Data and Univariate Summaries

Ryan Miller

Outline

1. Working with data
2. Describing a quantitative variable
3. Describing a categorical variable

Data

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

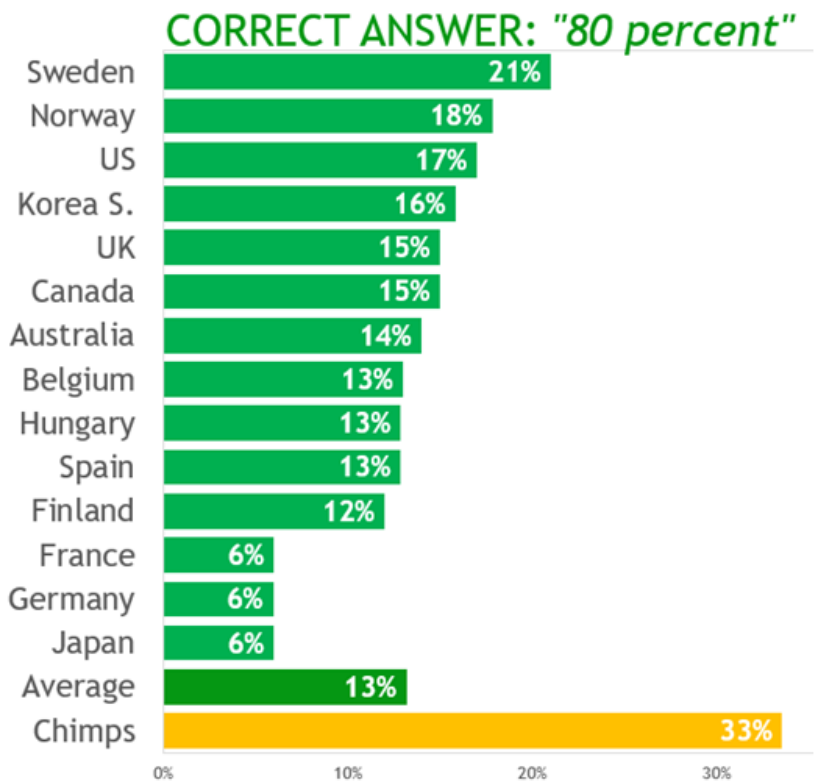
A) 20%
B) 50%
C) 80%

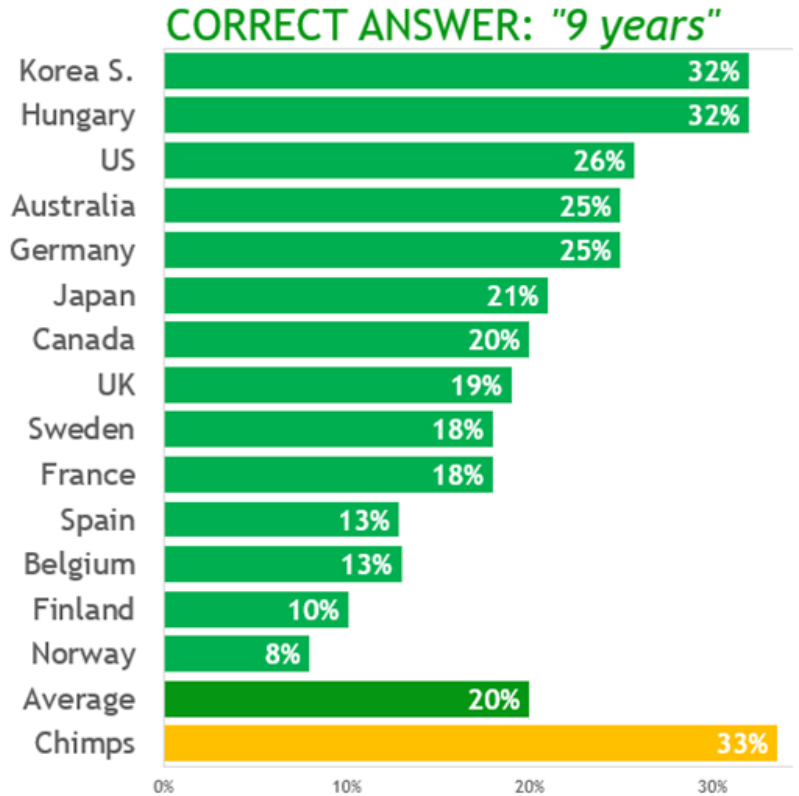
Question 2: Worldwide, 30-year-old men have 10 years of schooling, on average. How many years do women of the same age have?

