Chapter 1: Exploring Data

Introducing Statistics

- Doing Statistics can mean many different things.
    - **Exploring a Variable:** What values are typical? How much do values vary? What "shape" does the data make?
    - **Modeling Relationships:** Often, we want to understand the variation in data by finding what factors "explain" that variation. Finding predictors can help us better predict and understand the variable of interest.
    - **Testing Relationships:** Once we identify possible explanations, we might complete a statistical test to determine if the relationship we see in our data could be explained as random chance, or if it suggests a true underlying difference.
    - **Making Meaning:** Statistical work must also seek to understand why things relate the way they do. *Why* do these variables relate? Is one causing changes in the other, or associated with another factor that does? What implications does this have?

The best way to understand what statistics is about is to actually do a short statistical investigation!

Example: Consider how long it takes performers to sing the National Anthem before the Super Bowl. What does it look like to do statistics with this variable? What questions might we pose here?

Performer	Year	Time (seconds)
Mickey Guyton	2022	111
Eric Church / Jazmine Sullivan	2021	143
Demi Lovato	2020	109
Gladys Knight	2019	121
Pink	2018	113
Luke Bryan	2017	124
Lady Gaga	2016	129
Idina Menzel	2015	124
Renee Fleming	2014	114
Alicia Keys	2013	155
Kelly Clarkson	2012	94
Christina Aguilera	2011	114
Carrie Underwood	2010	107
Jennifer Hudson	2009	130
Jordin Sparks	2008	114
Billy Joel	2007	90

Vocabulary for talking about Data

- **Observational Unit vs. Variable**
  - An _____ represents one row of our spreadsheet. Observational units may be a person, a city, a community, a store, or something else!
  - A _____ is a characteristic of interest we gather from each observation through a question or measurement. Each variable will *typically* be one specific column of your spreadsheet. *Note: we call it a "variable" because the responses can vary. Not everyone provides the same response!*
- **Population vs. Sample**
  - A _____ is the entire group we have an interest in learning more about and making an inference toward. *The UIUC student body might be a population I'd like to know more about!*
  - A _____ is a subset of a population. *Maybe I can't reach everyone in the UIUC student body, but I could reach out to a subset of students!*

- **Parameter vs. Statistic**
  - A _____ is a numeric value that describes some characteristic about the population *What is the average amount of money spent on food each month for the UIUC student body?*
  - A _____ is a numeric value that describes some characteristic about a sample. *What is the average amount of money spent on food each month for the 100 students I surveyed?*
- **n** = shorthand for noting the size of a sample.

Practice: 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 observational unit is…_____ n =_____

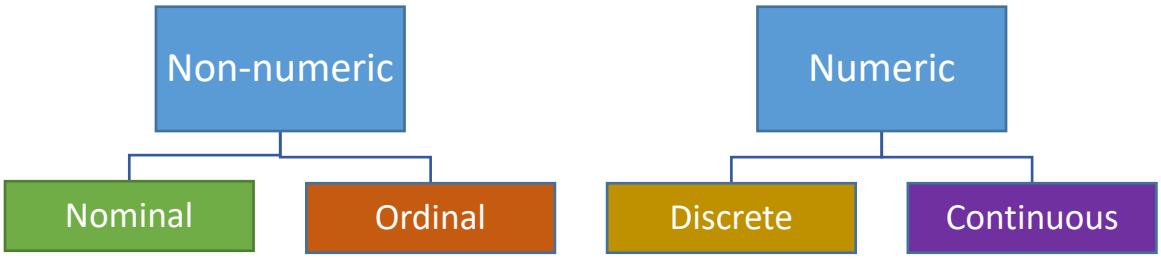
Our variable of interest is…_____

The sample statistic we gathered is…_____

Do we know what the population parameter is?

In case you are curious, 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

Identifying Different Types of Data



- **Nominal Data**
  - Data that falls into non-numeric categories that don't have any _____
    - In what places have you experienced pain since your knee surgery?
    - What fruits do you like to eat?
    - Does this state require photo ID to vote in elections?

