Week 1 - Introduction to Data and Summarization

Ryan Miller

Week #1 Outline

- Video #1
  - Data, the field of statistics, and types of variables
- Video #2
  - Summaries and graphs for quantitative variables
- Video #3
  - Summaries and graphs for categorical variables
- Video #4
  - Contingency tables

Data

Question 1: What percentage of the world's 1-year-old children have been vaccinated against at least one disease?

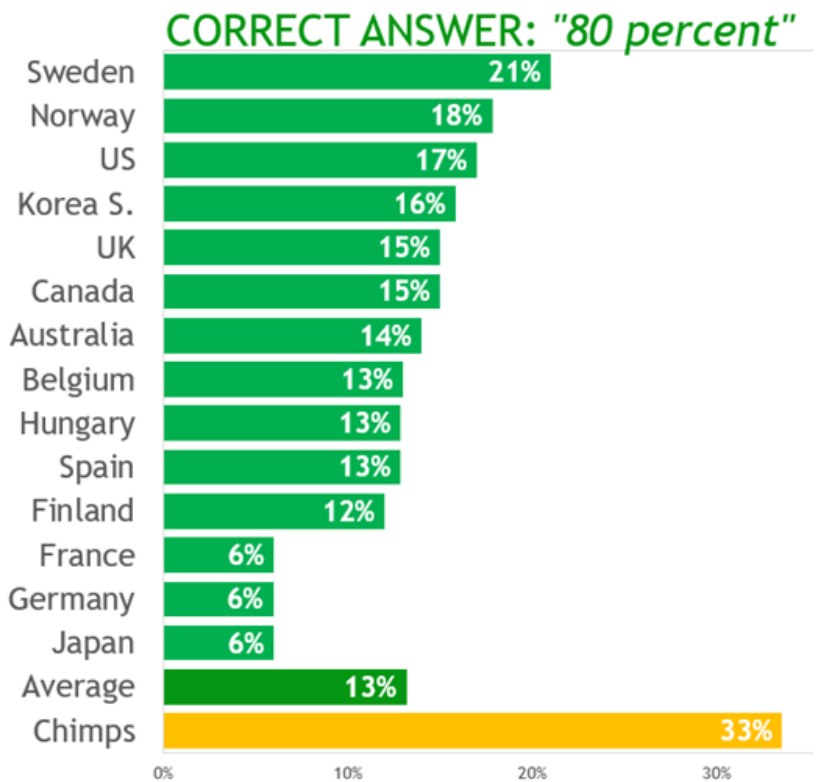
A) 20%
B) 50%
C) 80%

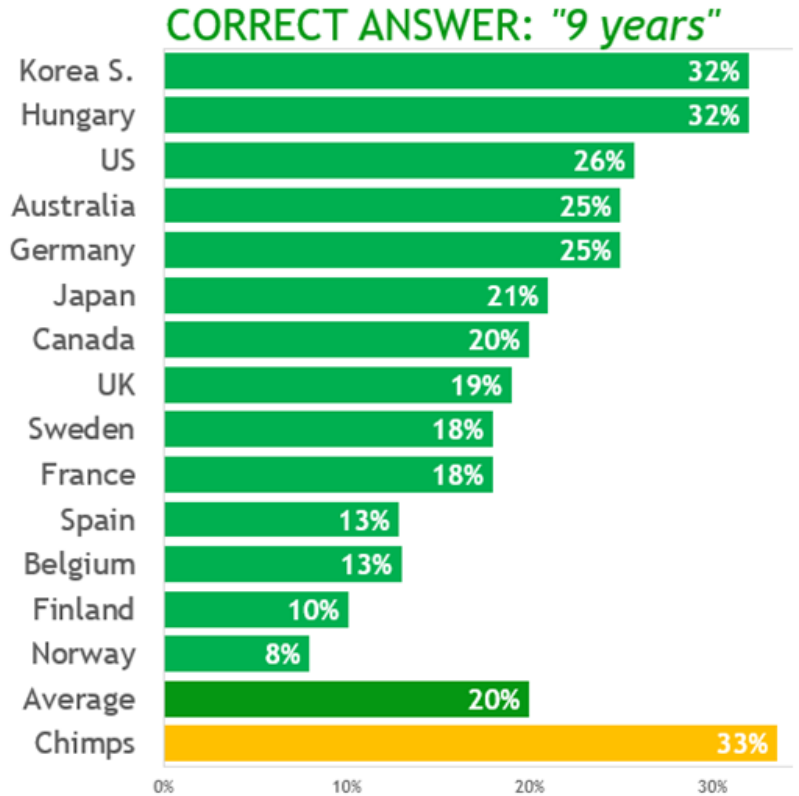
Question 2: Worldwide, 30-year-old men have 10 years of schooling, on average. How many years do women of the same age have?

A) 3 years
B) 6 years
C) 9 years

Data

Here's what the data show:





