Data Basics

Ryan Miller

Introduction

- A necessary precursor to practicing statistics is learning how to work with data
  - The first step in this process is learning the terminology that statisticians use to talk about data
- Our goal for today is to learn some basic definitions pertaining to data, and to practice using them to talk about real datasets
  - We'll also talk about ways to summarize and display our data in order to make it easier to talk about

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

Download the "Happy Planet" dataset here:
https://remiller1450.github.io/data/HappyPlanet.csv

- **Country**: Name of the country
- **Region**: Code for the region, 1 = Latin America, 2 = Western Nations, 3 = Middle East, 4 = Sub-Saharan Africa, 5 = South Asia, 6 = East Asia, 7 = Former Communist Countries
- **Happiness**: 0 to 10 score from Gallop World Poll data
- **LifeExpectancy**: Average life expectancy (years) from UN Department of Economic and Social Affairs
- **Footprint**: A measure of ecological footprint from *The Edition of the Global Footprint Networks National Footprint Accounts*, higher numbers indicate greater environmental impact
- **HLY**: Happy Life Years - measures life expectancy and well-being
- **HPI**: Happy Plant Index - a 0-100 score
- **HPIRank**: HPI rank of the country
- **GDPperCapita**: Gross Domestic Product per capita
- **HDI**: Human Development Index, the UN Report Office
- **Population**: Population (in millions)

Practice

Discuss the following with the person next to you:

1) What are the cases in this dataset?
2) What type of variable is "Population"?
3) What type of variable is "Region"?
4) Can you think of an analysis that would use these data, but the cases would differ from what you answered in #1?

Practice (solution)

1) Each case in these data is a country
2) "Population" is a quantitative variable, it is measured in millions of people (a numeric entity)
3) "Region" is categorical variable, it divides the cases into 7 geographic groups (categories)
4) You could group the countries by region, then treat the 7 regions as cases

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)

Summarizing Data

A restaurant server wanting to understand their income collects data on every table they serve. Data from 20 tables are displayed below. What do these data tell you?

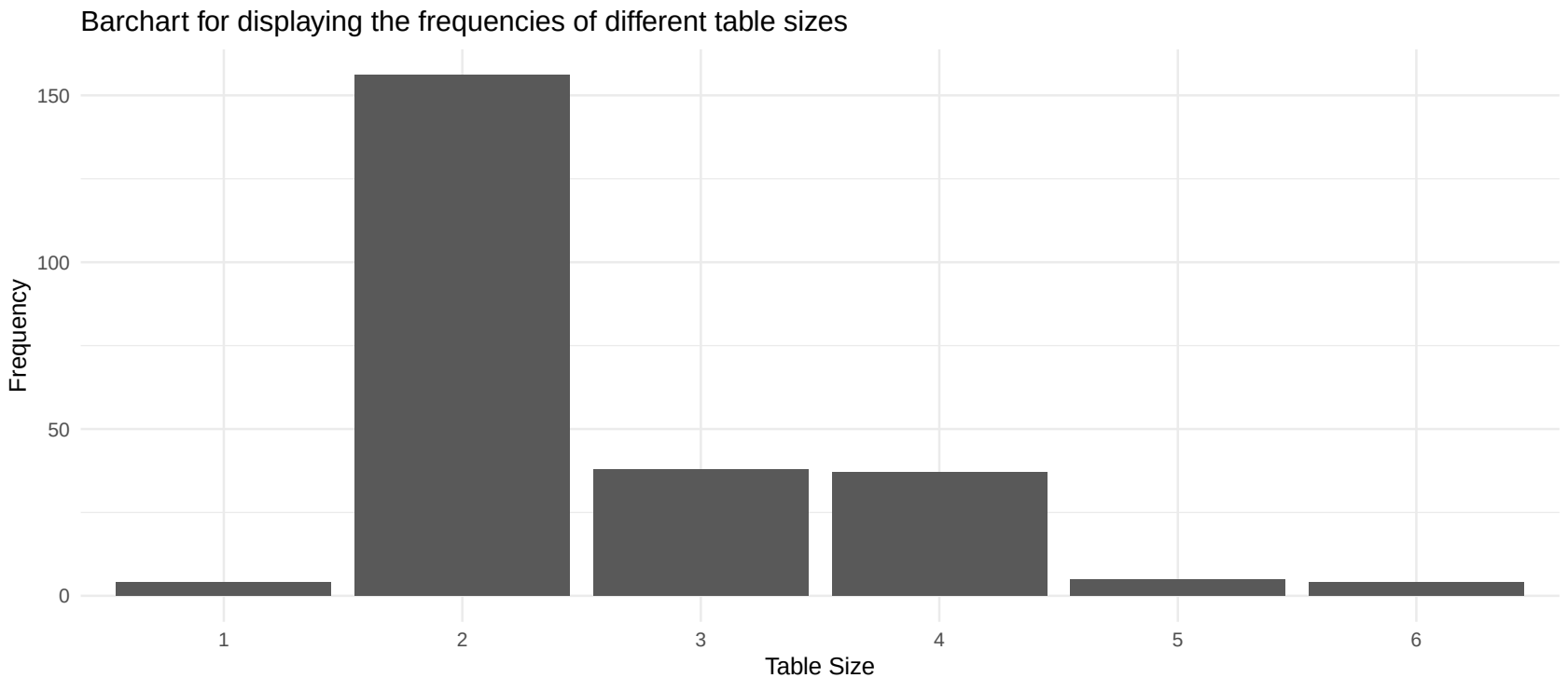
total_bill	tip	sex	smoker	day	time	size
12.69	2.00	Male	No	Sat	Dinner	2
17.29	2.71	Male	No	Thur	Lunch	2
7.51	2.00	Male	No	Thur	Lunch	2
11.35	2.50	Female	Yes	Fri	Dinner	2
10.07	1.25	Male	No	Sat	Dinner	2
14.00	3.00	Male	No	Sat	Dinner	2
10.33	2.00	Female	No	Thur	Lunch	2
11.17	1.50	Female	No	Thur	Lunch	2
24.52	3.48	Male	No	Sun	Dinner	3
27.05	5.00	Female	No	Thur	Lunch	6
20.27	2.83	Female	No	Thur	Lunch	2
12.03	1.50	Male	Yes	Fri	Dinner	2
44.30	2.50	Female	Yes	Sat	Dinner	3
13.27	2.50	Female	Yes	Sat	Dinner	2
21.16	3.00	Male	No	Thur	Lunch	2
15.01	2.09	Male	Yes	Sat	Dinner	2
22.76	3.00	Male	No	Thur	Lunch	2
16.47	3.23	Female	Yes	Thur	Lunch	3
17.31	3.50	Female	No	Sun	Dinner	2
18.43	3.00	Male	No	Sun	Dinner	4

Why summarize?

