Summarizing Data

Ryan Miller

Outline

1. Review vocabulary introduced during Lab #1 and discuss a few gray areas
2. Univariate graphical presentations of data
3. Univariate numerical summaries
4. Bivariate graphical presentations of data
5. Bivariate numerical summaries

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)

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, can you identify any interesting trends?

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

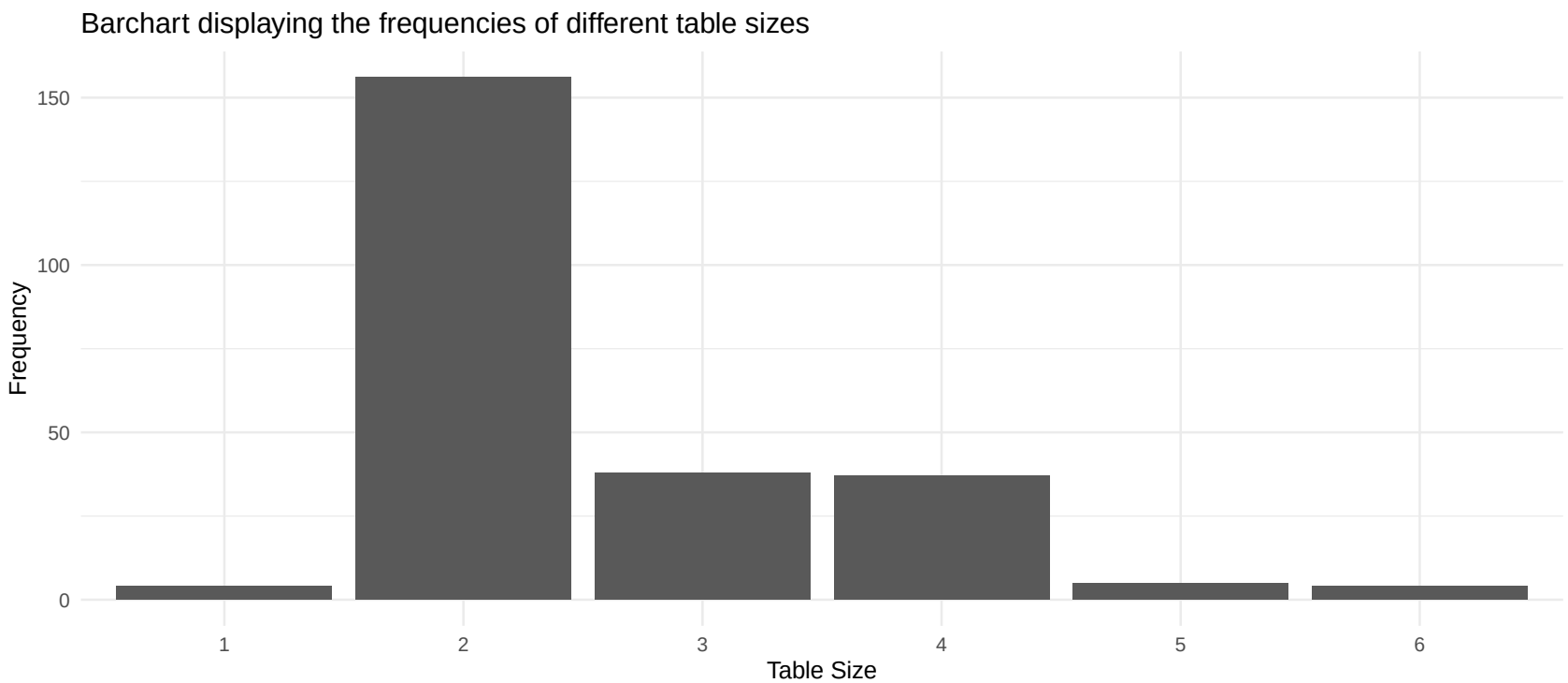
Data Presentation

Simply inspecting raw data is inefficient and is rarely useful, better strategies involve:

1. **Data visualization** - graphically displaying the data in ways that make trends more apparent
2. **Numerical summaries** - Deriving a single number, or small sets of numbers, that encapsulate certain aspects of the data

Univariate Graphs (One Categorical Variable)

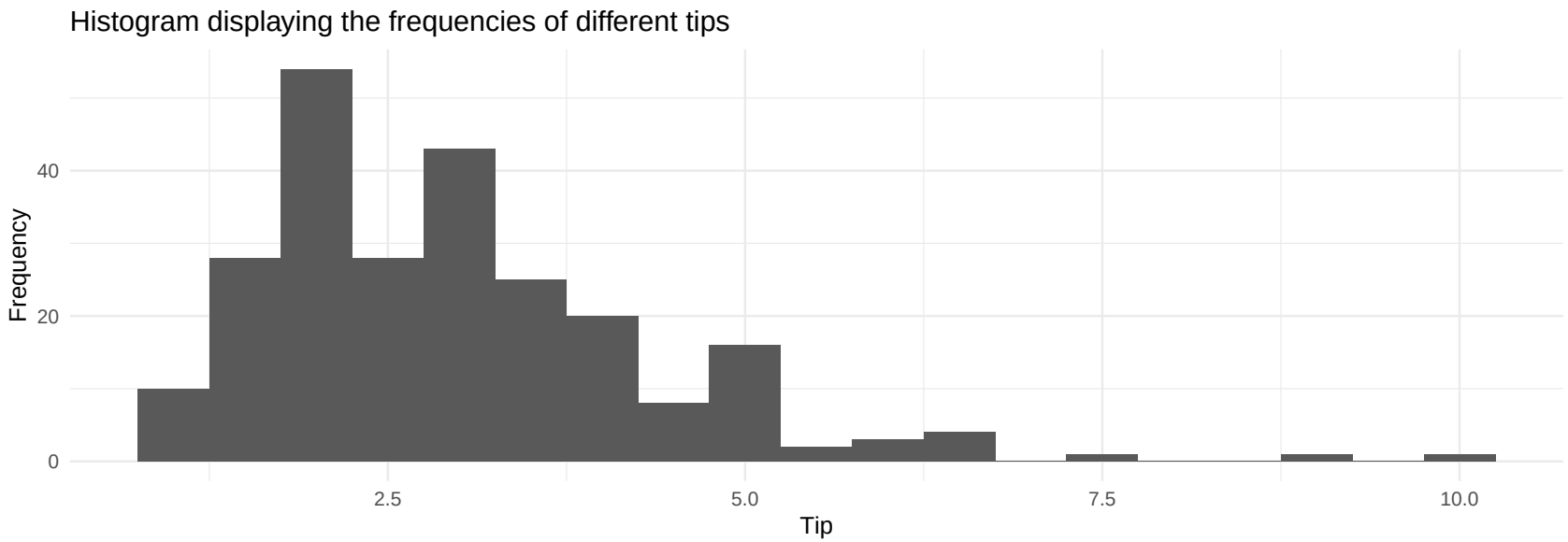
Barcharts are used to display a single categorical variable:



There aren't many technical terms or nuances involved in the *distribution* of a single categorical variable.

Univariate Graphs (One Numeric Variable)

Histograms categorize the numeric values of a quantitative variable into bins and display the frequencies of cases in each bin:

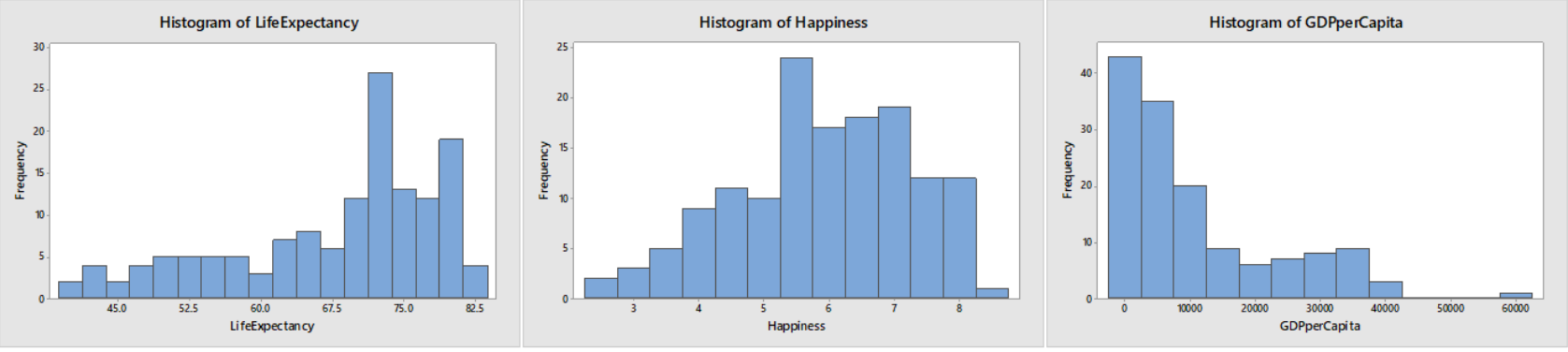


Key aspects of a variable's *distribution* visible in a histogram:

1. Shape (symmetric, skewed, bell-shaped, etc.)
2. Central Tendency (where are the data centered)
3. Variability (how spread out are the data)
4. Unusual data-points (outliers, excess zeros, etc.)

Distributional Shapes

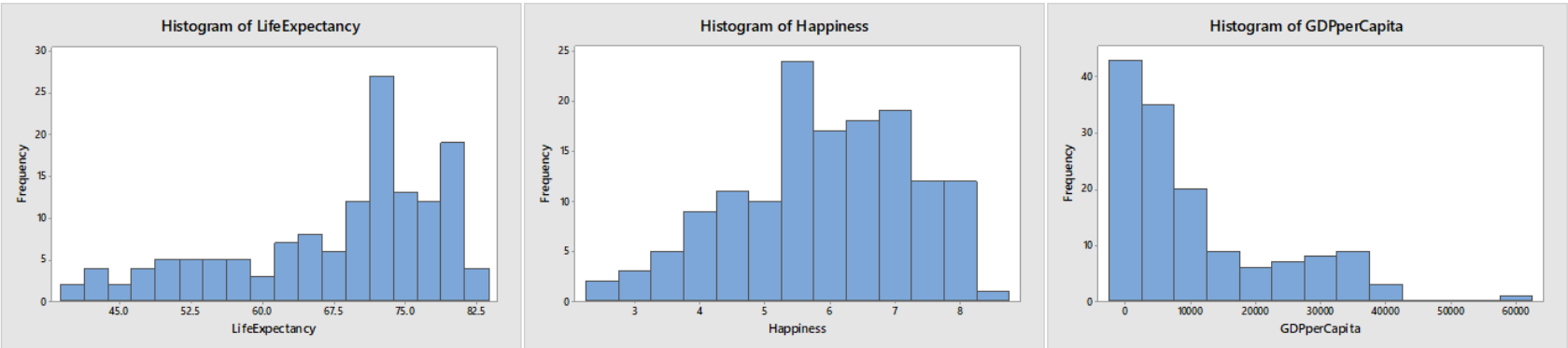
Depicted below are three important shapes:



- **Left skewed** - a long tail of smaller values (left panel)
- **Right skewed** - a long tail of larger values (right panel)

Distributional Shapes

Depicted below are three important shapes:



- **Left skewed** - a long tail of smaller values (left panel)
- **Right skewed** - a long tail of larger values (right panel)
- Happiness (center panel) is not skewed, it's also not perfectly **symmetric** or **bell-shaped**, but it's close enough that we might consider treating it *approximately bell-shaped*

Univariate Summaries of Categorical Variables

- There are really only two commonly used numerical summaries of categorical variables
  - **Frequencies**: counts of how many cases belong to a particular category
  - **Proportions**: fractions based upon frequencies, sometimes called *relative frequencies*
- These are usually organized in tabular form:

```
## A one-way frequency table
table(tips$day)
```

```
##
##  Fri  Sat  Sun Thur
##   19   87   76   62
```

```
## Proportions
prop.table(table(tips$day))
```

```
##
##        Fri        Sat        Sun       Thur
## 0.07786885 0.35655738 0.31147541 0.25409836
```

Univariate Summaries of Numeric Variables

Central Tendency:

- **Mean** - the arithmetic average of a numeric variable
  - $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ for *sample data* containing n cases
  - The mean can be influenced by outliers/skew

Univariate Summaries of Numeric Variables

Central Tendency:

- **Mean** - the arithmetic average of a numeric variable
  - $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ for *sample data* containing n cases
  - The mean can be influenced by outliers/skew
- **Median** - the 50th percentile of the data
  - The median is a *robust* measure of central tendency, it is not substantially influenced by outliers/skew

Univariate Summaries of Numeric Variables

Variability:

- **Variance** - the sum of squared deviations of the data from its average value
  - $s_x^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$ for *sample data* containing n cases

Univariate Summaries of Numeric Variables

Variability:

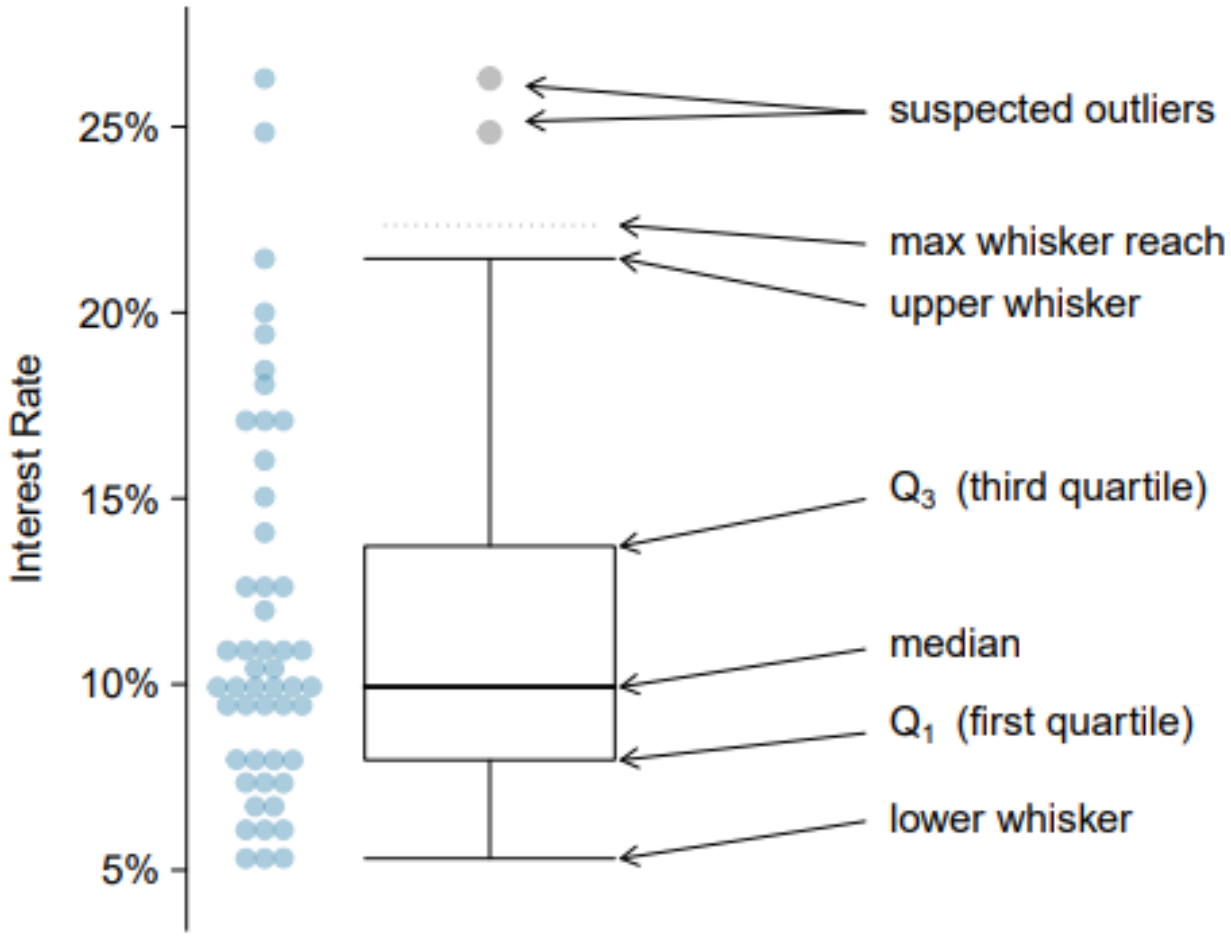
- **Variance** - the sum of squared deviations of the data from its average value
  - $s_x^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$ for *sample data* containing n cases
- **Standard Deviation** - the average deviation of the data from its average value (speaking loosely)
  - $s_x = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}}$ for *sample data* containing n cases
  - Standard deviation is *very sensitive* to outliers