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

Data

Here's what the data show:





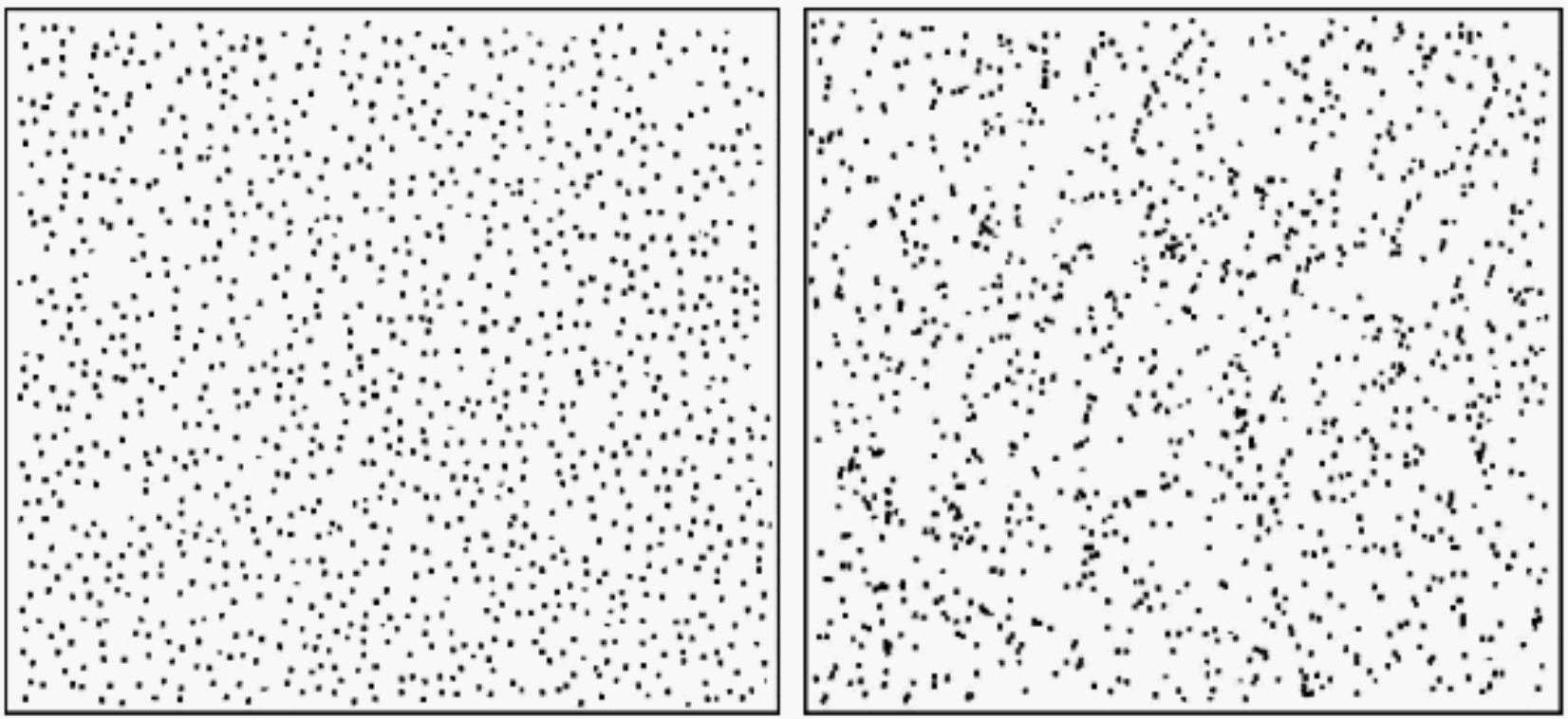
Source: Allan Rossman's JSM talk

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

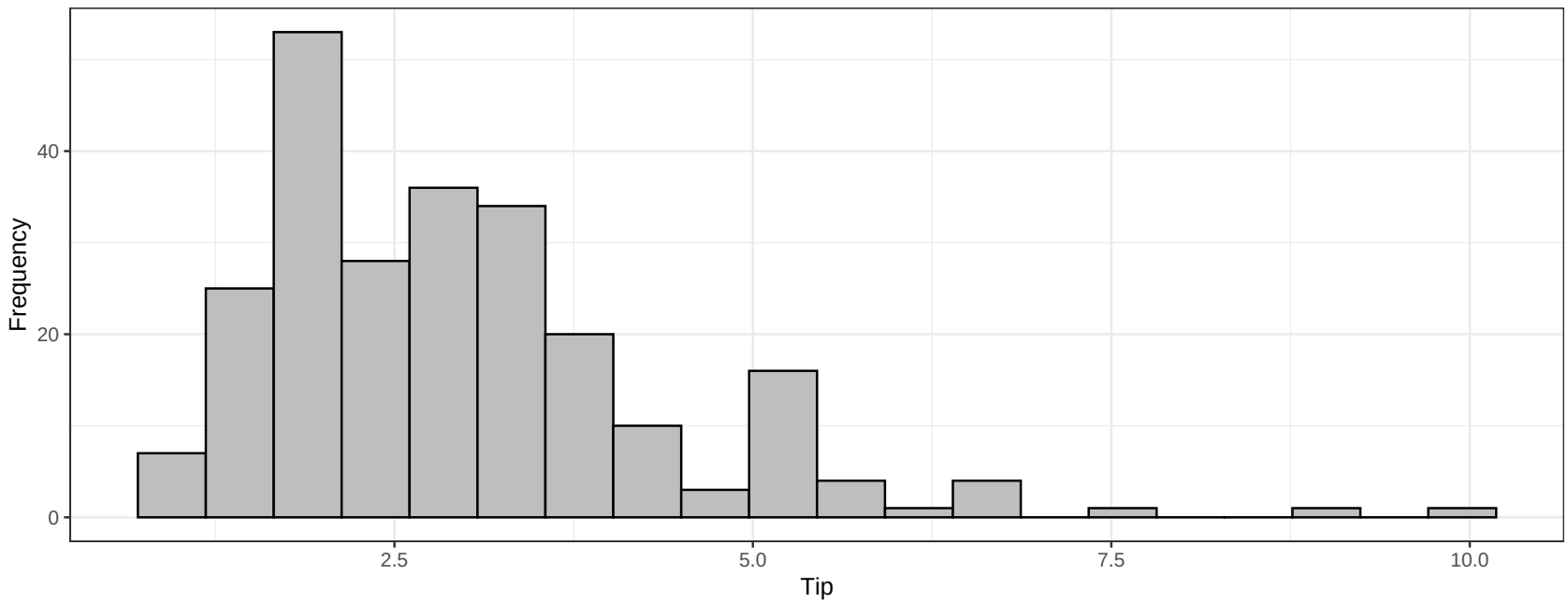
- **Raw data** can be difficult to make sense of
- **Summarization** refers to methods that simplify raw data into a more understandable form
  - Ideally, we can summarize a variable using one number, or a small set of numbers, in order to make informed judgments
- For now, we'll focus on **univariate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