- **Summarization** describes any process that reduces the data to a single number (or a small set of numbers)
  - Data is rarely useful without some degree of summarization, humans just aren't capable of processing that much information
  - In this class, we will focus mostly on **univariate** summaries (those involving a single variable) and **bivariate** summaries (those involving two variables)

Summarizing Categorical Variables

- Categorical variables are the simplest to summarize, there are really only two commonly used measures:
  - **Frequencies**: counts of how many cases belong to a particular category
  - **Proportions**: fractions based upon frequencies, sometimes called *relative frequencies*



Two Categorical Variables

- Frequencies of two categorical variables can be cross-tabulated and displayed in a **two-way table** (contingency table)
  - We will work with these tables extensively throughout the course
  - Note that proportions may now be calculated relative to row or column totals in this setup

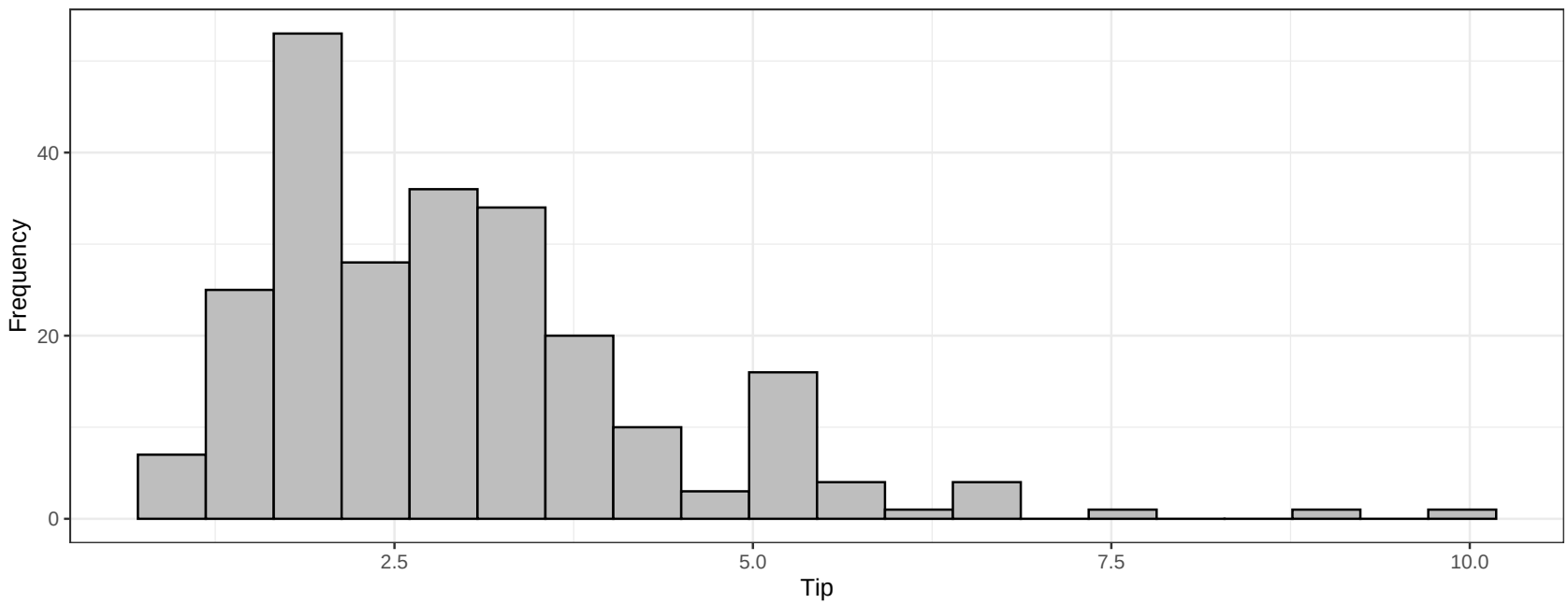
Size	Dinner	Lunch
1	2	2
2	104	52
3	33	5
4	32	5
5	4	1
6	1	3

Numeric Variables and Distributions

- Before diving into summarization of numeric variables, it is useful to talk about how data are *distributed*
  - A variable's **distribution** describes values that are possible and how frequently they occur

Numeric Variables and Distributions

- Before diving into summarization of numeric variables, it is useful to talk about how data are *distributed*
  - 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
  - \$2-3 tips are most common, larger tips of \$5+ do occasionally occur, tips over \$10 almost never occur



Distributions and Univariate Summaries

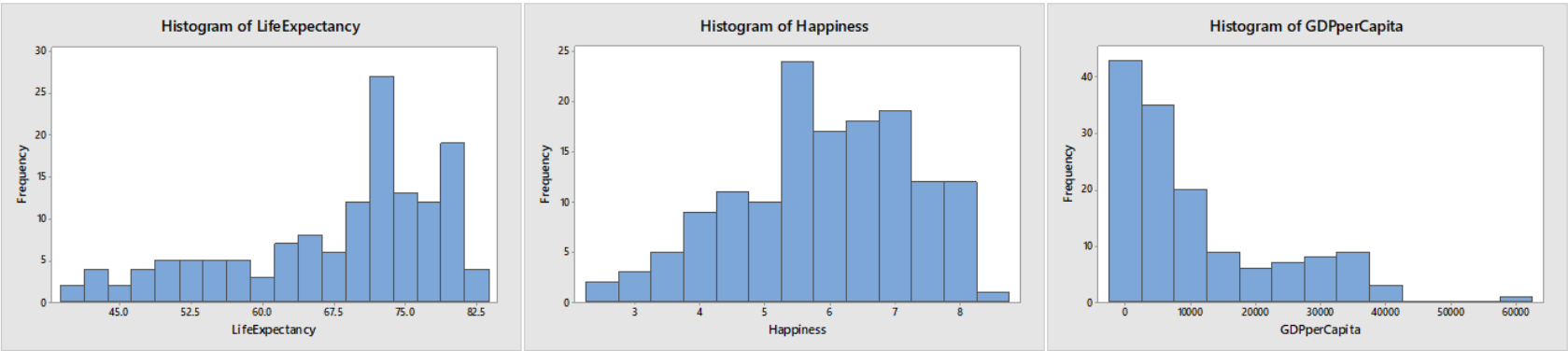
- Distributions aren't themselves a summary, but they can help us understand summarization
  - The *most common tips* could be more precisely characterized by the **mean** or **median**
  - The *less common larger tips* could be more precisely characterized by the **maximum** or **90% percentile**