Univariate Summaries of Numeric Variables

Variability:

- **Variance** - the sum of squared deviations of the data from its average value
  - $s_x^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$ for *sample data* containing n cases
- **Standard Deviation** - the average deviation of the data from its average value (speaking loosely)
  - $s_x = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}}$ for *sample data* containing n cases
  - Standard deviation is *very sensitive* to outliers
- **Interquartile Range** - the difference between the 75th (third quartile) and 25th (first quartile) percentiles of the data
  - The IQR is a *robust* measure of variability

Boxplots



Boxplots are a graphical presentation of summary statistics (not the data itself)

The 68-95-99 Rule

For bell-shaped distributions, the standard deviation describes the percentage of cases within a certain distance of the mean:

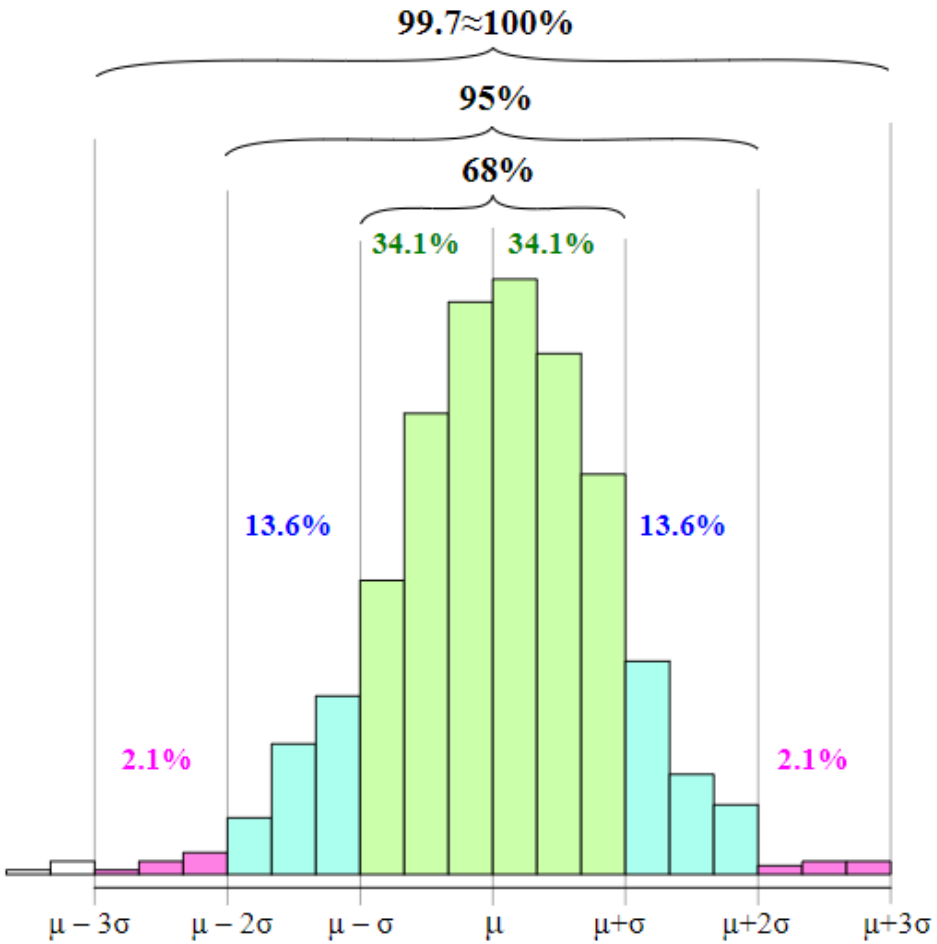


Image Source: https://en.wikipedia.org/wiki/68-95-99.7_rule

Practice #2

The "Diamonds" dataset (in the ggplot2 package) contains the sale prices and other attributes of nearly 54,000 diamonds sold by a large online retailer:

```
library(ggplot2)
data("diamonds")
summary(diamonds$price)
```