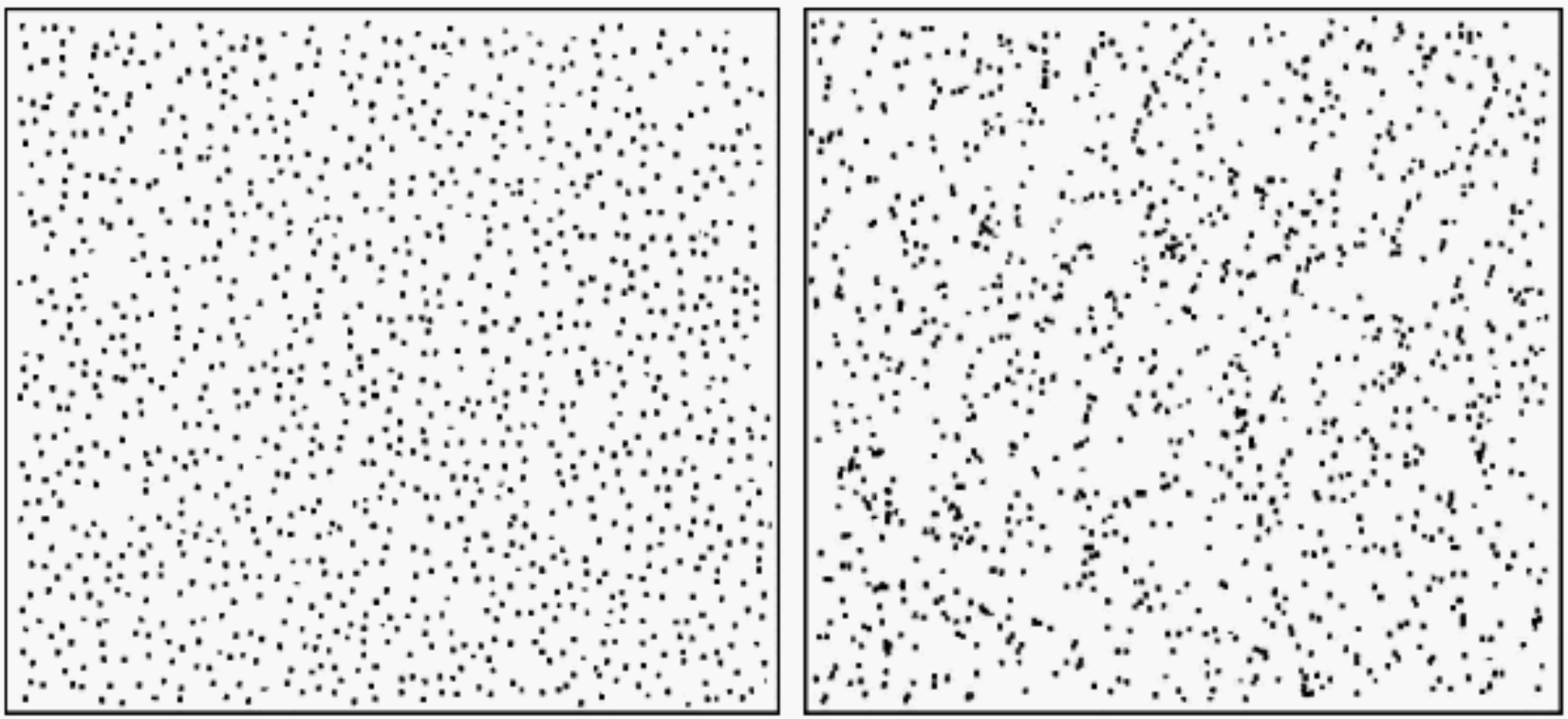
Source: Allan Rossman's JSM talk

Data

- The world has made remarkable progress in the last 20 years
  - Due to biases and a lack of exposure to quality data, most people aren't away of this
- Data empowers us to *objectively understand reality*

What about statistics?

- In most situations simply having data isn't enough, humans are too good at finding non-existent patterns
  - Which panel do you think displays randomly generated data?



What about statistics?

- Statistics is often defined as the science of *understanding uncertainty*
  - More specifically, it's a way of thinking combined with collection of tools and methods that can be used to understand uncertainty in order to make judgements about the world

What about statistics?

- Statistics is often defined as the science of *understanding uncertainty*
  - More specifically, it's a way of thinking combined with collection of tools and methods that can be used to understand uncertainty in order to make judgements about the world
  - ie: What can we learn from experiment that involved only 30 people? How accurately can a poll of 1000 registered voters predict an election?

What about statistics?

- Statistics is often defined as the science of *understanding uncertainty*
  - More specifically, it's a way of thinking combined with collection of tools and methods that can be used to understand uncertainty in order to make judgements about the world
  - ie: What can we learn from experiment that involved only 30 people? How accurately can a poll of 1000 registered voters predict an election?
- But before we can get to answer these questions, we need to learn the vocabulary of Statisticians

Vocabulary

- **Case**: the subject/object/unit of observation
  - Usually data is organized so that each case is represented by a *row* (but not always!)
- **Variable**: any characteristic that is recorded for each case (generally stored in a *column*)

Vocabulary

- **Case**: the subject/object/unit of observation
  - Usually data is organized so that each case is represented by a *row* (but not always!)
- **Variable**: any characteristic that is recorded for each case (generally stored in a *column*)
- **Categorical Variable**: a variable that divides the cases into *groups*
  - **Nominal**: many categories with no natural ordering
  - **Binary**: two exclusive categories
  - **Ordinal**: categories with a natural order
- **Quantitative Variable**: a variable that records a *numeric* value for each case
  - **Discrete**: countable (ie: integers)
  - **Continuous**: uncountable (ie: real numbers)

Practice

1) Download and open the "Happy Planet" dataset from our course website or this link
2) Identify the cases
3) What type of variable is "Population"?
4) What type of variable is "Region"?

Practice (solution)

- Each case is a country
- "Population" is a quantitative variable, it is measured in millions of people (a numeric entity)
- "Region" is categorical variable, it divides the cases into 7 geographic groups (categories)

Grey Areas

Sometimes there are situations where a variable is technically one type, but it more useful to analyze it as if it were another. For example:

- "Year" might be a discrete quantitative variable, but if the data only contain 2 or 3 years we might treat it is as categorical
- A Likert Scale question is be an ordinal categorical variable, but we might translate it into numeric scores and treat it is a quantitative

Grey Areas

Sometimes there are situations where a variable is technically one type, but it more useful to analyze it as if it were another. For example:

- "Year" might be a discrete quantitative variable, but if the data only contain 2 or 3 years we might treat it is as categorical
- A Likert Scale question is be an ordinal categorical variable, but we might translate it into numeric scores and treat it is a quantitative

"An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem." - John Tukey (Statistician, 1915-2000)

Closing Remarks (Data and Statistics)

- The first step in any statistical analysis is to understand the big picture aspects of the data you are working with
  - Identifying the cases and variable types will enable you to choose the proper analytic methods for your specific scenario
  - In the next couple of weeks, we'll cover the *summary statistics* and *graphs* that statisticians use when working with certain types of variables

Motivation

- Shown below are a few *quantitative variables* from the "Tips" dataset, but how useful is this display?

TotBill	Tip	Size
13.37	2.00	2
17.29	2.71	2
7.51	2.00	2
11.35	2.50	2
10.07	1.25	2
14.00	3.00	2
10.33	2.00	2
11.17	1.50	2
24.52	3.48	3
27.05	5.00	6
20.27	2.83	2
12.03	1.50	2
44.30	2.50	3
13.27	2.50	2
21.16	3.00	2
15.01	2.09	2
22.76	3.00	2
16.47	3.23	3
17.31	3.50	2
15.42	1.57	2

Summarization

- **Raw data** can be difficult to make sense of
- **Summarization** refers to methods that simplify raw data into a more understandable form
  - Ideally, we can summarize a variable using one number, or a small set of numbers, in order to make informed judgments