Distributions and Univariate Summaries

- Distributions aren't themselves a summary, but they can help us understand summarization
  - The *most common tips* could be more precisely characterized by the **mean** or **median**
  - The *less common larger tips* could be more precisely characterized by the **maximum** or **90% percentile**
- Each of the four bolded terms is a different *univariate* summary measure
  - During future labs we'll get some practice working with these (and other) summary measures

Skew and Univariate Summaries

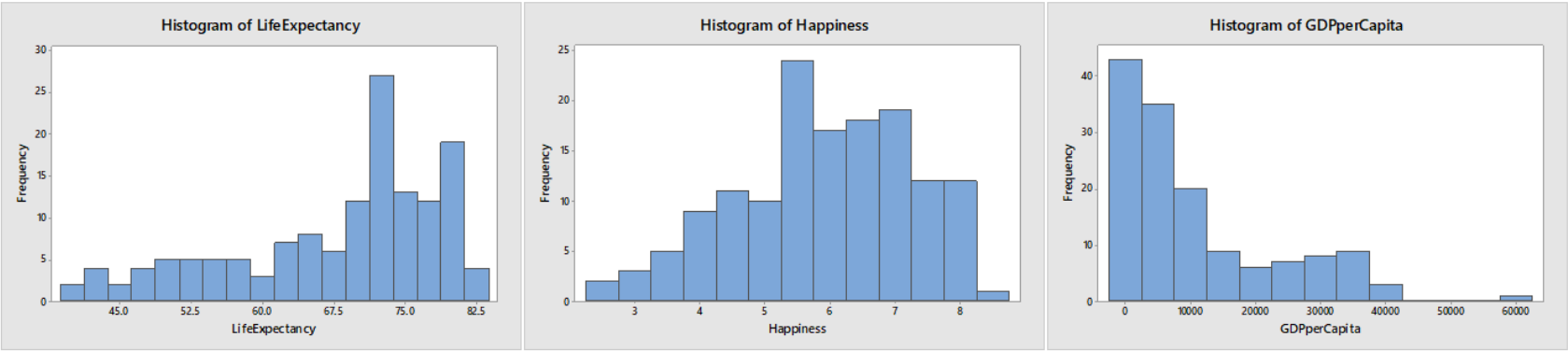
- Distributions, by definition, display how frequently certain values occur
  - An important descriptive characteristic of a distribution is its *shape*



- **Question**: How does a distribution's shape impact how its mean and median compare?

Skew and Univariate Summaries

- Distributions, by definition, display how frequently certain values occur
  - An important descriptive characteristic of a distribution is its *shape*



- **Question**: How does a distribution's shape impact how its mean and median compare?
  - Skew will pull the mean away from the median in that direction (ie: skewed right = mean > median)

Variability

- Distributions also display **variation** in the data, a fundamental concept in statistics
  - Variation is most commonly measured using **standard deviation**, which roughly corresponds to the *average distance of each data-point from the mean*
  - The *sample standard deviation* is calculated:

$$s_x = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}}$$

Where $\bar{x}$ is the sample average of the variable denoted by x

The 68-95-99 Rule

For symmetric, bell-shaped distributions, the standard deviation is related to the percentage of cases within a certain distance of the mean

