data points, I now have 1 piece of information about the variability. The denominator accounts for this.

Multivariate Investigations

- **Univariate vs. Multivariate Investigations**
  - **Univariate Questions:** Ask about characteristics of... one variable in isolation
  - **Multivariate Questions:** Ask about the... relationship between two variables

- Identifying a response variable
  - A **response variable** is a variable that we have an interest in better understanding or predicting. It is an outcome of interest.
  - 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...Exam score

The explanatory variable is...Whether one attends class or not

Comparing Proportions

Investigation: A study 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. https://www.sciencedaily.com/releases/2017/10/171010124112.htm

Unit of observation: One Adult (in one of these countries)

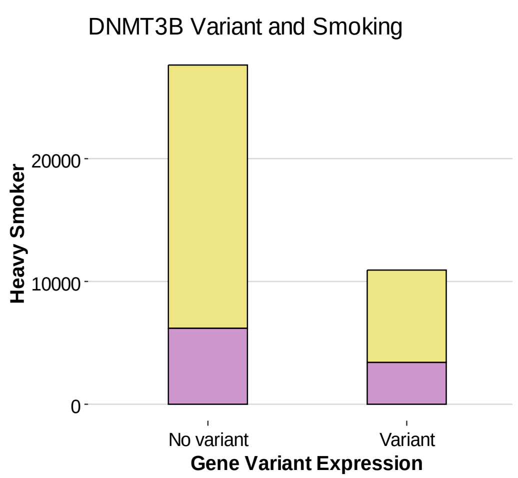
Response variable (and type): Being a heavy smoker (categorical)

Explanatory variable (and type): Having gene variant (categorical)

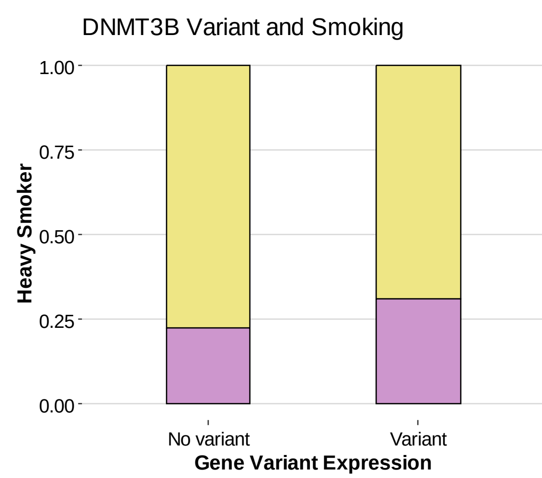
Parameter of interest: $\pi_1 - \pi_2$

Statistic of interest: $\hat{p}_1 - \hat{p}_2$

Stacked Barplot

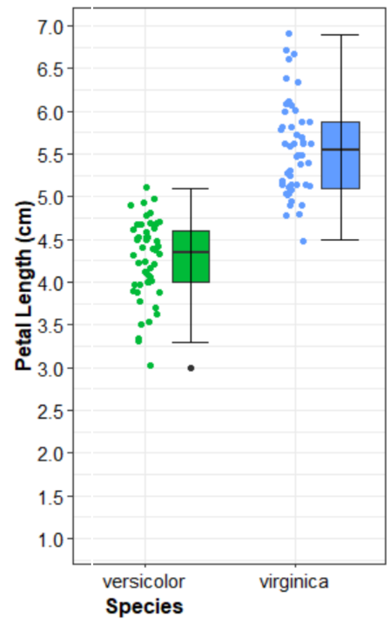


100% Stacked Barplot



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 certain iris species might be higher on average than others.



Unit of observation: One plant

Response variable (and type): Petal length (numeric)

Explanatory variable (and type): Species (categorical)