Summarization

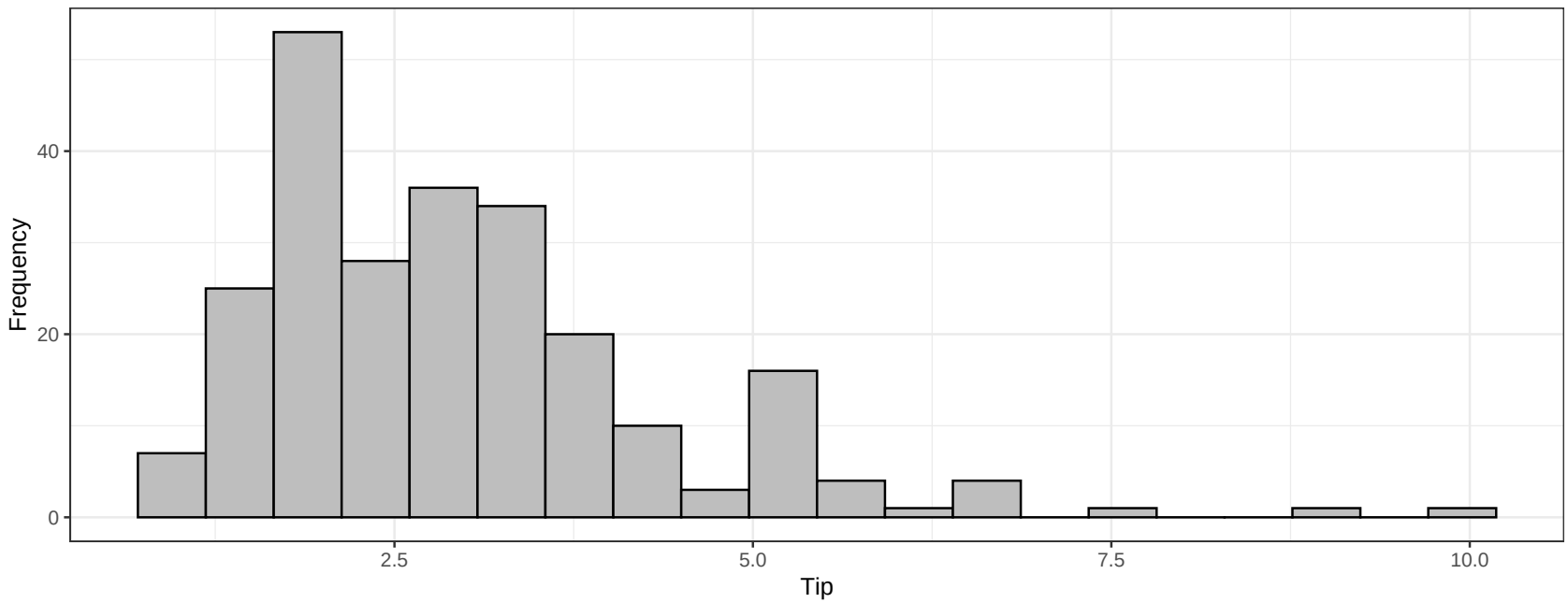
- **Raw data** can be difficult to make sense of
- **Summarization** refers to methods that simplify raw data into a more understandable form
  - Ideally, we can summarize a variable using one number, or a small set of numbers, in order to make informed judgments
- For now, we'll focus on **univariate summaries**, or those involving only a single variable
  - Later we'll start dealing with more interesting stuff involving multiple variables

Distributions

- Before getting to far into summarization, we need to introduce the idea of *distributions*
  - A variable's **distribution** describes *values that are possible* and *how frequently they occur*

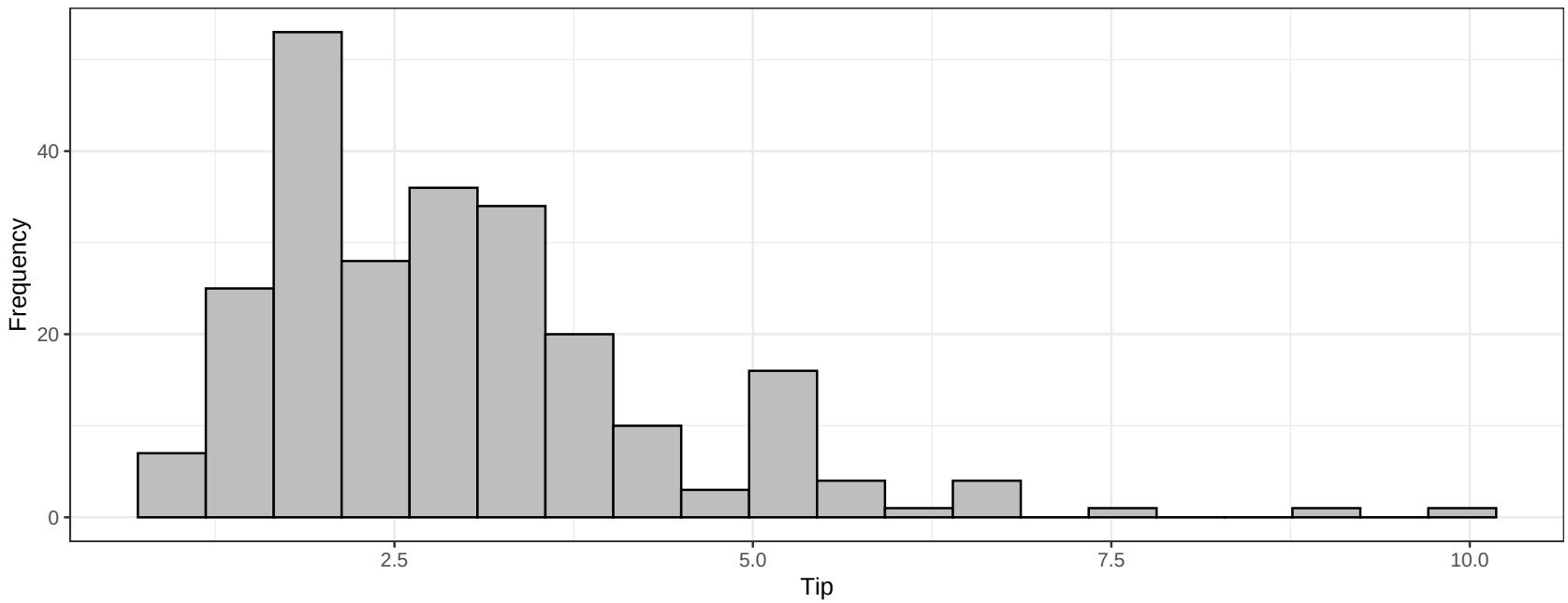
Distributions

- Before getting to far into summarization, we need to introduce the idea of *distributions*
  - A variable's **distribution** describes *values that are possible* and *how frequently they occur*
- Below is a **histogram**, one way of showing a distribution of a quantitative variable



Histograms

- A histogram works by dividing the quantitative variable of interest into **bins**, or equal length intervals
  - The number of cases that belong to each bin are graphed on the y-axis
- Notice how \$2-3 tips are most common, larger tips of \$5+ do occasionally occur, tips over \$10 almost never occur