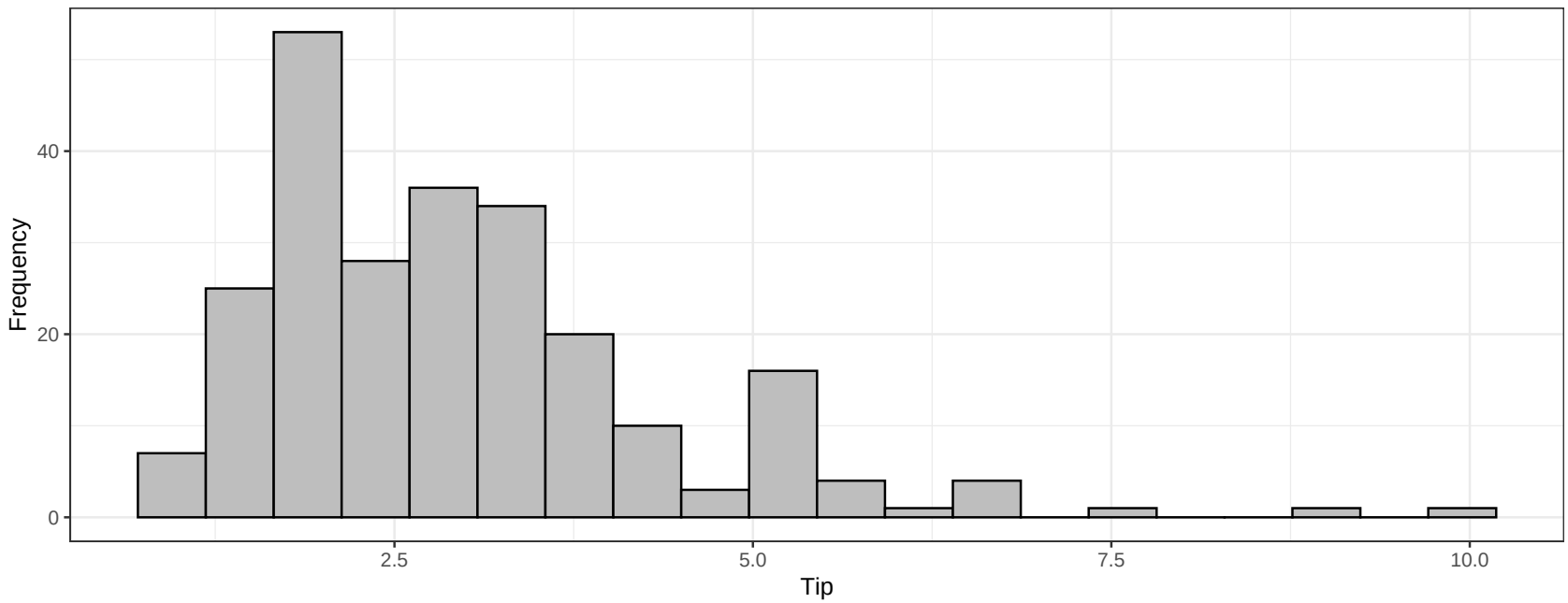
Distributions

- 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

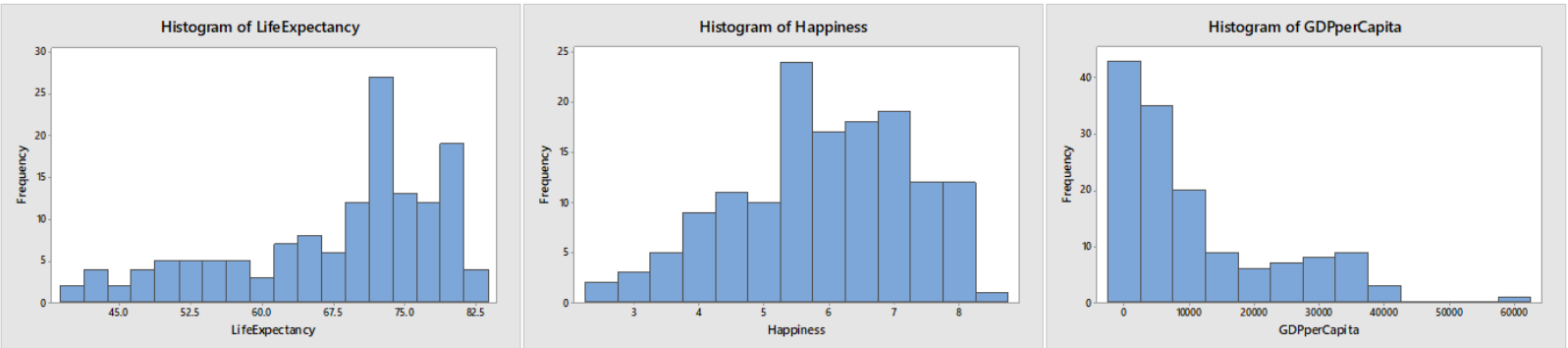


Judging Shape from Histograms