Parameter of interest: $\mu_1 - \mu_2$

Statistic of interest: $\bar{x}_1 - \bar{x}_2$

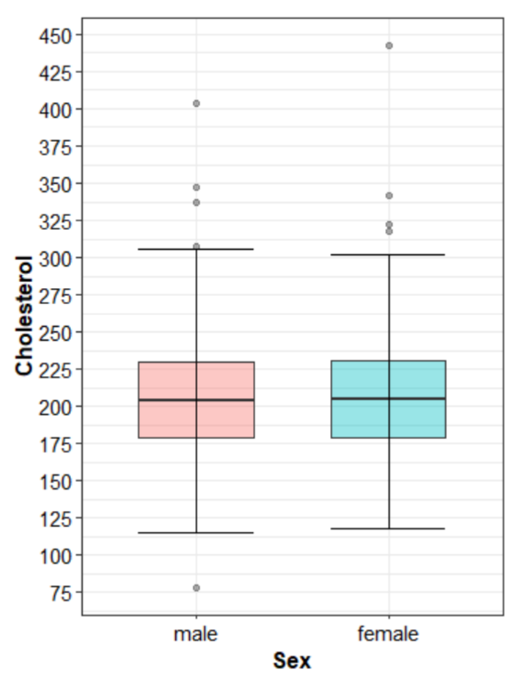
Just using the graphs and summary statistics here, do you think the predictor in this investigation is explaining a lot of variability in the response variable?

	Versicolor	**Virginica**
Mean	4.26	5.55
SD	0.470	0.552

*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.

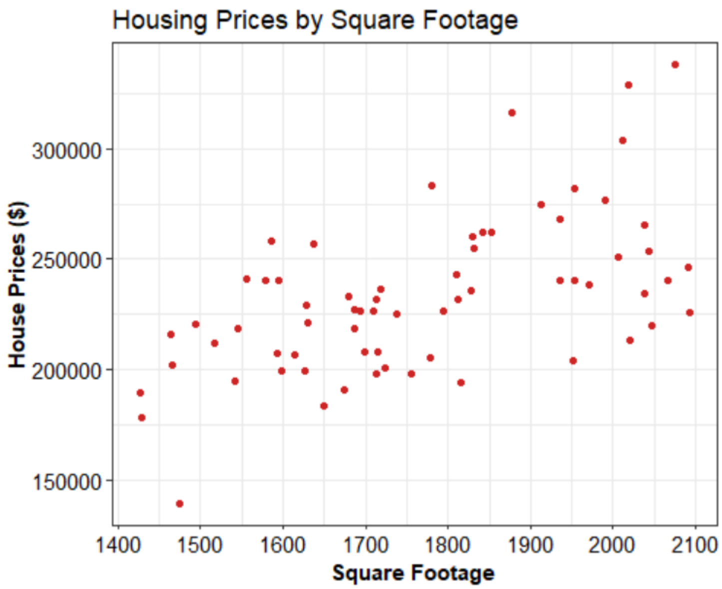
	Male	**Female**
Mean	207.5	208.3
SD	45.5	43.7



Just using the graphs and summary statistics here, do you think the predictor in this investigation is explaining a lot of variability in the response variable?

Measuring Correlation

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: One house (pop being all houses in community)

Response variable (and type)

Price (numeric)

Explanatory variable (and type)

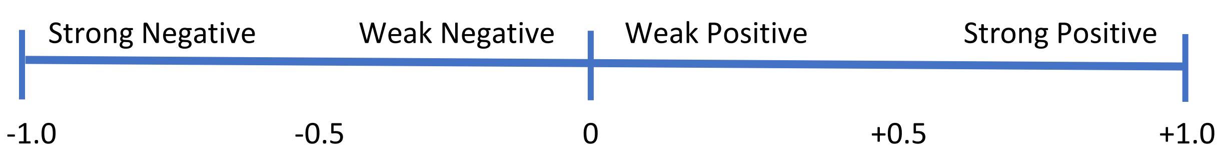
Square footage (numeric)

Statistic of interest

We need a new way to measure this!

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

- Pearson's Correlation Coefficient (r)
  - A statistic between -1 and +1 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!

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 NON-LINEAR relationship.

Practice! For each investigation below, identify 1) what the unit of observation is, 2) what variable or variables are involved (and what type of variable you think you would collect for each), 3) what statistic we might use complete your investigation, and 4) what type of graph might be helpful to represent the data. *Answers may differ based on how you might approach your investigation—your choices may vary!*

Investigation: Do higher temperatures increase the metabolism rate for humans? *Note that metabolism levels are commonly measured using a calorimeter, which calculates how many calories you burn at rest based on your breathing.*

Unit of Observation: One human

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

Metabolism rate (numeric) as predicted by temperature (numeric)

Statistic of interest: r (correlation coefficient)

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: One dating app user

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

Discrepancy in height listed from real height (numeric) as predicted by gender (categorical)