- **Ordinal Data**
  - Data that falls into categories that have a meaningful ordering (but not on a _____ 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 Data**
  - Data that follows a numeric scale, but only takes limited values (like _____). These are typically things that are countable.
    - What year of school is this for you? (notice this is similar to the Freshman, sophomore,… question, but now it has a clear numeric scale).
    - 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?
    - Likert-scale items *if* presented *numerically* to survey-takers

- **Continuous Data**
  - Numeric and _____ (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 data
    - Binary data would typically **not** be thought of as discrete…that is because you can't have meaningful "ordering" with only two categories. It would always be nominal.
    - Just because data is numeric does *not necessarily* mean it is discrete/continuous. _____, or numbered items in which the order bears no meaning, may better be thought of as nominal.

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."

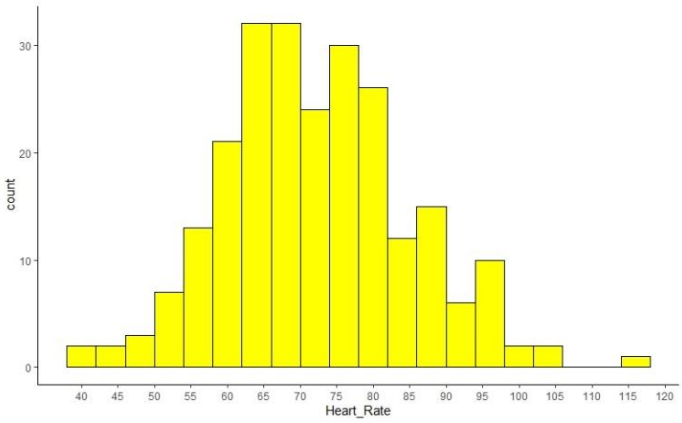
Identify the variable of interest: _____

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

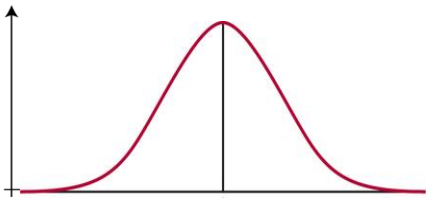
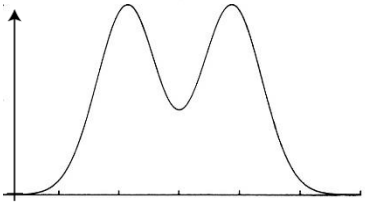
- **Distributions – Visualizing Numeric data**
  - In a previous semester, we asked students to record the number of times their heart beat in one minute. The results are presented below.

65	78	90	75	80	95	95	78	94	64	70	66	74	76	81	84	64	74	80	68	78	70
70	68	70	56	80	82	72	68	74	98	62	71	90	74	78	64	70	74	74	72	67	72
64	78	73	50	95	96	76	55	68	88	86	68	53	76	70	76	75	98	84	87	61	64
64	83	84	63	78	64	66	70	62	51	76	92	64	60	88	65	94	56	78	66	72	82
62	56	61	98	60	46	72	80	86	72	54	76	58	70	54	69	54	65	82	104	90	80
72	78	62	98	90	56	94	81	60	76	87	96	70	41	80	66	70	55	90	73	81	70
66	78	64	80	80	76	45	99	81	63	62	60	70	84	60	80	56	54	82	70	80	74
66	80	60	74	65	58	90	80	80	90	88	58	77	64	86	68	58	99	85	76	60	76
68	75	70	86	82	78	80	63	73	60	50	70	54	50	70	72	74	73	84	62	65	63
68	80	67	75	62	66	42	76	80	56	66	96	88	70	94	75	60	66	72	68	60	74
76	64	78	90	67	105	117	59	90	60	66	92	70	64	70	58	64	78	82	84		

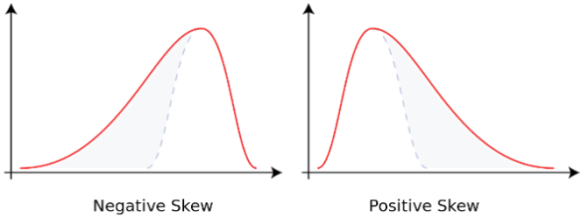
  - It's hard to make sense of a list of raw data—let's visualize it instead!
  - **Histograms** are a common representation to represent a numeric variable. The following histogram below represents the resting heart rate of a class of approximately 120 students.
  - The **variable** is represented in the **x axis**, and the *y-axis* represents a "*count*": how many observations are in each particular bin.