- 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



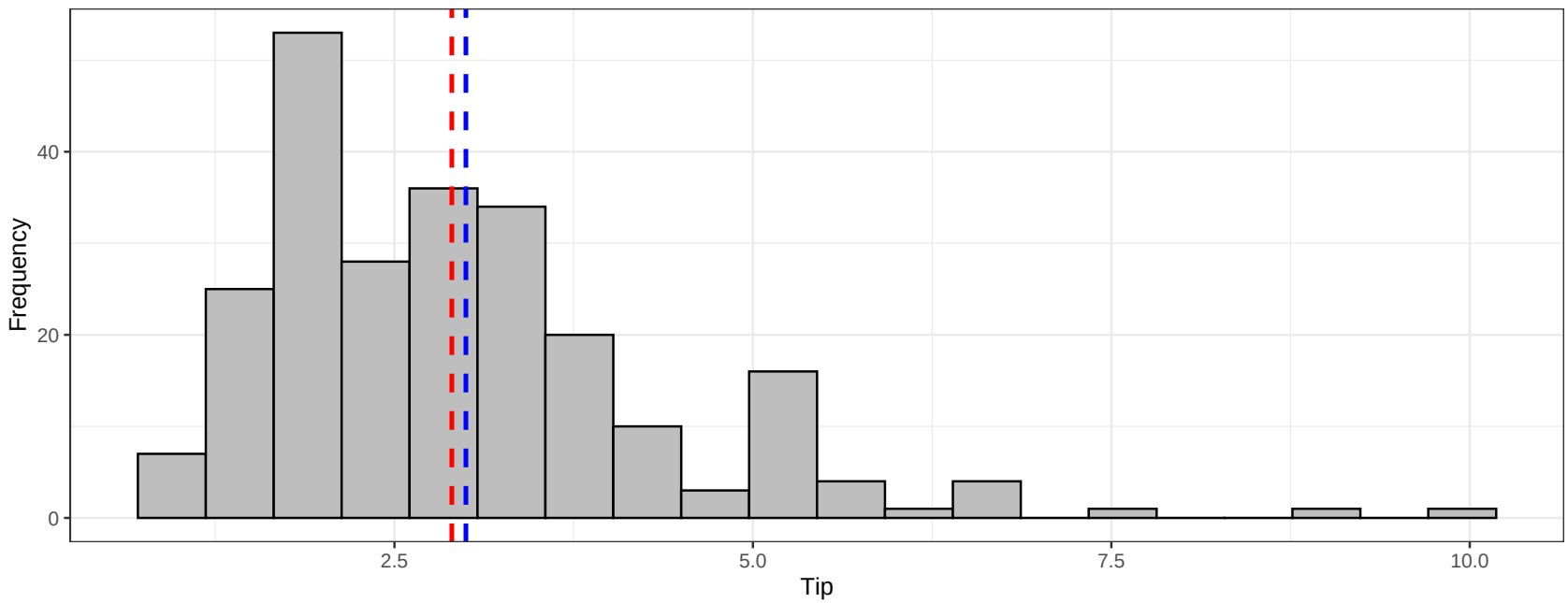
The Mean

- 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

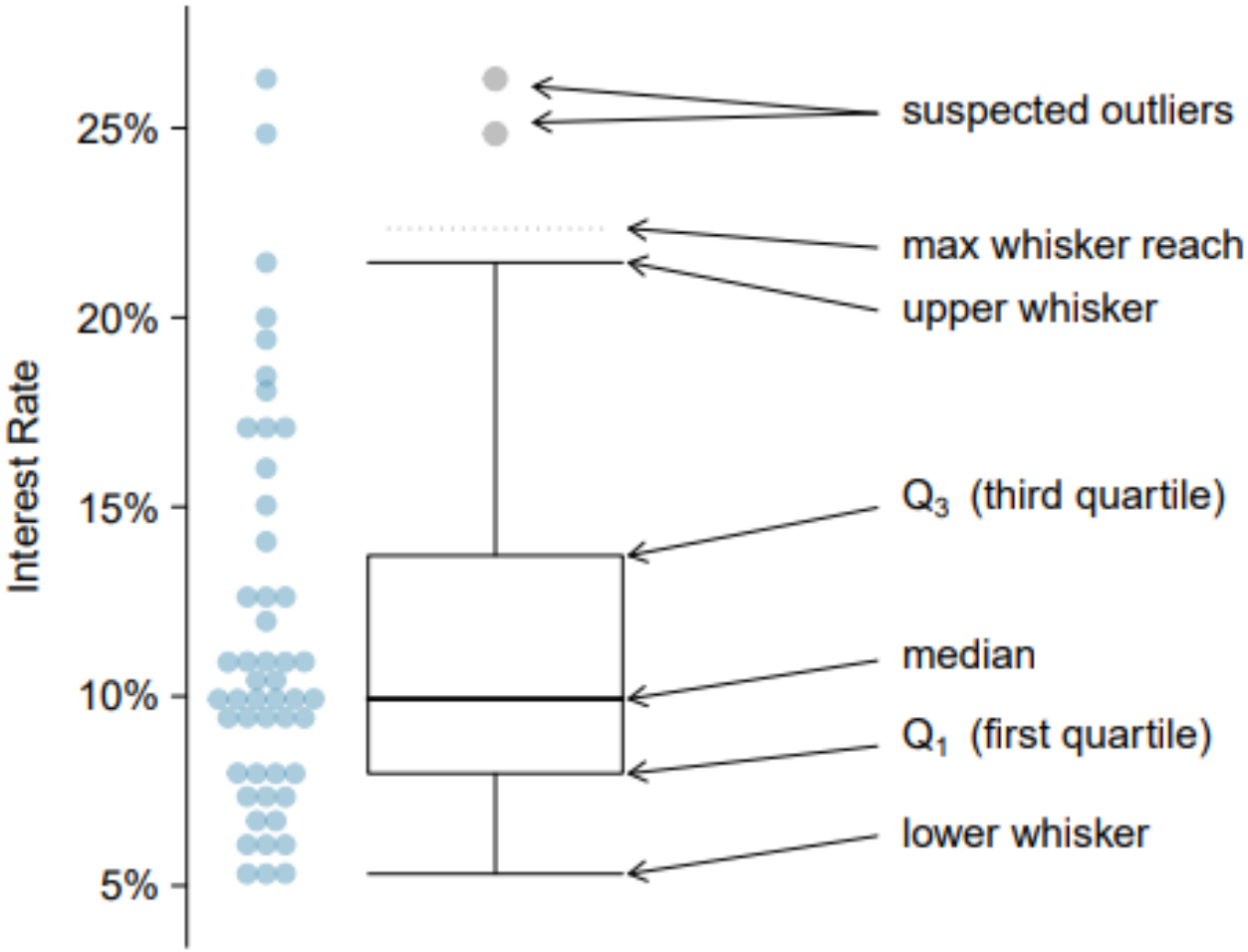
- Very often we'd like to know more about a variable than simply the center of its distribution
- The **minimum** and **maximum** are self-explanatory summaries of the variable's most extreme values