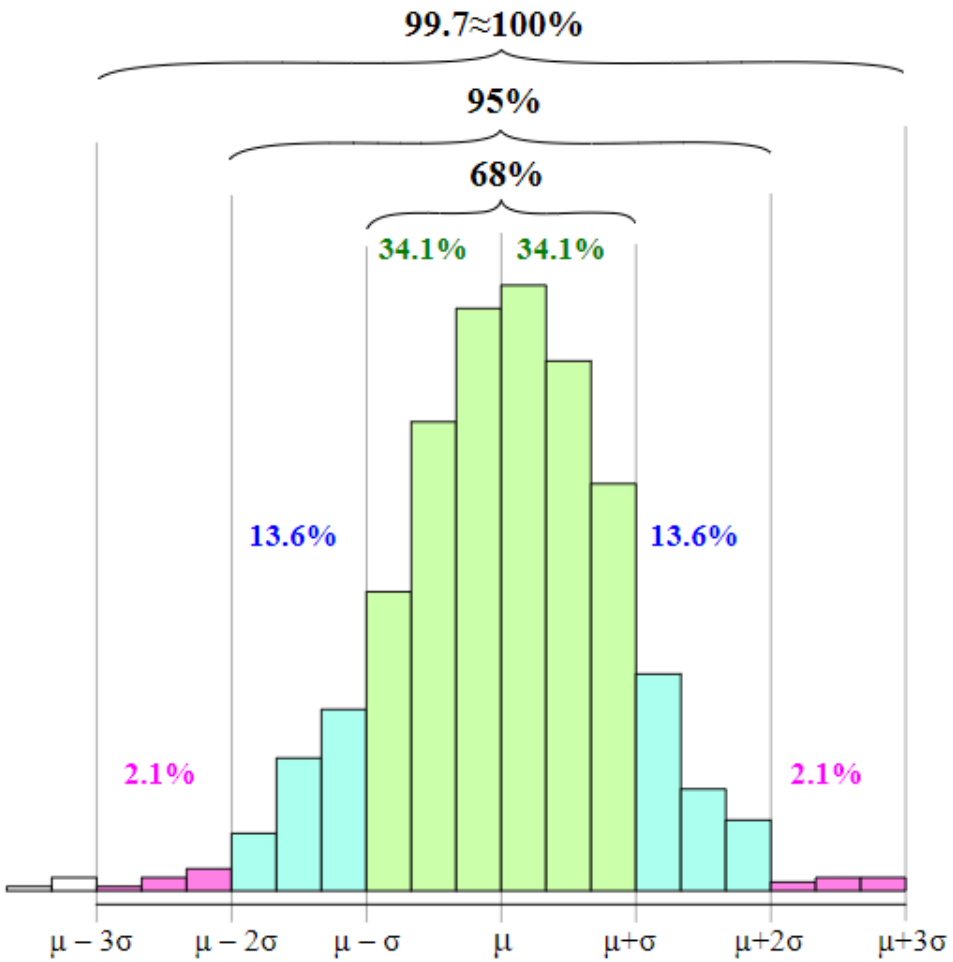
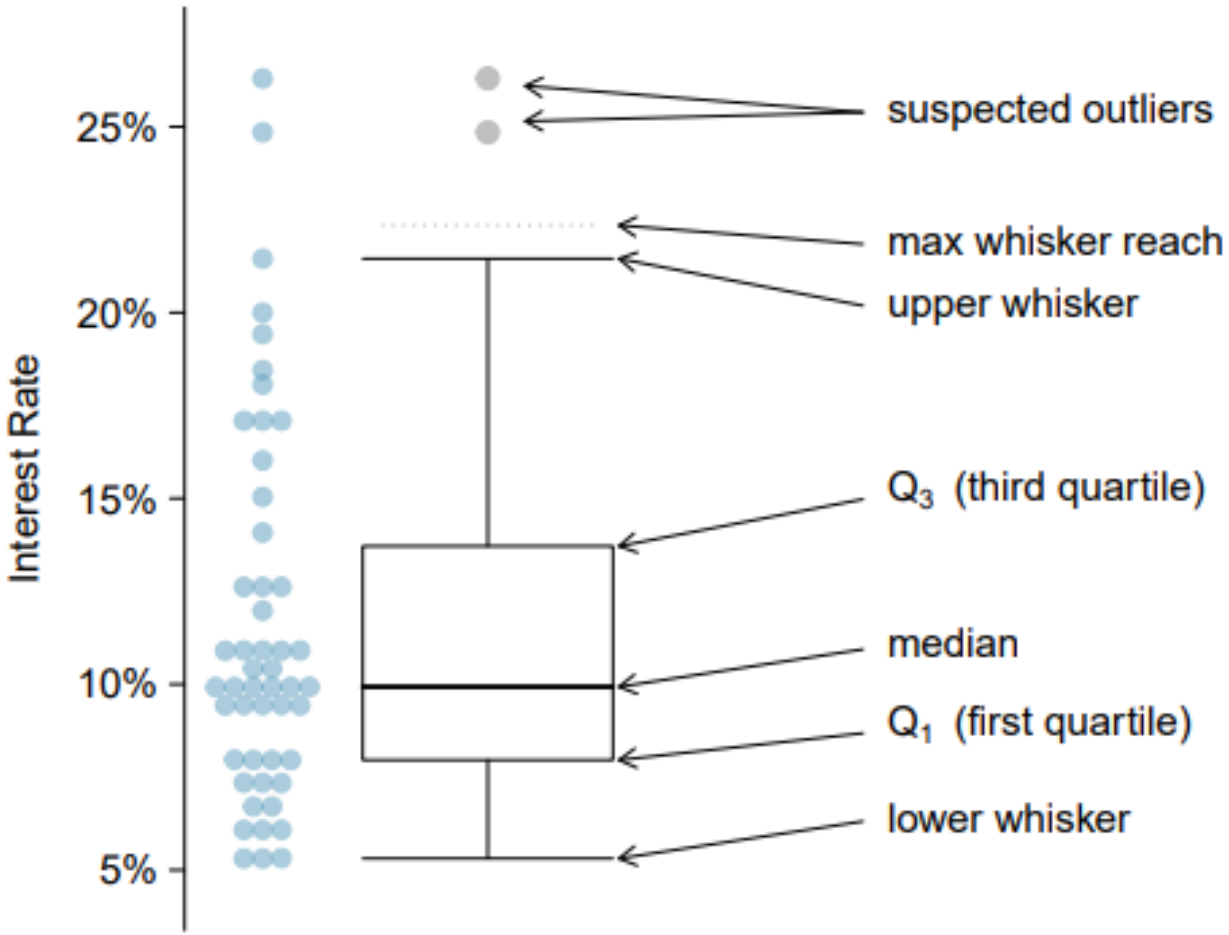