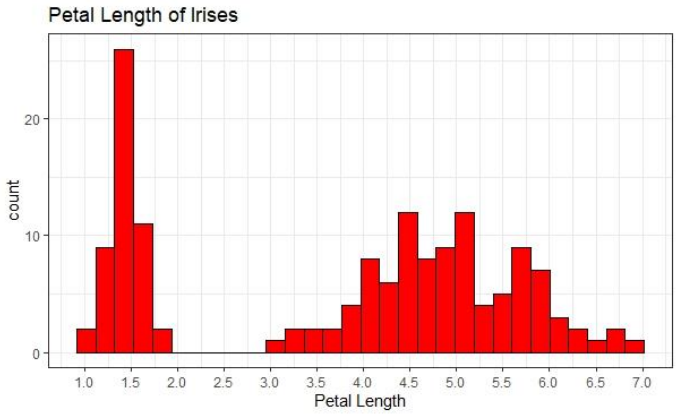
  - **Symmetry vs. Skewness**
    - Many distributions are fairly symmetrical, meaning that they balance at some central point. Consider these different examples below of symmetric distributions.


    - But many distributions may instead be **Skewed,** meaning that the data is stretched out sparsely to one direction while being heavily concentrated on the other side.
    - Identifying the direction of skew:
      - Data that skews to the left side may be called "left" skewed or "negative" skewed.
      - Data that skews to the right side may be called _____ or _____



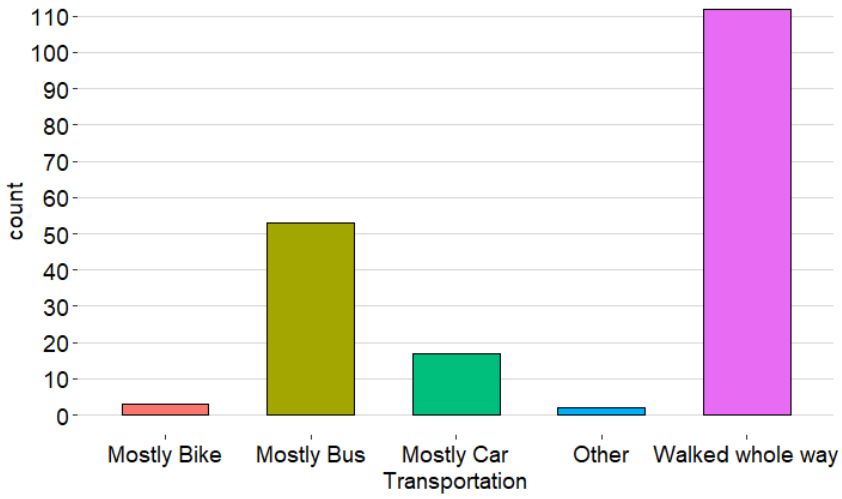
- **Mode:** The most common data value, or where the distribution _____
  - Distributions are commonly unimodal (one peak), but consider this distribution below representing petal lengths of iris plants in a particular dataset.
  - This picture is an example of a bimodal (2 peak) distribution. *Why might a distribution separate like this?*



- **Distributions – Visualizing Non-numeric Data**
  - In a previous semester, we asked students how they got to class that morning. But this time, their responses were not numeric, but a list of categorical responses where they could choose one. Their options were:
    - ❖ Mostly Bus
    - ❖ Mostly Car
    - ❖ Mostly Bike
    - ❖ Walked whole way
    - ❖ Other

Mostly Bike	Mostly Bus	Mostly Car	Walked Whole Way	Other
3	53	17	112	2

  - Bar plots are the most common representations of categorical data.
  - Much like histograms, the y axis represents the count for how many observations fit each category.



Numerically Summarizing Data

- We can also calculate statistics from our data that quickly describe characteristics about our sample