Statistic of interest: $\bar{x}_1 - \bar{x}_2$

This is best visualized with:

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

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

Unit of Observation: One Illinois student

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

Money spent on food (numeric)

Statistic of interest: Not explicitly stated, but could try sample mean or median

This is best visualized with:

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

Chapter 1 Additional Practice

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 the full report on Gallup's poll to Americans on gun policy :
https://news.gallup.com/poll/268016/americans-stricter-laws-gun-sales.aspx

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)

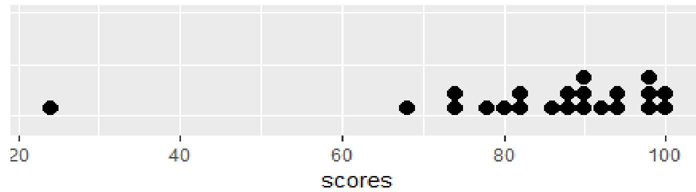
Judges score musicians across a number of different criteria using four choices: "superior," "excellent," "good," or "needs work."
Identify the variable of interest: ____
Nominal, Ordinal, Discrete, or Continuous (circle one)

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

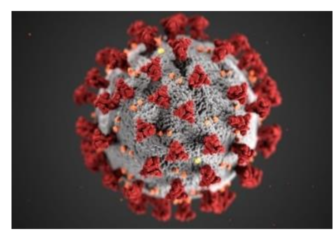


If we removed that score of 24, which value do you think would be most affected: mean or median?

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:

Chapter 1 Learning Goals

After this chapter, you should be able to…

- Identify statistical investigations as those that try to draw an insight about a population using a sample of data—inferring a parameter from a statistic
- Recognize the uncertainties that pervade statistical investigations and what causes them:
    - Uncertainty in the **precision** of our statistic (is this only a sample of the population?)
    - Uncertainty in the **accuracy** of our statistic (does our sample have a bias?)
    - Uncertainty in the **suitability** of our statistic and the choices we made as researchers (is this design or approach suitable for this investigation?)
- Distinguish terms associated with a statistical investigation: population, sample, unit of observation, variable, statistic, parameter
- Distinguish different types of variables that we might collect from each unit in our investigation: Categorical (Nominal or Ordinal) and Numeric (Discrete and Continuous)
- Complete a univariate investigation involving a categorical variable
    - Calculate and interpret a proportion
    - Recognize π as a population proportion and $\hat{p}$ as a sample proportion
    - Use a barplot to visualize a categorical variable and identify the proportion
- Complete a univariate investigation involving a numeric variable
    - Interpret the mean, median, and standard deviation of a numeric variable
    - Recognize μ, M, and σ as parameters, and $\bar{x}$, m, and s as statistics
    - Use a histogram to identify features of a numeric variable visually, including center, variability, and skewness
    - Use a boxplot to identify the distribution of a numeric variable through summary values (the 5-number summary)
- Identify what kind of investigation should be done based on the variable(s) involved
    - Identify a univariate investigation as having only one variable involved, and a multivariate investigation as the examination of how multiple variables relate to each other
    - In multivariate investigations, distinguish between the response variable (the target outcome) and the explanatory variable (the predictor)
    - Distinguish different statistics/visualizations that are appropriate for different types of variables comparisons
- Complete a multivariate investigation involving two categorical variables
    - Use a stacked barplot to visually compare proportions from two or more groups
    - Calculate a difference in proportions $\hat{p}_1$ - $\hat{p}_2$, as an estimate for the true difference in proportions at the population level π_1 - π_2
- Complete a multivariate investigation involving one numeric and one categorical variable
    - Use side by side boxplots or jitter plots to visually compare the numeric distributions of two groups
    - Visually identify differences in center or variability
    - Recognize whether an investigation might best be answered in comparing differences in center (mean or median) or differences in variability (standard deviation)
    - Recognize $\bar{x}_1$ - $\bar{x}_2$, as an estimate for the true difference in means for the population: μ_1 - μ_2

- Recognize s_1 - s_2 as an estimate for the true difference in standard deviations for the population: σ_1 - σ_2

- Complete a multivariate investigation involving two numeric variables
  - Use a scatterplot to visually compare the relationship between two numeric variables
  - Identify the correlation coefficient (r) of our sample as an estimate of the true correlation (ρ) for the entire population