Image Source: https://en.wikipedia.org/wiki/68-95-99.7_rule

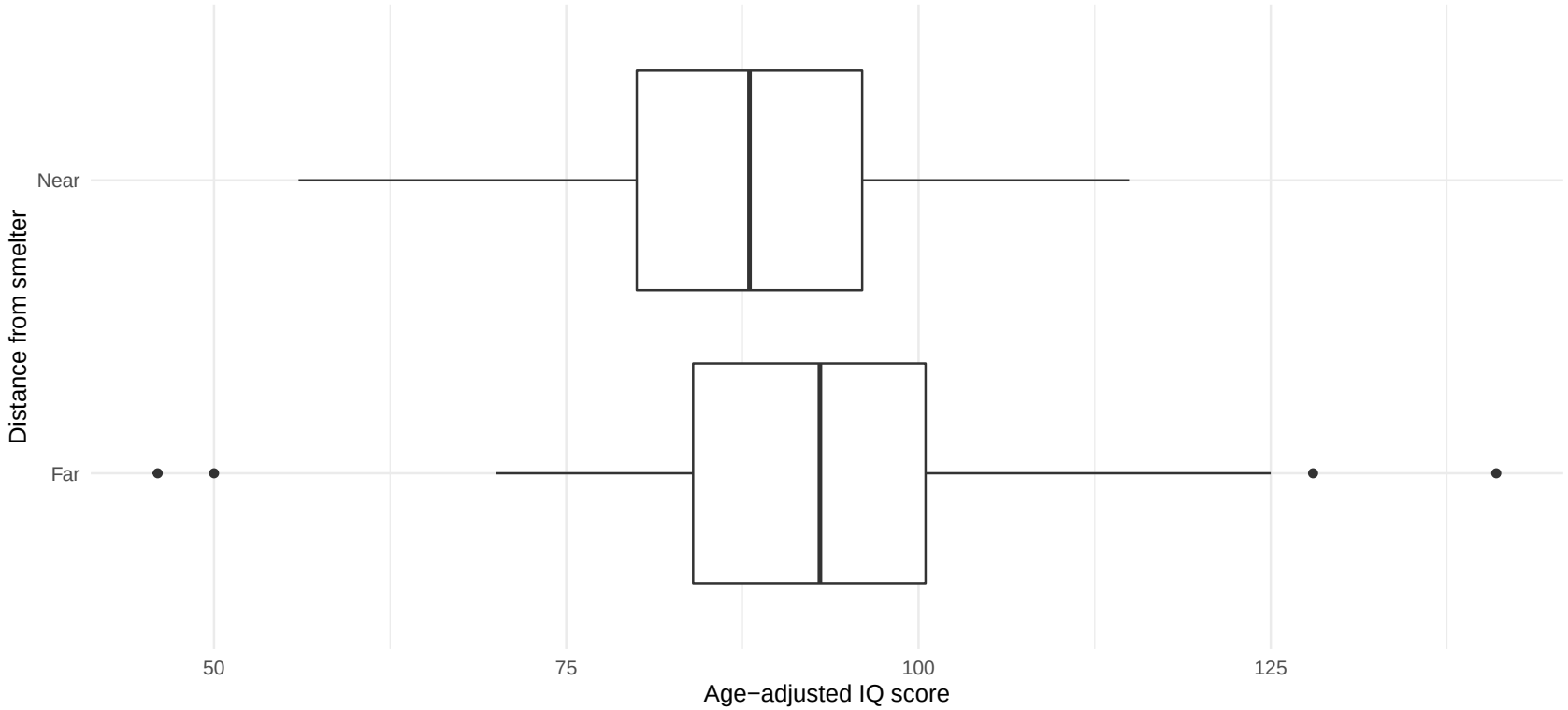
Boxplots

- One particularly good method of describing variable is the **boxplot**, which effectively conveys a number of summary statistics that effectively capture the variable's overall distribution



Relationships between Variables

- This graph shows the age-adjusted IQs of 114 children living El Paso, TX and their proximity to a local lead smelter
- **Practice**: Describe the depicted variables - 1) what are they? 2) are they categorical or quantitative? 3) are they related?



Practice (solution)

1) There are two variables
2) The quantitative variable is "IQ" and the categorical variable is "Proximity"
3) There appears to be some relationship, the distribution of IQ scores is different for each group

Association

- Two variables are **associated** if certain values of one variable tend to correspond with certain values of the other variable

Association

- Two variables are **associated** if certain values of one variable tend to correspond with certain values of the other variable
- For example, the **two-way frequency table** below suggests "table size" and "time of day" *are associated*
  - Using **column proportions** we see that 76.5% of lunches have size = 2, while only 59.1% of dinners have size = 2

Size	Dinner	Lunch
1	2	2
2	104	52
3	33	5
4	32	5
5	4	1
6	1	3

Explanatory and Response Variables

- When discussing association, we often think about *cause and effect*
  - "time" could influence "tip", but could "tip" influence "time"?

Explanatory and Response Variables

- When discussing association, we often think about *cause and effect*
  - "time" could influence "tip", but could "tip" influence "time"?
- In this spirit, an **explanatory variable** is one that is used to understand or predict a **response variable**

Explanatory and Response Variables

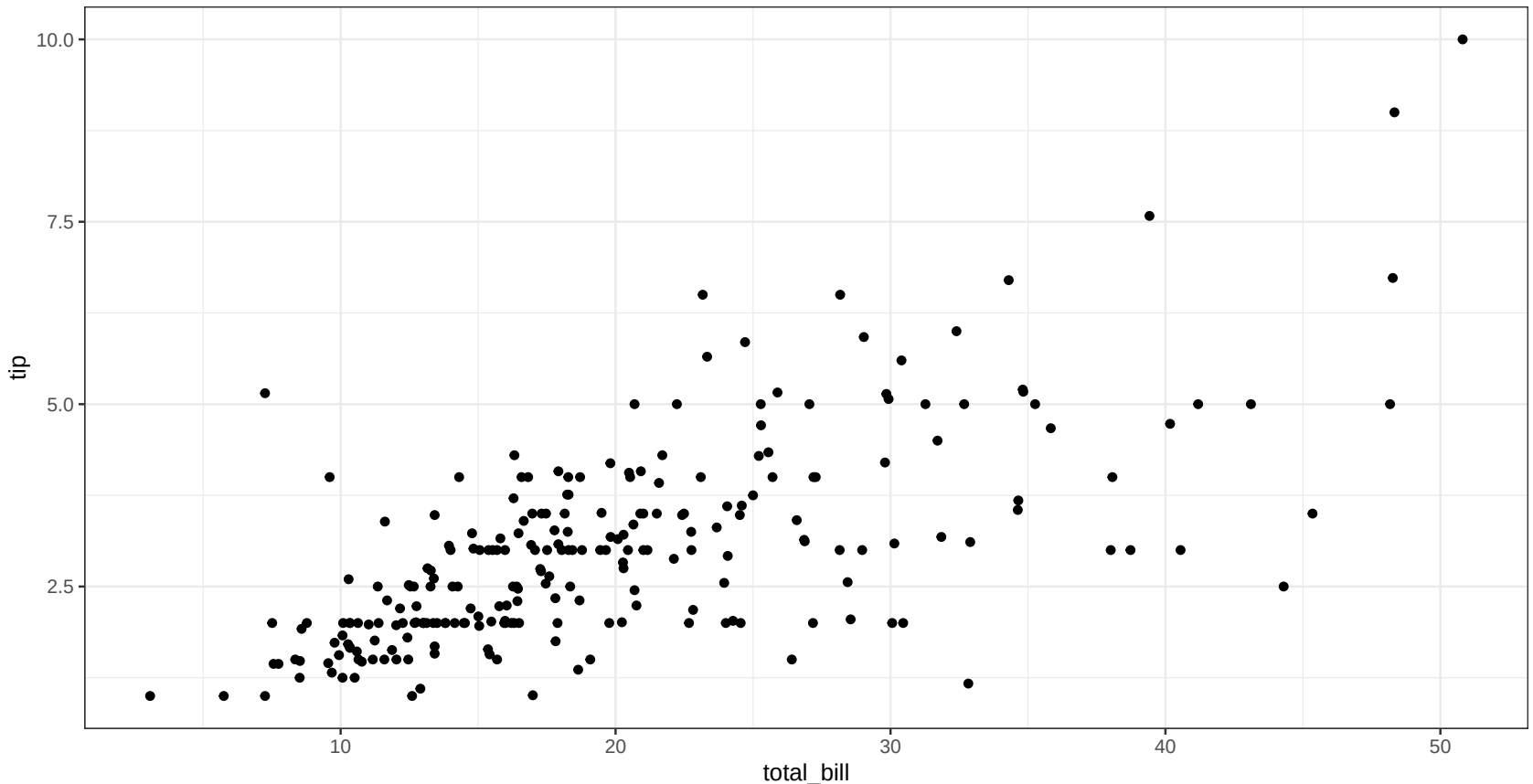
- When discussing association, we often think about *cause and effect*
  - "time" could influence "tip", but could "tip" influence "time"?
- In this spirit, an **explanatory variable** is one that is used to understand or predict a **response variable**
  - Not every two-variable relationship requires the designation of explanatory and response variables
  - Systolic blood pressure is strongly associated with diastolic blood pressure, but neither "explains" the other