65	78	90	75	80	95	95	78	94	64	70	66	74	76	81	84	64	74	80	68	78	70
70	68	70	56	80	82	72	68	74	98	62	71	90	74	78	64	70	74	74	72	67	72
64	78	73	50	95	96	76	55	68	88	86	68	53	76	70	76	75	98	84	87	61	64
64	83	84	63	78	64	66	70	62	51	76	92	64	60	88	65	94	56	78	66	72	82
62	56	61	98	60	46	72	80	86	72	54	76	58	70	54	69	54	65	82	104	90	80
72	78	62	98	90	56	94	81	60	76	87	96	70	41	80	66	70	55	90	73	81	70
66	78	64	80	80	76	45	99	81	63	62	60	70	84	60	80	56	54	82	70	80	74
66	80	60	74	65	58	90	80	80	90	88	58	77	64	86	68	58	99	85	76	60	76
68	75	70	86	82	78	80	63	73	60	50	70	54	50	70	72	74	73	84	62	65	63
68	80	67	75	62	66	42	76	80	56	66	96	88	70	94	75	60	66	72	68	60	74
76	64	78	90	67	105	117	59	90	60	66	92	70	64	70	58	64	78	82	84		

- **Measures of Center**
  - **The Mean** (The Average)
    - The mean may be thought of as the _____
    - If you think about the set of all heart rate values in a population, there does exist a true population mean. This measure would be represented as $\boldsymbol{\mu}$ **(mu).**
    - However, we only have a sample of heart rate measurements from the larger population, so the value that we could actually calculate is a sample mean $\bar{x}$ **(x-bar)**
    - The formula to calculate the mean: $\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$
    - Where x_1 = the 1st observation in our data set. x_2 = the 2nd observation…
    - The sample mean for our heart rate data comes out to **73.075**

  - **The Median** (The Midpoint)
    - The median represents the value of the middle or average person/unit. If you arrange all observations in numeric order, then the median is the point such that 50% of the data falls at or below that value and 50% falls at or above that value.
    - If we sorted the heart rate data and found the middle data value, it would be **72**.

Practice: Consider this smaller set of data. What is the Mean? What is the Median?

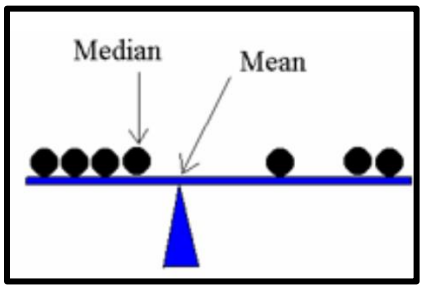
12 17 21 22 23 25 25 27 32

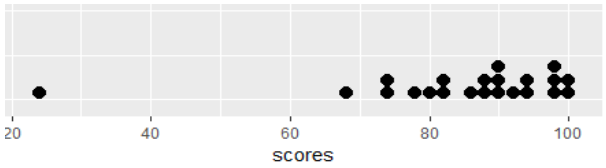
What if a data set has 10 numbers instead of 9? (larger question, what if a data set has an even number of observations instead of an odd number of observations?). We now find the median as the average of the two middle values.

12 17 21 22 23 25 25 27 32 35

- **Mean vs. Median**
  - The mean is responsive to *all* data points. Changing just one value, especially on either extreme, will change the mean!

$$13, 45, 78, 96 \;:\; \bar{x} = 58 \quad and \; M = 61.5$$
$$2, 45, 78, 96 \;:\; \bar{x} = 55.25 \; and \; M = 61.5$$





Practice: Consider the following 22 scores for a recent test.

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

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

- **Quartiles – Measures of Position**
  - Quartiles are three numbers that partition a data set into 4 approximately equal parts.
  - Q_1 is the 25th percentile. Also think of it as the median of the *lower* half of the data
  - Q_2 is the 50th percentile. Which also makes it the _____ of the entire set of data!
  - Q_3 is the 75th percentile. Also the median of the *upper* half of the data
  - In the case where there is an odd number of data points, we will use an "inclusive" method where the median data value is included in both the lower half and upper half when calculating Q_1 and Q_3
  - 5-Number Summary
    - **The 5-number summary**: (Minimum data point, Q_1, Q_2, Q_3, maximum data point)
    - This is a common set of numeric summaries to quickly get a sense of where your data falls.

Find the 5-number summary of test score data ___, ___, ___, ___, ___

- **Measures of Variability**
  - In addition to identifying various measures of center in our data, we can also measure how much data varies.