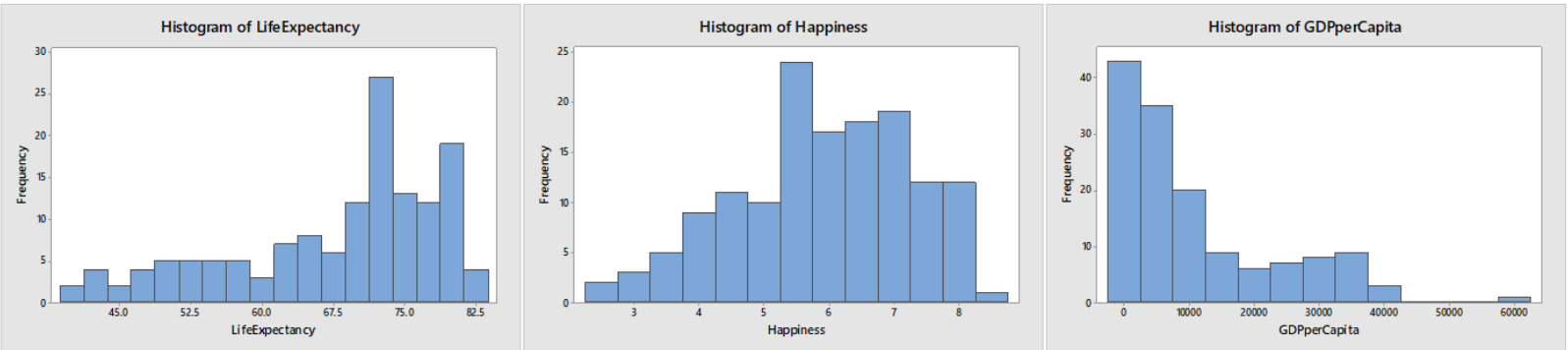


Judging Shape from Histograms

- The first noteworthy characteristic of a variable's distribution is its shape
  - A distribution is **symmetric** if it can be folded over a center line with both sides roughly matching each other

Judging Shape from Histograms

- The first noteworthy characteristic of a variable's distribution is its shape
  - A distribution is **symmetric** if it can be folded over a center line with both sides roughly matching each other
- A distribution is **skewed** if most of the data is piled up in one area and there's a long tail containing smaller amounts of data in the opposite direction



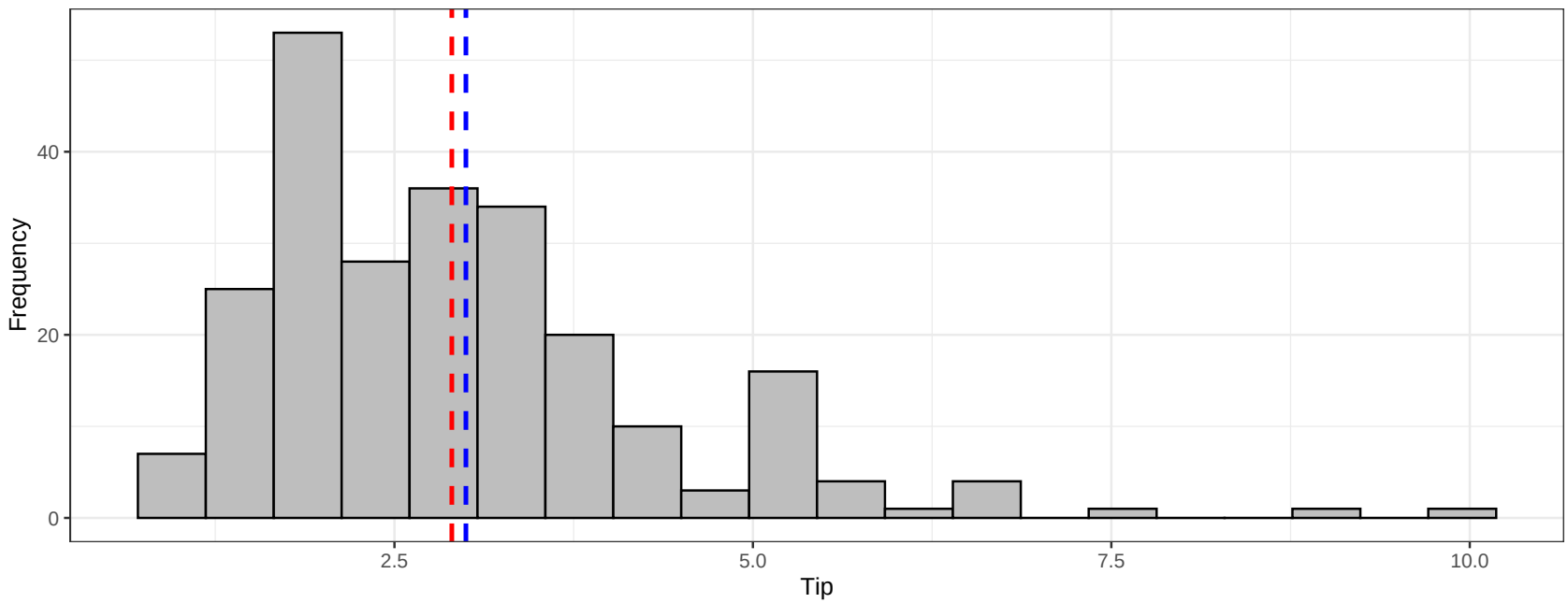
The Mean

- Distributions aren't a summary, but they can help us understand the purpose of summarization
- The **mean**, or arithmetic average, is way of describing the *center of a distribution*, or its *central tendency*
  - The mean can provide us a sense of what is typical for a quantitative variable

$$\text{Mean} = \frac{\text{Sum across all cases}}{\text{Number of cases}}$$

The Median

- Another way approach to describing the center of a distribution is the **median**, or the midpoint if the variable's values were arranged from smallest to largest
- The histogram below shows the mean tip (blue) and the median tip (red)
  - Why is the mean larger?