Explanatory and Response Variables

- When discussing association, we often think about *cause and effect*
  - "time" could influence "tip", but could "tip" influence "time"?
- In this spirit, an **explanatory variable** is one that is used to understand or predict a **response variable**
  - Not every two-variable relationship requires the designation of explanatory and response variables
  - Systolic blood pressure is strongly associated with diastolic blood pressure, but neither "explains" the other
- We will revisit *cause and effect* soon, for now we'll use the general term "association" when discussing relationships between variables, and we'll avoid reading too much into *why* associations exist (a key topic for the rest of the semester)

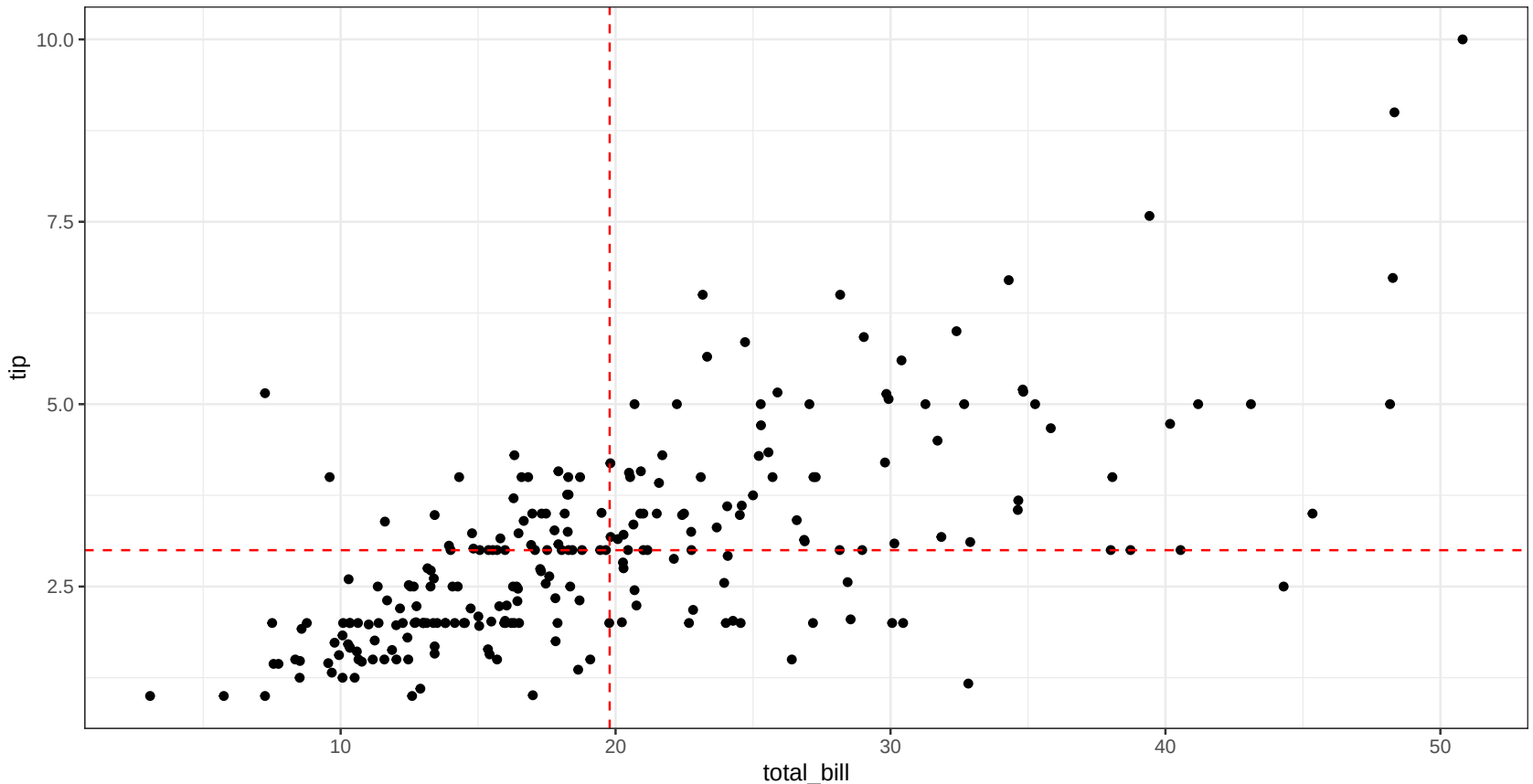
Practice

Using the scatterplot below, are the variables "total_bill" and "tip" associated? Why or why not? Which variable makes more sense to consider as an explanatory variable?



Practice (Solution)

Dividing the scatterplot into quadrants (using each variable's mean), an association is evidenced by the abundance of data in the upper-right and lower-left quadrants.



Correlation

Dividing the scatterplot into quadrants also enables us to understand the **correlation coefficient** (Pearson's correlation) as measure summarizing the relationship between two numeric variables (denoted x and y):

$$r_{xy} = \frac{1}{n-1} \sum_i \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- The correlation coefficient measures the strength of a *linear association*
  - Pearson's correlation is poorly suited for summarizing non-linear relationships

Correlation

From Cook & Swayne's *Interactive and Dynamic Graphics for Data Analysis*:

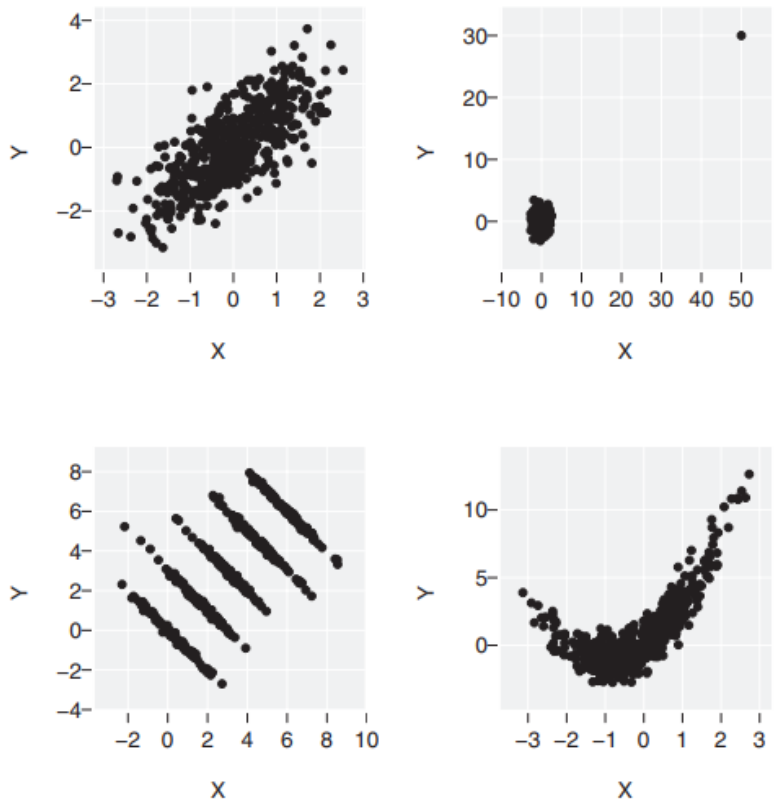
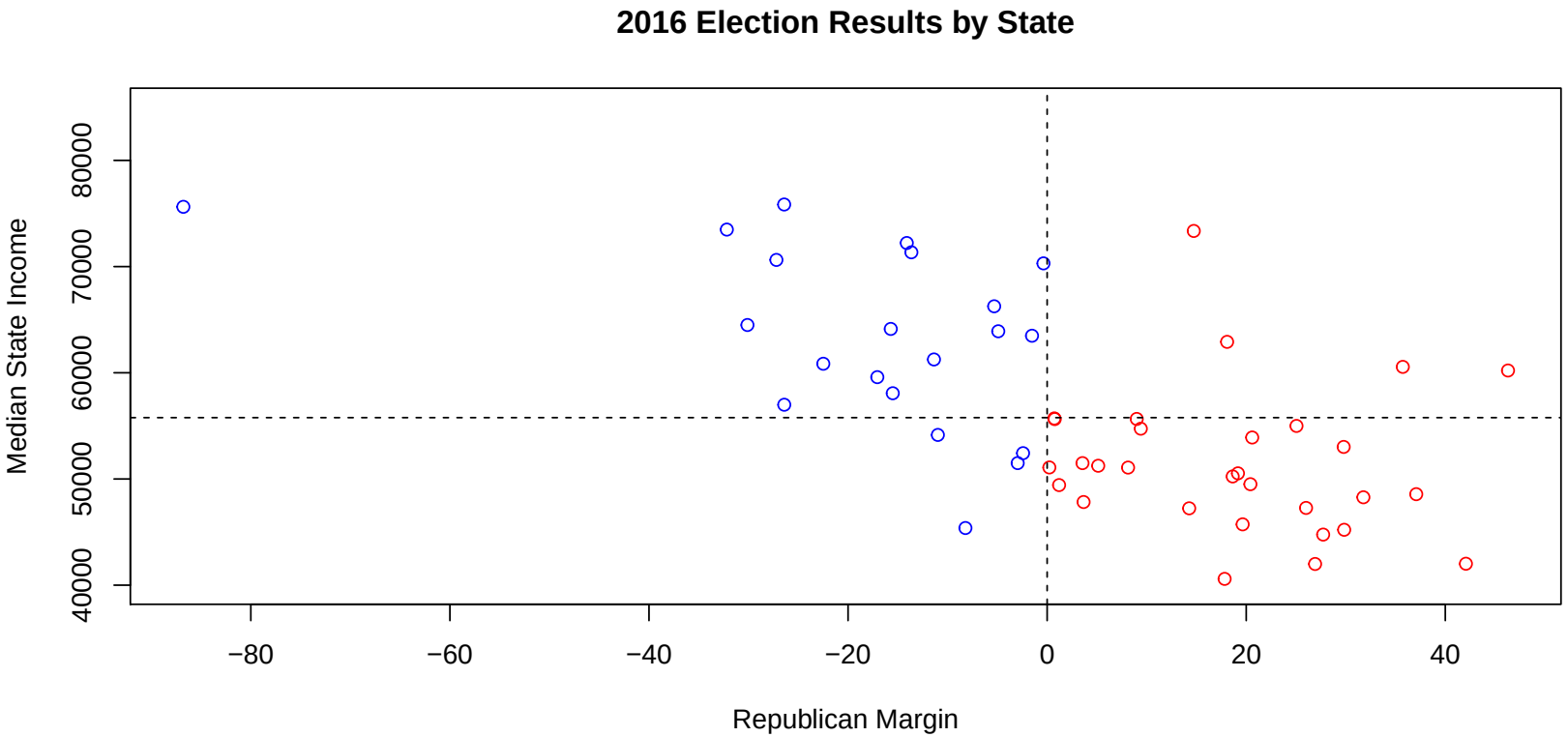


Fig. 6.1. Studying dependence between X and Y. All four pairs of variables have correlation approximately equal to 0.7, but they all have very different patterns. Only the top left plot shows two variables matching a dependence modeled by correlation.

Ecological Correlations

- **Ecological correlations** compare variables at an ecological level (ie: The cases are aggregated data - like countries or states)
- Let's look at the correlation between a US state's median household income and how that state voted in the 2016 presidential election

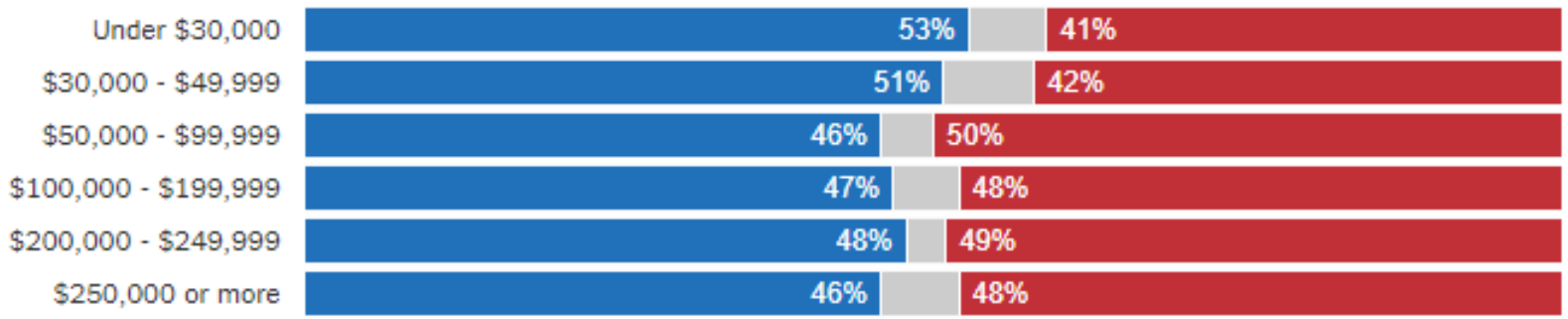
Ecological Correlations



- $r = -.63$, so do republicans earn lower incomes than democrats?