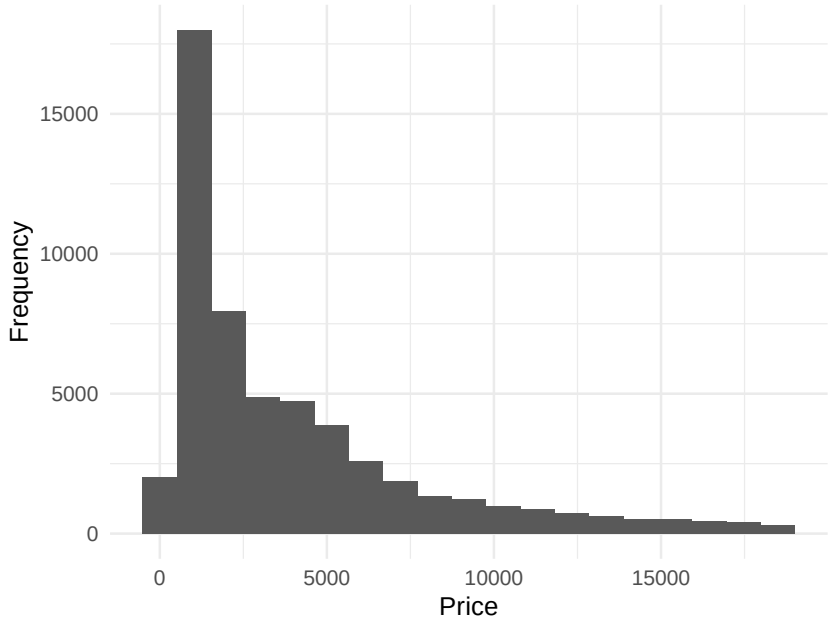
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     326     950    2401    3933    5324   18823
```

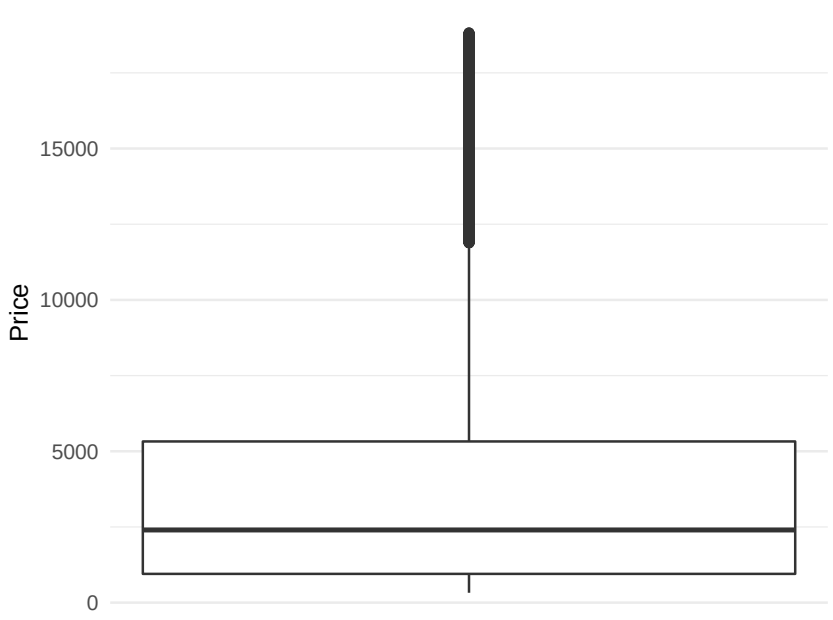
Use this summary information to answer the following:

1) Find and interpret the interquartile range of the variable "price".
2) Explain why the mean is so much larger than the median?

Practice #2 (continued)

Displayed below are a histogram and boxplot of the variable "price":





3) Describe the *shape* of the distribution of "price". Which graph is more helpful in judging shape?
4) How does the shape of this distribution relate to your answer to #2 (on the previous slide)?
5) Which graph is more helpful in judging the *center* and *variability* of "price"?

Practice #2 (solutions)

1) The IQR is $5324 - 950 = 4374$, this tells us that the middle 50% of diamonds in these data differed in price by at most \$4374
2) Initially, it seems like an outlier might be pulling the mean towards larger values
3) The distribution is very *right-skewed*, meaning it has a long tail of large values
4) Based upon the histogram, there isn't really any single distinct outlier, instead its the rightward skew that is influencing the mean
5) The boxplot shows the IQR and the median, so it is more helpful (while the histogram is more helpful for judging shape)

Relationships between Variables

Two variables, X and Y, are said to be **associated** if the values of X share a relationship with the values of Y

- Usually, we designate an **explanatory variable** (suspected cause) and a **response variable** (suspected outcome)
- This is often done based upon practical knowledge (ie: could time of day cause tip? could tip cause time of day?)

Relationships between Variables

Two variables, X and Y, are said to be **associated** if the values of X share a relationship with the values of Y

- Usually, we designate an **explanatory variable** (suspected cause) and a **response variable** (suspected outcome)
- This is often done based upon practical knowledge (ie: could time of day cause tip? could tip cause time of day?)

Note:

1. Association is general term, there many more specific types of association (ie: linear, non-linear, etc.)
2. Observing an association between X and Y does not imply that X causes Y, or that Y causes X, *causation* is a complex topic that we'll discuss next week

Two-way Frequency Tables

For comparisons of *two categorical variables*, two-way tables are the most straightforward approach to express associations:

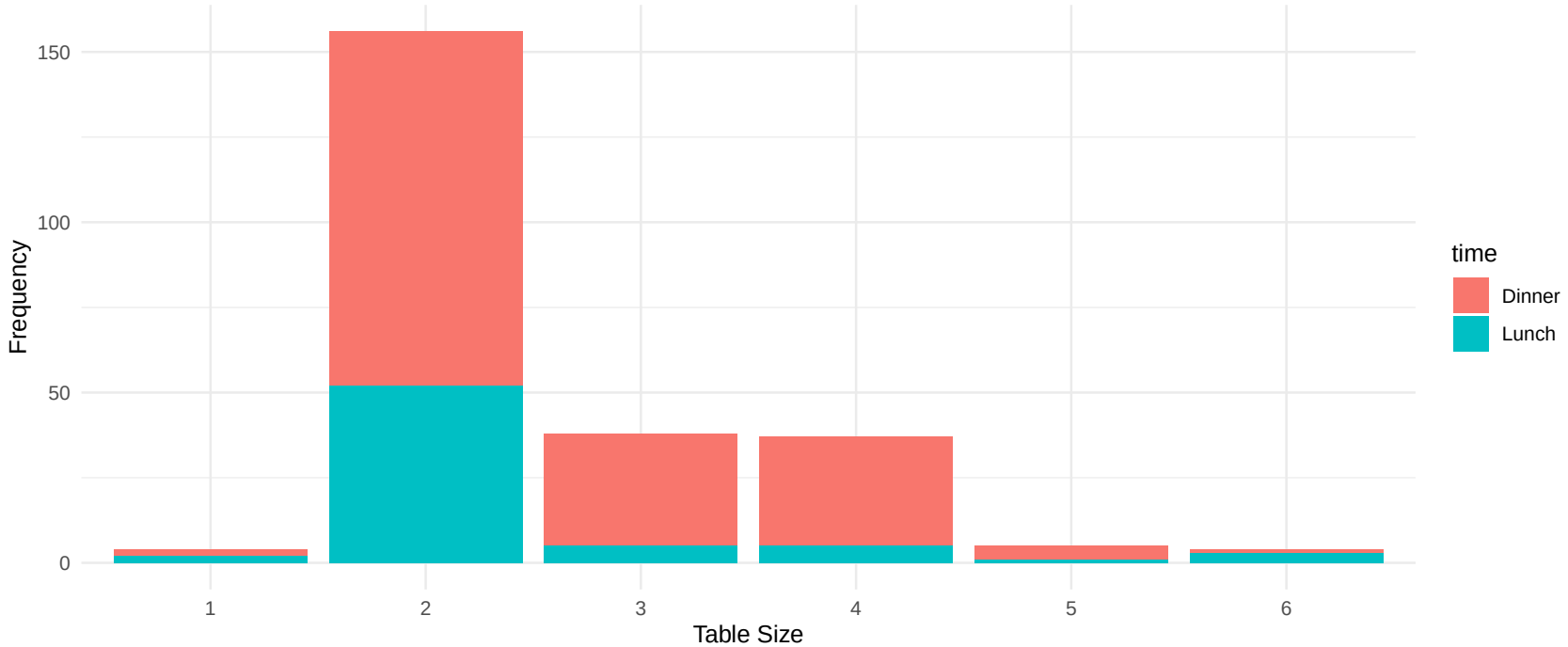
```
table(tips$day, tips$size)
```