Mean vs. Median

- The median is considered a *robust* measure of the center of a distribution because it is not heavily influenced by *extreme values* known as *outliers*
  - The table below demonstrates the impact of adding a 100-dollar tip to the Tips dataset

	Mean	Median
Original	3.0	2.90
With $100 tip	3.4	2.96

Other Summaries

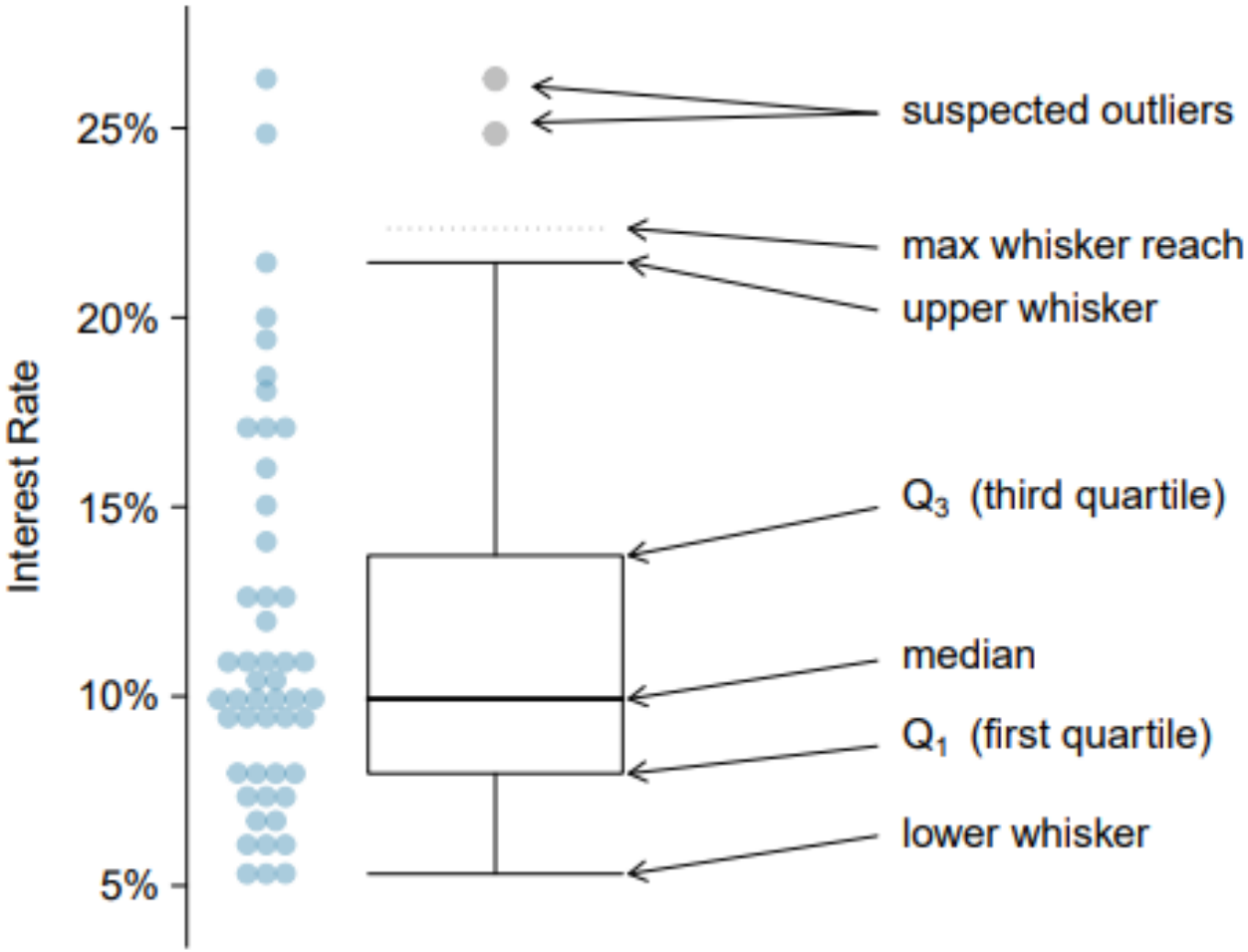
- Very often we'd like to know more about a variable than simply the center of its distribution
- The **minimum** and **maximum** are self-explanatory summaries of the variable's most extreme values

Other Summaries

- Very often we'd like to know more about a variable than simply the center of its distribution
- The **minimum** and **maximum** are self-explanatory summaries of the variable's most extreme values
- **Percentiles** describe a cutoff value for which P data falls below
  - The median is the 50^{th} percentile
  - The 25^{th} and 75^{th} percentiles are called the **first quartile**, or Q1, and the **third quartile**, or Q3

Boxplots

- The summary measures presented on the previous slide can be used to construct a visualization known as a **boxplot**



Spread

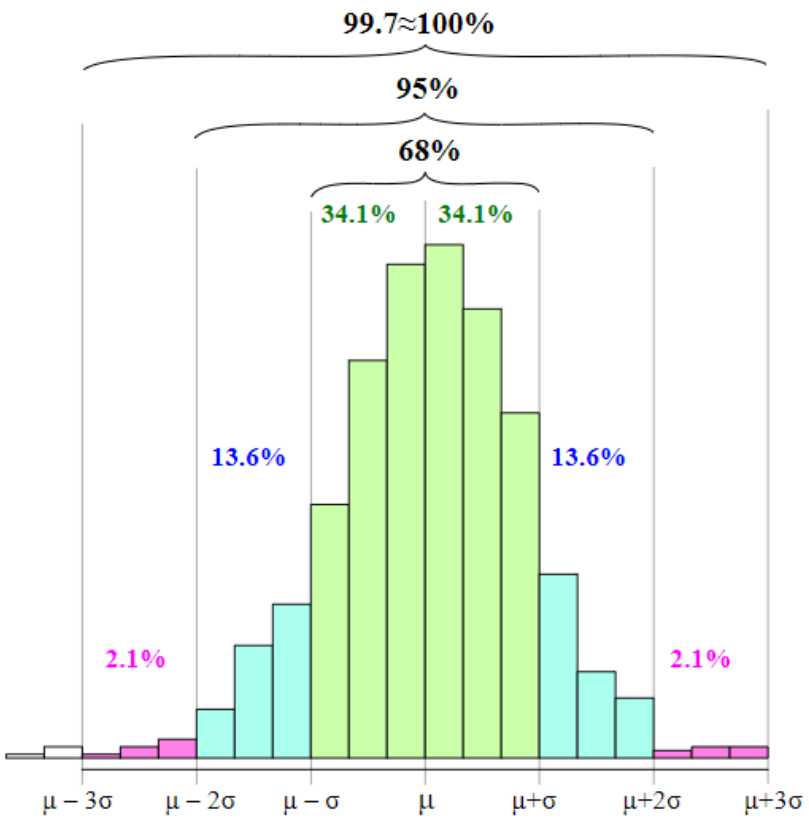
- The mean and median summarize the *center* of a distribution
- It is also useful to summarize the *spread*, or how the data values tend to vary around the center

Spread

- The mean and median summarize the *center* of a distribution
- It is also useful to summarize the *spread*, or how the data values tend to vary around the center
  - The **range** is the difference between the minimum and maximum
  - The **interquartile range**, or **IQR**, is the difference between the third and first quartiles (Q1 and Q3)