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

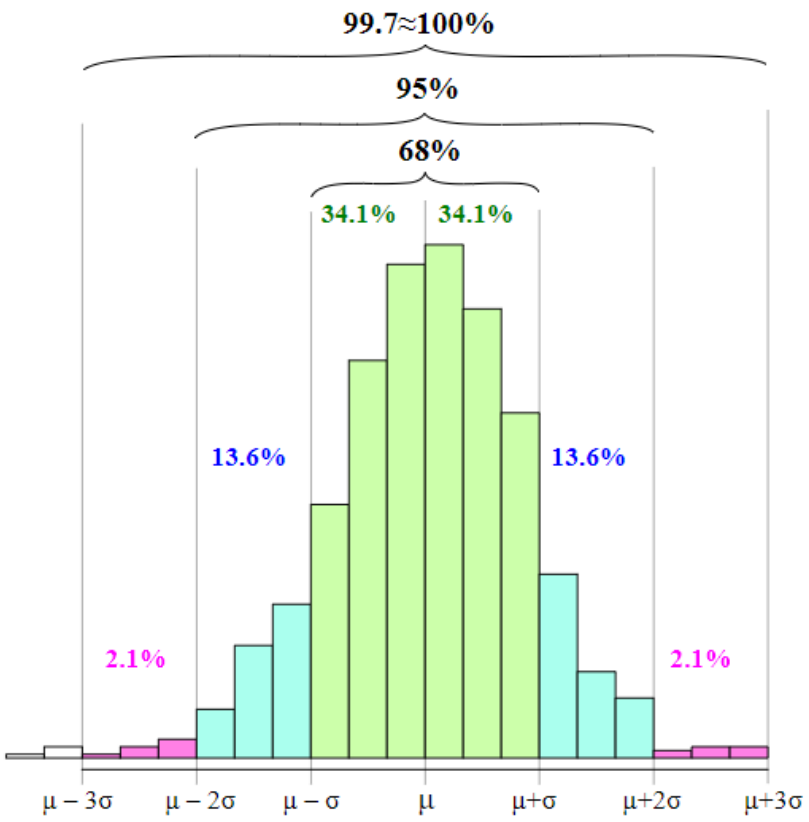
- The mean and median summarize the *center* of a distribution
- It is also useful to summarize the *spread*, or how the data values tend to vary around the center