The Ecological Fallacy

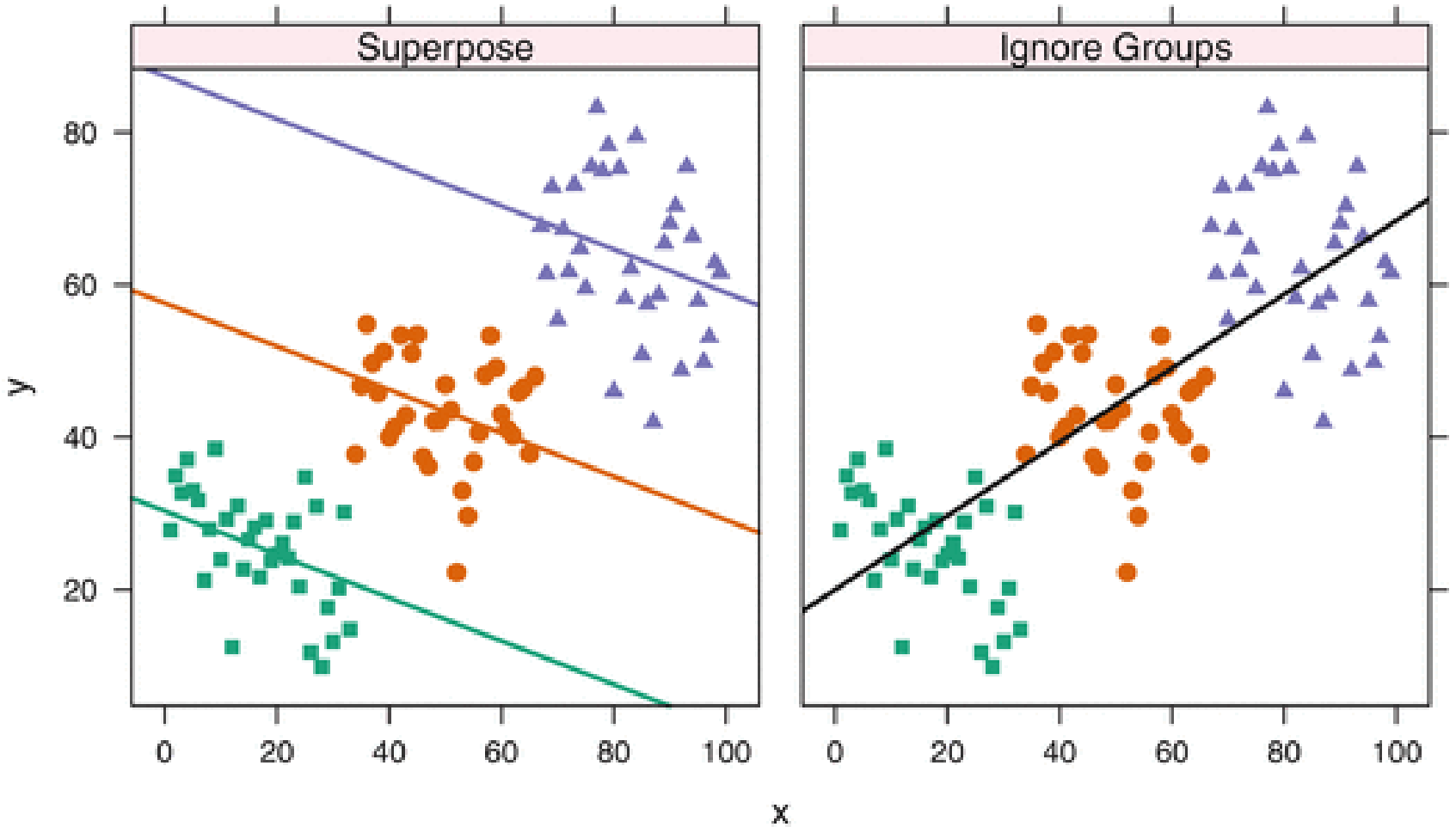
Using 2016 exit polls, conducted by the NY Times (Link), we can get a sense of how party vote and income are related *for individuals*:



- Looking at individuals as cases there is an opposite relationship between political party and income
- This "reversal" is an example of the **ecological fallacy**
  - Inferences about individuals cannot necessarily be deduced from inferences about the groups they belong to
  - The lesson here is we should use data where the cases align with who/what we're aiming to describe

Ecological Fallacy

The ecological fallacy can result from ignoring an important grouping variable:



Measuring Association (summary)

Association can be quantified numerically depending upon the types of the variables in question:

- For two categorical variables, association can be measured using **differences in proportions**
  - The proportion of tables with exactly 2 patrons is 0.174 higher for lunches than for dinners

Measuring Association (summary)

Association can be quantified numerically depending upon the types of the variables in question:

- For two categorical variables, association can be measured using **differences in proportions**
  - The proportion of tables with exactly 2 patrons is 0.174 higher for lunches than for dinners
- For one quantitative and one categorical variable, it can be measured using **differences in means**
  - The mean tip is \$1.6 higher for dinners than it is for lunches

Measuring Association (summary)

Association can be quantified numerically depending upon the types of the variables in question:

- For two categorical variables, association can be measured using **differences in proportions**
  - The proportion of tables with exactly 2 patrons is 0.174 higher for lunches than for dinners
- For one quantitative and one categorical variable, it can be measured using **differences in means**
  - The mean tip is \$1.6 higher for dinners than it is for lunches
- For two quantitative variables, it can be measured using the **correlation coefficient**
  - The correlation between tip and total bill is 0.676, suggesting higher bills are associated with higher tips
  - More info on the correlation coefficient is coming later in the course

Visualizing Association (summary)

Similarly, the best way to graphically display an association also depends upon the types of variables you're considering:

- Two categorical variables -> stacked barcharts
- One categorical variable and one quantitative variable -> boxplots
- Two quantitative variables -> scatterplots