Standard Deviation

- The most widely used measure of spread is the **standard deviation**, which roughly corresponds to the *average distance of each data-point from the mean*
- For *bell-shaped distributions*, the standard deviation is related to the percentage of cases within a certain distance from the mean



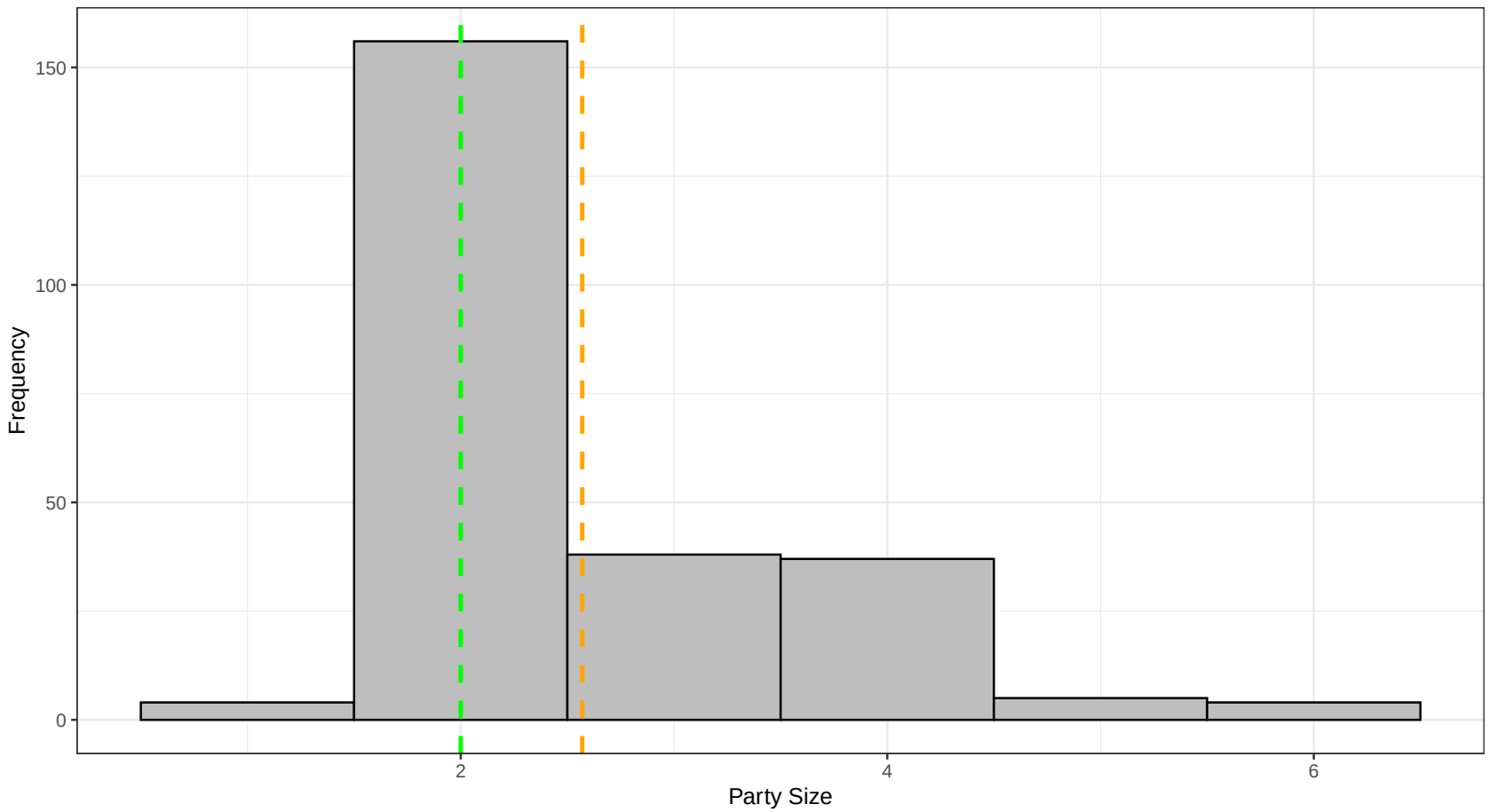
Standard Deviation vs. IQR

- Similar to how the median is more robust to outliers than the mean, the IQR is more robust than the standard deviation

	Mean	Median	StDev	IQR
Original	3.00	2.9	1.38	1.56
With $100 tip	3.37	2.9	6.35	1.56

Practice

Using the graph below, answer the following: 1) What is the name of this graph? 2) How many bins are displayed? 3) Which color line marks the mean and which marks the median? 4) Is this variable's distribution *skewed* or *symmetric*?



Practice (solution)

1) Histogram
2) 8 bins (note that one of them has zero cases in it)
3) green = median, orange/yellow = mean
4) Skewed right (mean $>$ median, long tail of larger values)

Closing Remarks (Quantitative Variables)

- Understanding a quantitative variable is inherently tied to understanding its distribution
  - *Histograms* and *boxplots* provide a visual display of the distribution
  - The *mean* and *median* describe the *central tendency*
  - The *standard deviation* and *IQR* describe the *spread* or variability

Summarizing Categorical Variables

- Categorical variables tend to be much simpler to summarize (relative to quantitative variables)
- The only summaries we'll consider are *frequencies* and *proportions*
  - **Frequencies** are a tally of how many cases belong to a particular category
  - **Proportions** are the fraction of the total cases that belong to a particular category

Frequency Tables

Frequencies are typically displayed for *all categories* of a variable in a **frequency table**

Day	Frequency
Fri	19
Sat	87
Sun	76
Thu	62

Dividing each frequency by the total number of cases (244 for this dataset) yields *proportions* for each category