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



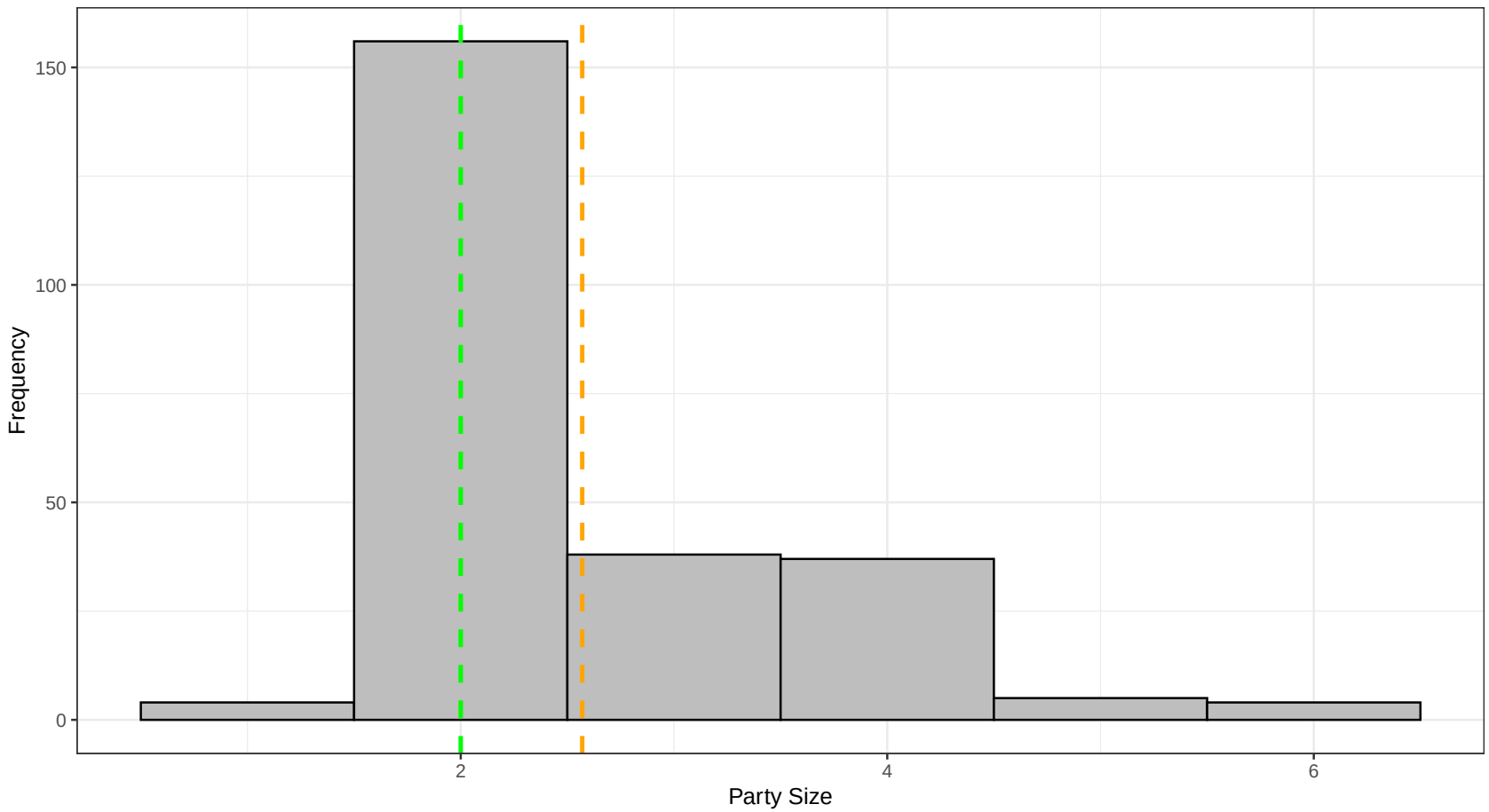
Standard Deviation vs. IQR

- 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

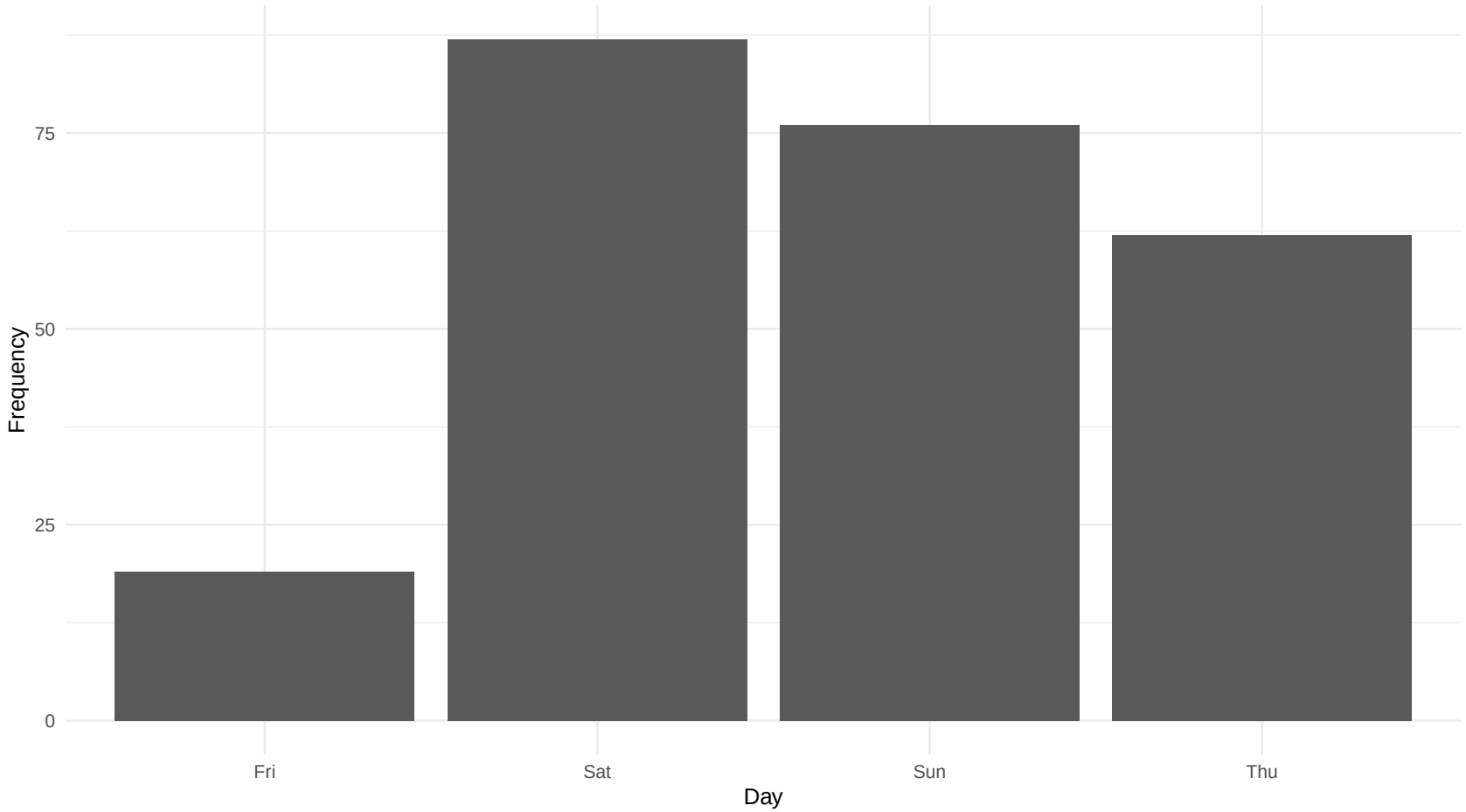
Proportions vs. Percentages

- 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%

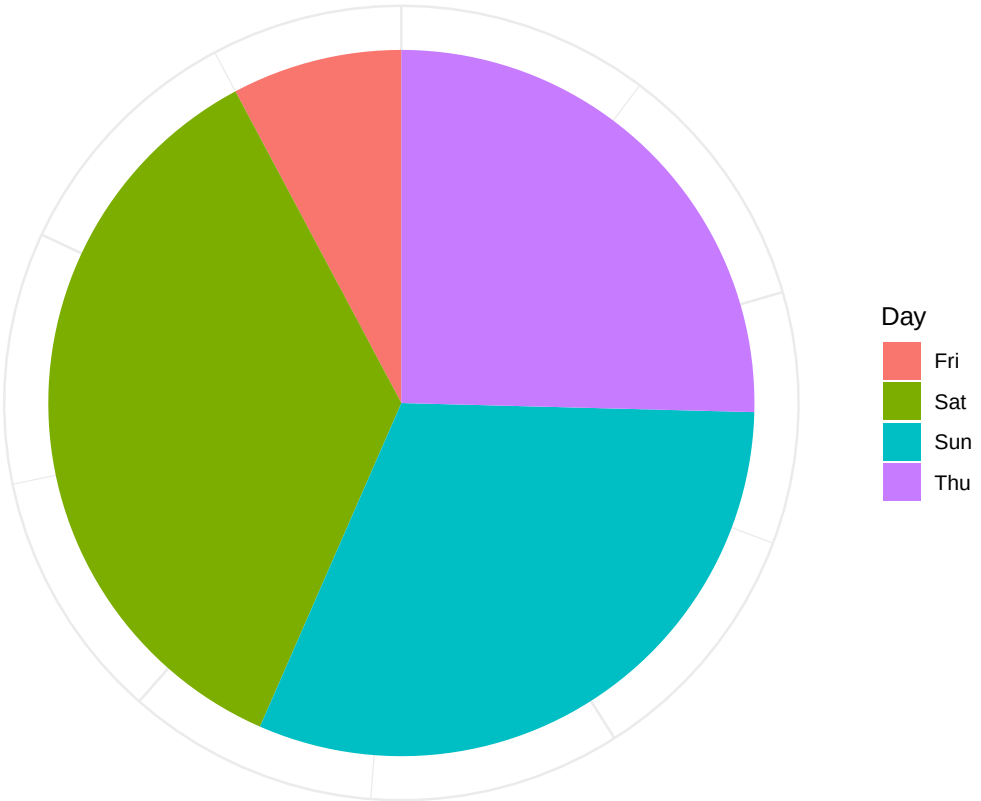
Barcharts

- 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



Practice

In this course we'll use a web-based statistical program known as StatKey to create graphs and calculate descriptive statistics. You can access StatKey at https://www.lock5stat.com/StatKey/

1) If necessary, re-download the Happy Planet dataset
2) Using the "Descriptive Statistics and Graphs" section of StatKey, provide a univariate summary of the variable "LifeExpectancy" (Hint: this is a quantitative variable, so you should describe its shape, center, and spread)
3) Next, provide a univariate summary of the variable "Region".

1	2	3	4	5	6	7
Latin America	Western Nations	Middle East	Africa	South Asia	East Asia	Former Soviet States

Practice (solution)

1) Not shown
2) Life Expectancies are skewed left, centered at a mean of roughly 67 years, and have a standard deviation of 11 years and an IQR of roughly 14 years.
3) Region #5 (South Asia) contains the fewest countries (roughly 5%), while region #4 (Africa) contains the most (roughly 23%)