```
##
##          1  2  3  4  5  6
##   Fri    1 16  1  1  0  0
##   Sat    2 53 18 13  1  0
##   Sun    0 39 15 18  3  1
##   Thur   1 48  4  5  1  3
```

- **Conditional proportions** (ie: row proportions or column proportions) are used to express observed associations
  - The proportion of 2-person tables on Fridays is $16/19 = 0.84$, while the same proportion on Sundays is only $39/76 = 0.51$
  - Because these proportions are different, "size" and "day" are associated

Bivariate Graphs (two categorical variables)

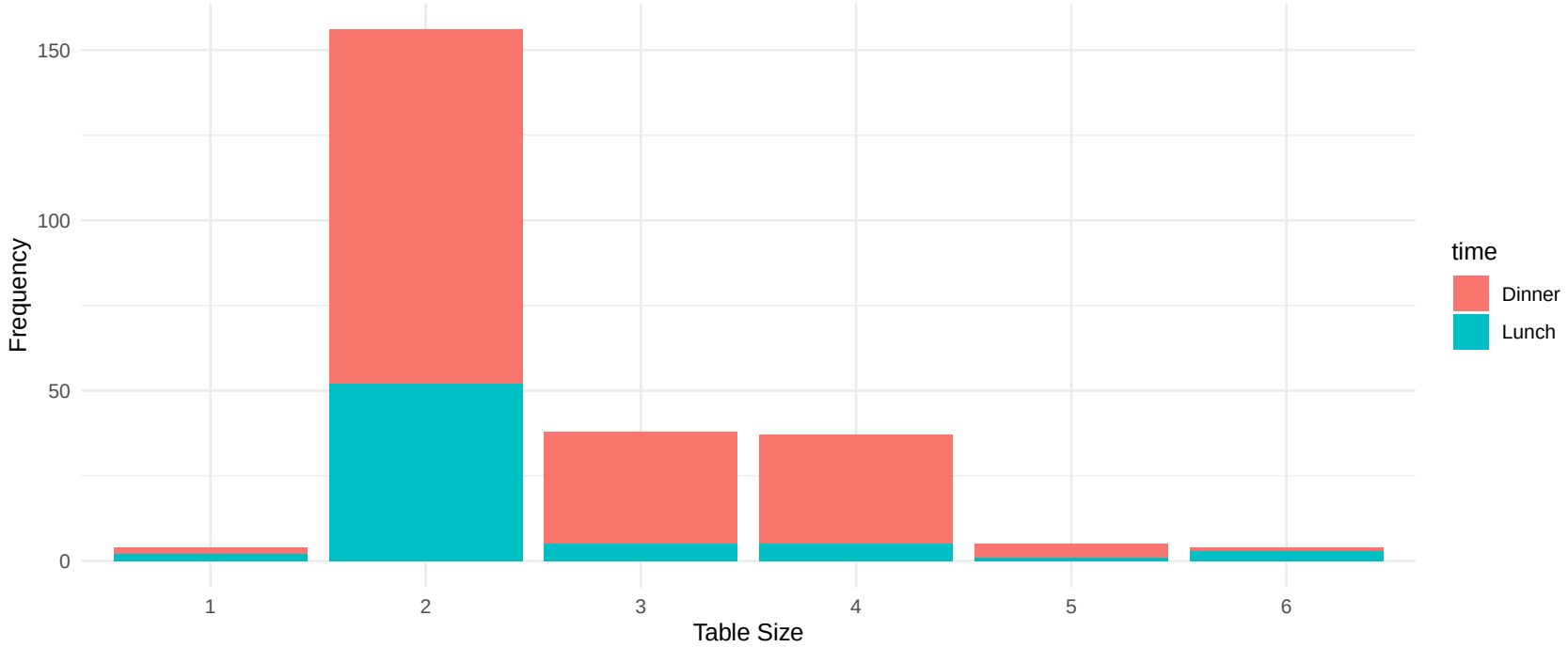