Day	Proportion
Fri	0.08
Sat	0.36
Sun	0.31
Thu	0.25

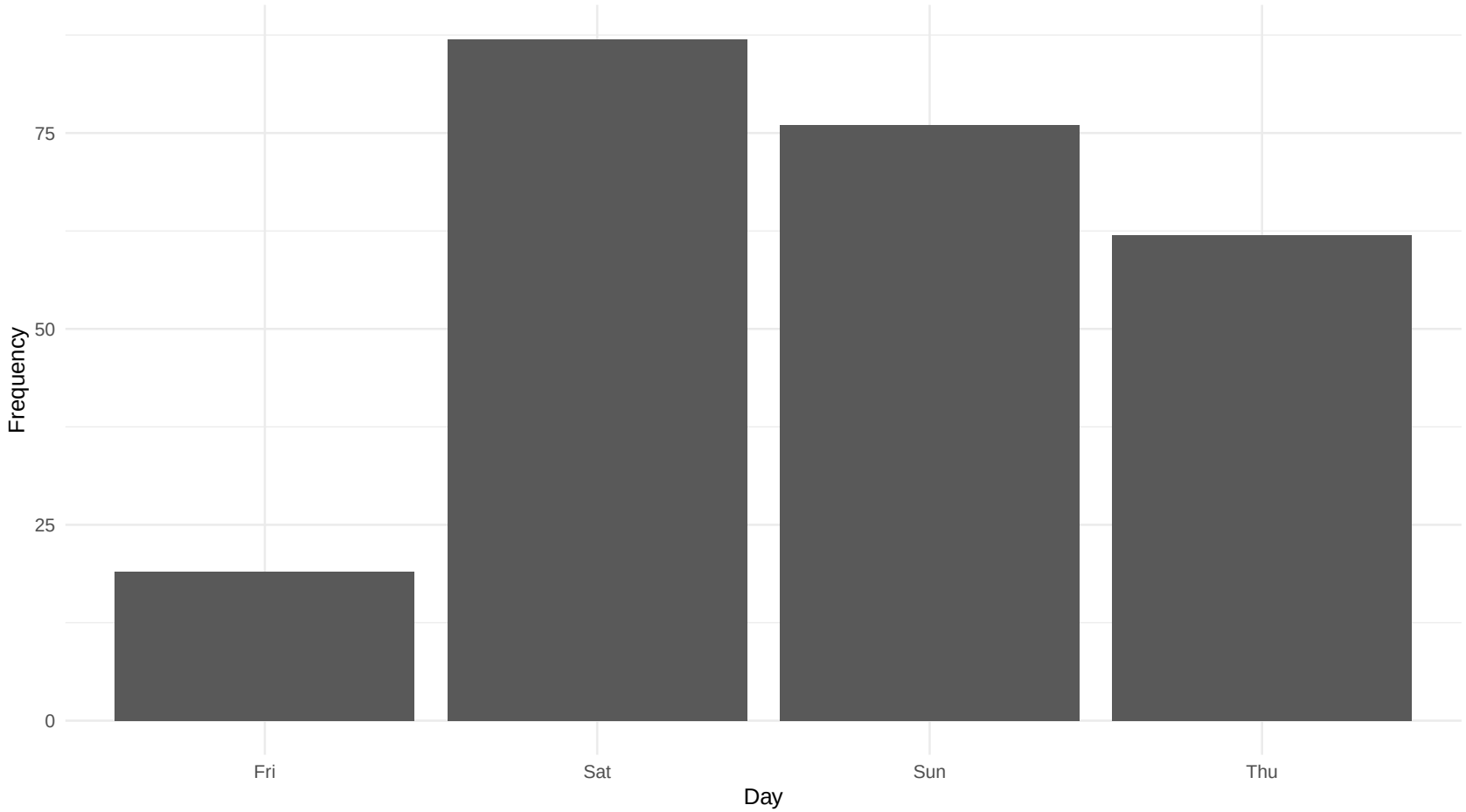
Proportions vs. Percentages

- Recognize that proportions are simply percentages divided by 100
  - Statisticians prefer proportions in most situations because of their connection with probability (a topic for another time)

Day	Proportion	Percentage
Fri	0.08	8%
Sat	0.36	36%
Sun	0.31	31%
Thu	0.25	25%

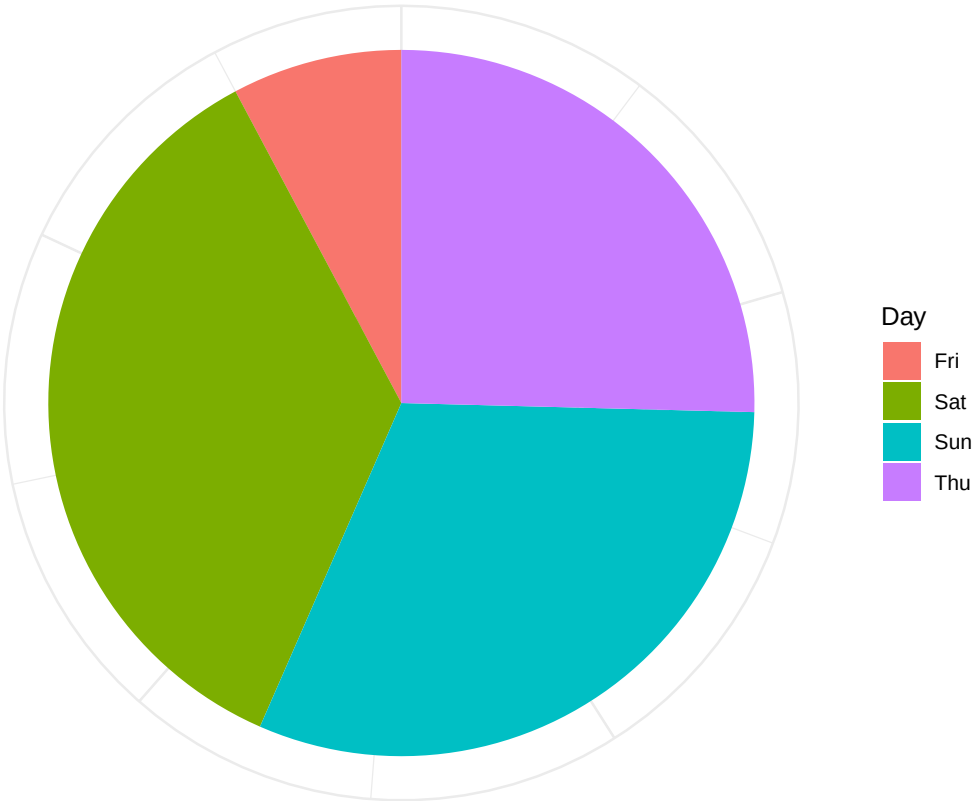
Barcharts

- The best and most common way to visualize categorical variables is the **bar chart**:



Piecharts

- An alternative is the **pie chart**, but research has shown that readers perceive information more accurately from bar charts



Closing Remarks (Categorical Variables)

- There isn't a whole lot to say about summarization for a single categorical variable
  - *Frequencies* and *proportions* describe how often different categories occur within a dataset
  - *Bar charts* and *pie charts* provide a visual depiction

Associations Between Two Variables

- The *univariate* summaries and graphs we've discussed are very important
  - They are often the first steps in any statistical analysis

Associations Between Two Variables

- The *univariate* summaries and graphs we've discussed are very important
  - They are often the first steps in any statistical analysis
- However, most statistical analyses are focused on establishing relationships between *multiple variables*
  - We'll use the term **association** to describe an observed relationship or correspondence between two or more variables

Contingency Tables

- Consider a dataset containing *two categorical variables*
  - We can display frequencies for *each combination* of the variables in a **contingency table** (also called a two-way frequency table)
- Below is a two-way frequency table describing the historic Golden State Warriors 2015-16 season:

	Win	Loss
Home	39	2
Away	34	7

Practice

What do you think the raw data that was used to construct this table looks like? Try writing out a few rows.