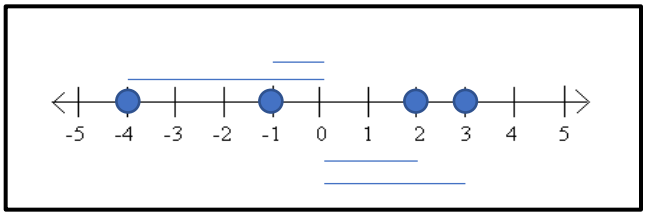
  - **Range – Distance from Maximum to Minimum**
    - Simply take the Maximum value – the Minimum Value
    - What is the range of the test score data? _____
    - Range can be greatly affected by outliers though, and will also be wider for larger sample sizes. It's not a great measure for comparing variability across groups!
  - **IQR (Inter-Quartile Range) – Distance from the 75th to 25th percentile**
    - IQR is considered a more *robust* measure of variability, meaning that it is unlikely to be thrown off by more extreme values or differences in sample size.
    - The IQR of your data is calculated by taking $Q_3 - Q_1$

Between what two test scores does the middle 50% of the class fall between?

What is the Interquartile Range for the class' test scores?

  - **Mean Absolute Deviation from the Mean (MAD)**
    - Another measure of variability we might be interested in is the average deviation from the mean in our dataset.
    - In other words, we want to know, on average, how far are all of the data points away from the mean? If we picked a person at random, how far off from the mean should we expect their numeric result to be?

Consider the following dataset representing the heights in inches of 10 high-school boys. What is the *mean absolute deviation from the mean* (MAD) in this dataset?



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

- **Standard Deviation and Variance**
  - While the MAD is a very intuitive measure of variability, it is not considered a standard method in statistics due to the absolute value not actually being an algebraic operation.
  - Instead, statisticians use a formula that does something similar. Squaring values also ensures that distances come out as positive values.

    The **Variance** is: $\sigma^2 = \frac{\sum(x_i-\mu)^2}{n} = \frac{(x_1-\mu)^2+(x_2-\mu)^2+\ldots+(x_n-\mu)^2}{n}$
  - Note that Σ = sum (as in, keep repeating this operation for all data values).
  - This represents the _____ from the mean.

- **Standard Deviation** represents the *"typical deviation"* from the mean.

  Standard deviation is calculated: $\sigma = \sqrt{\frac{\sum(x_i-\mu)^2}{n}} = \sqrt{\frac{(x_1-\mu)^2+(x_2-\mu)^2+\ldots+(x_n-\mu)^2}{n}}$

  Why do we need both "variance" and "standard deviation"?

  - Standard deviation is a more practical measure to directly *interpret* since it is scaled in the units we are measuring. *Easier to interpret!*
  - Variance is a simpler measure (no square root operation needed) and is often a measure of variability within more complex calculations. *Easier in mathematics!*
  - For the heart rate data, the standard deviation comes to 12.62.

- **Parameter vs. Statistics:** σ and σ^2 represent population parameters for standard deviation and variance while s and s^2 represent sample statistics
  - The formula for the sample statistics s and s^2 have a different _____: instead of "n", it is "n-1"

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

Why divide by n-1?

We call n – 1 the "degrees of freedom" for these measures. That is because you need two data points before you have 1 piece of information about variability in each case. If dividing by n, we will consistently **underestimate** the true variability.

They said the weather around here was lovely, on average. I should have asked about the standard deviation!

- **Describing Non-numeric data**
  - When data falls into non-numeric categories, we might choose to measure what proportion fit into different bins.
  - Instead of calculating a mean, we calculate a special case of center called a **proportion.** This represents the number of cases that fit a category of interest out of the total.
    - p is a _____, representing the proportion of a population in a category
    - $\hat{p}$ is a _____, representing the proportion of a sample in a category
  - Consider the traveling to school data from earlier. What **proportion** of this sample of students walked the whole way to school?

Mostly Bike	Mostly Bus	Mostly Car	Walked Whole Way	Other
3	53	17	112	2

  - You might think about a proportion as a special case of a mean, where our data is binary (0's and 1's).

  - This also helps us think about measuring variability. The variance of binary data can be calculated with our earlier formula, but plugging in 0's and 1's. Or we can take a shortcut formula and calculate the following:
    - $\sigma = \sqrt{p(1-p)}$
    - $s = \sqrt{\hat{p}(1-\hat{p})}$

If our data were all 0's or all 1's, what would s come out to be?

If exactly half of our data was 0's and half were 1's, what would s come out to be?

Practice: Consider the National Anthem data from the first page, but instead re-record the data categorically as whether the performance took more than 2 minutes or not.

What is the standard deviation in responses when evaluating in this particular binary form?