Stacked barcharts are used to visually display relationships between two categorical variables



In the "Tips" dataset, do table size and time of day appear associated?

Bivariate Graphs (two categorical variables)

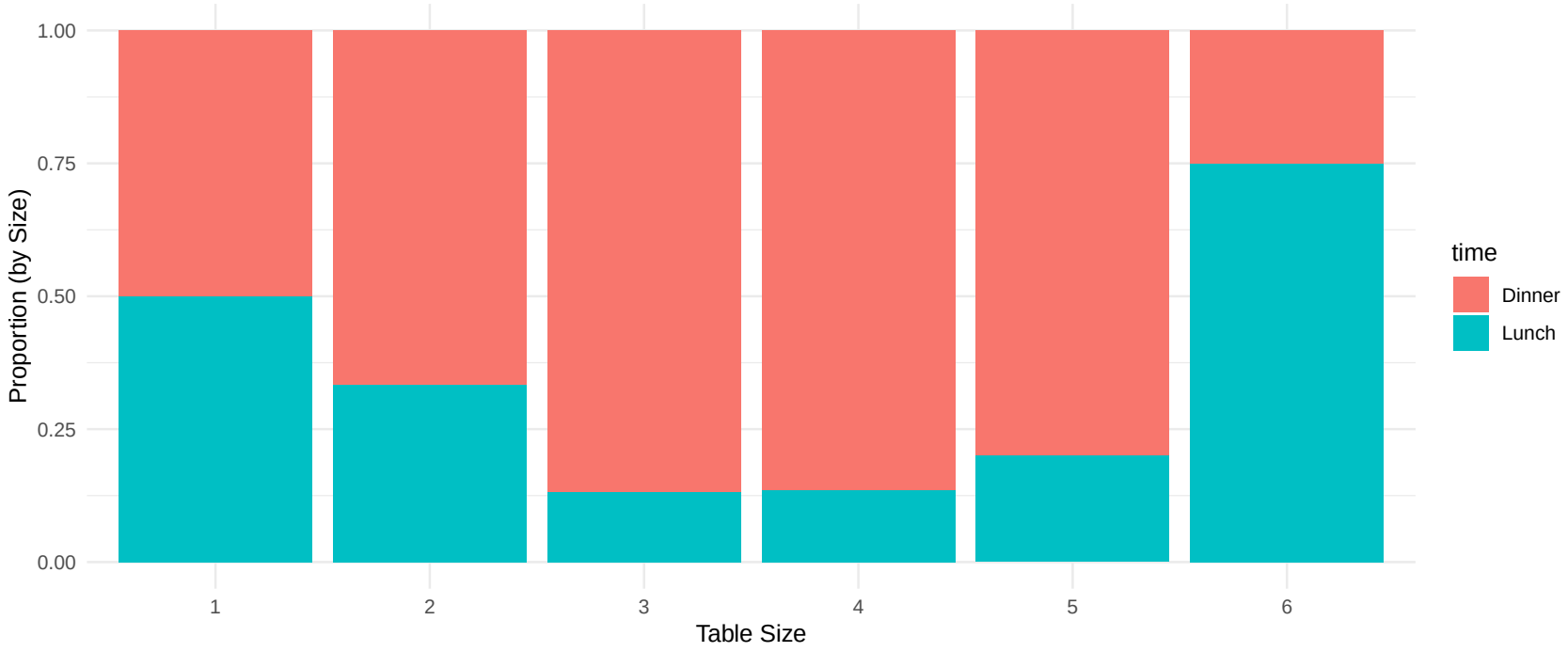
Stacked barcharts are used to visually display relationships between two categorical variables



In the "Tips" dataset, do table size and time of day appear associated? Yes, two person tables are relatively more common at lunch

Bivariate Graphs (two categorical variables)

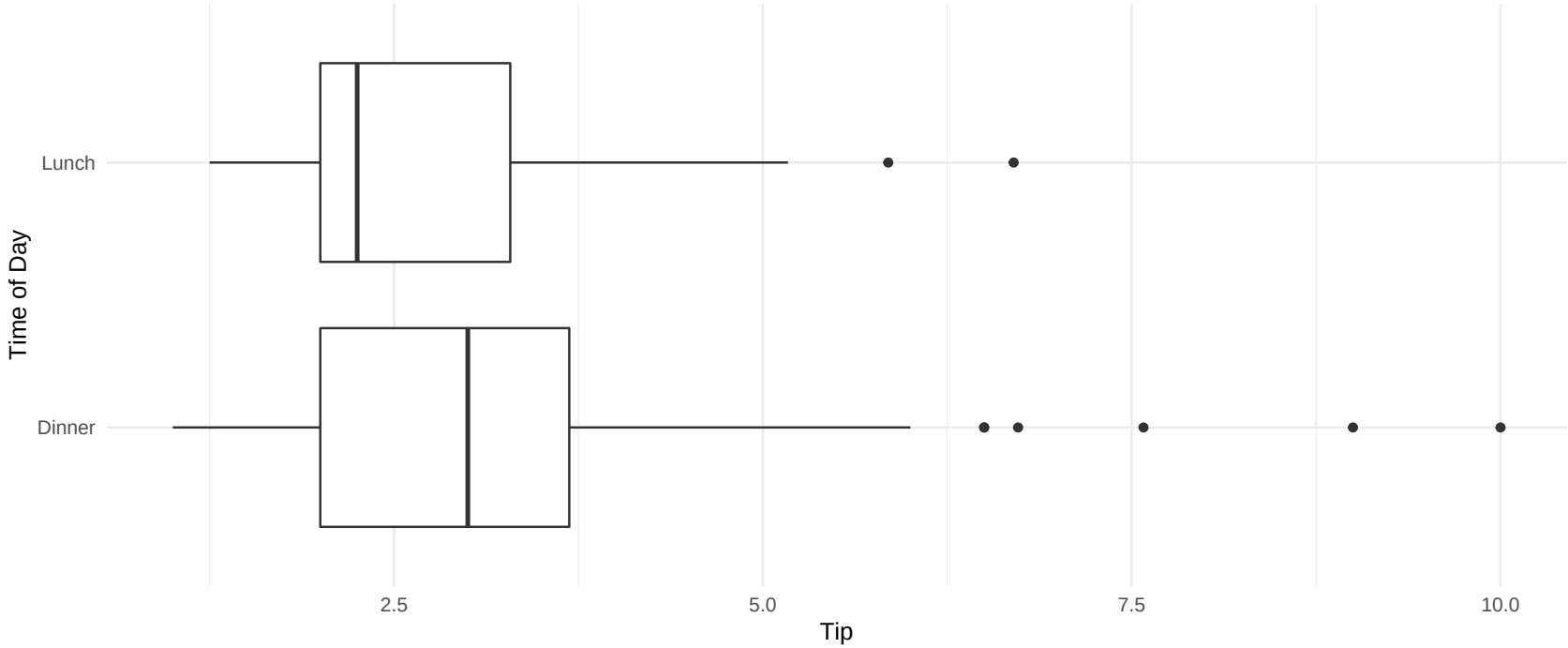