	Win	Loss
Home	39	2
Away	34	7

Practice (solution)

Recognize you're only able to discern the last two columns from the contingency table on the prior slide

Date	Opp	Location	Win
10/27/2015	NOP	Home	W
10/30/2015	HOU	Away	W
10/31/2015	NOP	Away	W
11/2/2015	MEM	Home	W
11/4/2015	LAC	Home	W
11/6/2015	DEN	Home	W
11/7/2015	SAC	Away	W
11/9/2015	DET	Home	W
11/11/2015	MEM	Away	W
11/12/2015	MIN	Away	W
11/14/2015	BRK	Home	W
11/17/2015	TOR	Home	W
11/19/2015	LAC	Away	W
11/20/2015	CHI	Home	W
11/22/2015	DEN	Away	W
11/24/2015	LAL	Home	W
11/27/2015	PHO	Away	W
11/28/2015	SAC	Home	W
11/30/2015	UTA	Away	W
12/2/2015	CHO	Away	W
12/5/2015	TOR	Away	W
12/6/2015	BRK	Away	W
12/8/2015	IND	Away	W
12/11/2015	BOS	Away	W
12/12/2015	MIL	Away	L
12/16/2015	PHO	Home	W
12/18/2015	MIL	Home	W

Margins

A useful preliminary step when working with contingency tables is to add *table margins*:

	Win	Loss	Row Total
Home	39	2	41
Away	34	7	41
Column Total	73	9	82

Margins

A useful preliminary step when working with contingency tables is to add *table margins*:

	Win	Loss	Row Total
Home	39	2	41
Away	34	7	41
Column Total	73	9	82

- These row/column totals are sometimes called **marginal distributions**
  - The marginal distribution of the "win" variable (win/loss) is characterized by the frequencies $\{73, 9\}$ and the proportions $\{0.89, 0.11\}$
  - The marginal distribution of the "location" variable (home/away) is characterized byt the frequencies $\{41, 41\}$ and the proportions $\{0.5, 0.5\}$

Conditional Proportions

- Starting with a contingency table, we can look at **conditional proportions** to determine if the two variables displayed are *associated*
- There are two types of conditional proportions: **row proportions** are calculated using each row's total, the bottom table show how to calculate these

	Win	Loss	Row Total
Home	39	2	41
Away	34	7	41
Column Total	73	9	82

	Win	Loss	Row Total
Home	$39/41 = 0.95$	$2/41 = 0.05$	1
Away	$34/41 = 0.83$	$7/41 = 0.17$	1
Column Total	$73/82 = 0.89$	$9/82 = 0.11$	1

Conditional Proportions

- **Column proportions** are calculated in a similar way

	Win	Loss	Row Total
Home	39	2	41
Away	34	7	41
Column Total	73	9	82

	Win	Loss	Row Total
Home	$39/73 = 0.53$	$2/9 = 0.22$	$41/82 = 0.5$
Away	$34/73 = 0.47$	$7/41 = 0.78$	$41/82 = 0.5$
Column Total	1	1	1

Conditional Distributions and Association

- Two variables are **associated** if the distribution of one variable depends upon the value of the other variable
- For example, we might compare the distribution of win/loss proportions *conditional upon a game being at home* with the distribution of win/loss proportions *conditional upon a game being away*
  - If these distributions differ, the variables "location" and "win" are associated

Practice

1. Using the row proportions given below, do you think there is an association between whether the Warriors were home/away and winning?
2. How would you explain this association?

	Win	Loss	Row Total
Home	0.95	0.05	1
Away	0.83	0.17	1
Column Total	0.89	0.11	1

Practice (solution)

1. Yes, there is an association between "location" and "win"
2. The warriors look to be *more likely* to win when playing at home. In other words, the distribution of wins/losses for home games differs from the distribution of wins/losses for away games.

Remarks

- Recognize that row and column proportions tell you fundamentally different things about your data
  - In our example, *row proportions* can describe the *proportion of wins conditional on the game being at home*
  - Contrast that with *column proportions*, which can describe the *proportion of home games conditional on that game being a win*

Remarks

- Recognize that row and column proportions tell you fundamentally different things about your data
  - In our example, *row proportions* can describe the *proportion of wins conditional on the game being at home*
  - Contrast that with *column proportions*, which can describe the *proportion of home games conditional on that game being a win*
- The row proportions suggest how often home games were won, while the column proportions suggest how often wins were home games
  - This distinction doesn't seem to matter much here, but let's look at another example

Practice

- Were crew members on the Titanic more likely to survive than 1st class passengers?
  - Use row or column proportions from the contingency table below to support your answer

	Survived	Died
Crew	212	673
1st Class	203	122

Practice (solution)

- No, using *row proportions* we see that $\frac{212}{623+212} = 0.24$, or 24% of the crew survived; while $\frac{203}{122+203} = 0.62$, or 62% of first class passengers survived

	Survived	Died
Crew	212	673
1st Class	203	122

Practice (solution)

- No, using *row proportions* we see that $\frac{212}{623+212} = 0.24$, or 24% of the crew survived; while $\frac{203}{122+203} = 0.62$, or 62% of first class passengers survived

	Survived	Died
Crew	212	673
1st Class	203	122

- Notice that this particular question *cannot be answered* using column proportions
  - The proportion of survivors who were crew is $\frac{212}{212+203} = 0.51$, while the proportion of survivors who were first class passengers is $\frac{203}{212+203} = 0.49$

Practice (solution)

- No, using *row proportions* we see that $\frac{212}{623+212} = 0.24$, or 24% of the crew survived; while $\frac{203}{122+203} = 0.62$, or 62% of first class passengers survived

	Survived	Died
Crew	212	673
1st Class	203	122

- Notice that this particular question *cannot be answered* using column proportions
  - The proportion of survivors who were crew is $\frac{212}{212+203} = 0.51$, while the proportion of survivors who were first class passengers is $\frac{203}{212+203} = 0.49$
  - Conditioning on the column variable is problematic here because the *marginal distribution* of 1st class/crew is *skewed towards crew*
  - In other words, most of the survivors were crew members because there were so many more crew members, not because individual crew members were more likely to survive