Stacked conditional barcharts are can be more useful in displaying associations



However, you should be careful not read too much into sparsely populated categories (such as 1 or 6 person tables)

Bivariate Graphs (one categorical and one numeric variable)

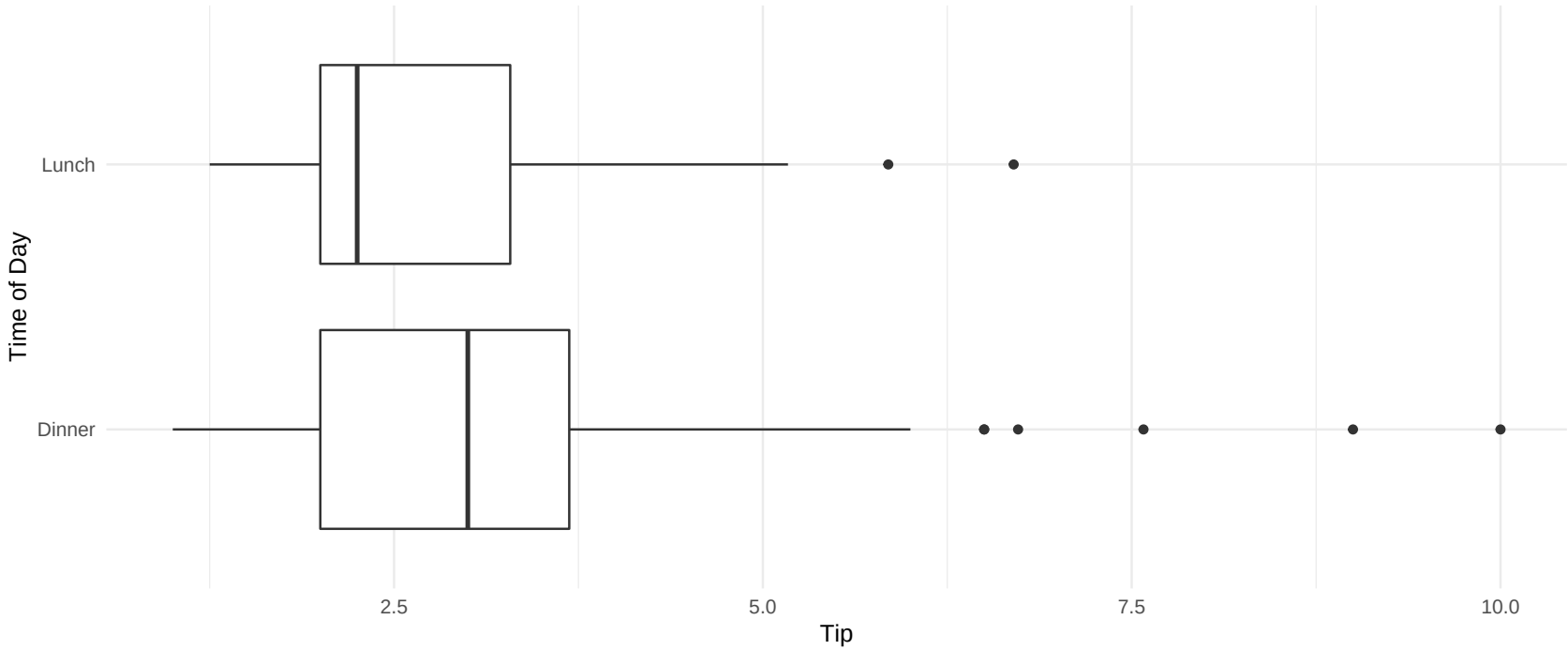
Side-by-side univariate graphs are used to graphically depict relationships between one categorical and one numeric variable



In the "Tips" dataset, do tip and time of day appear associated?

Bivariate Graphs (one categorical and one numeric variable)

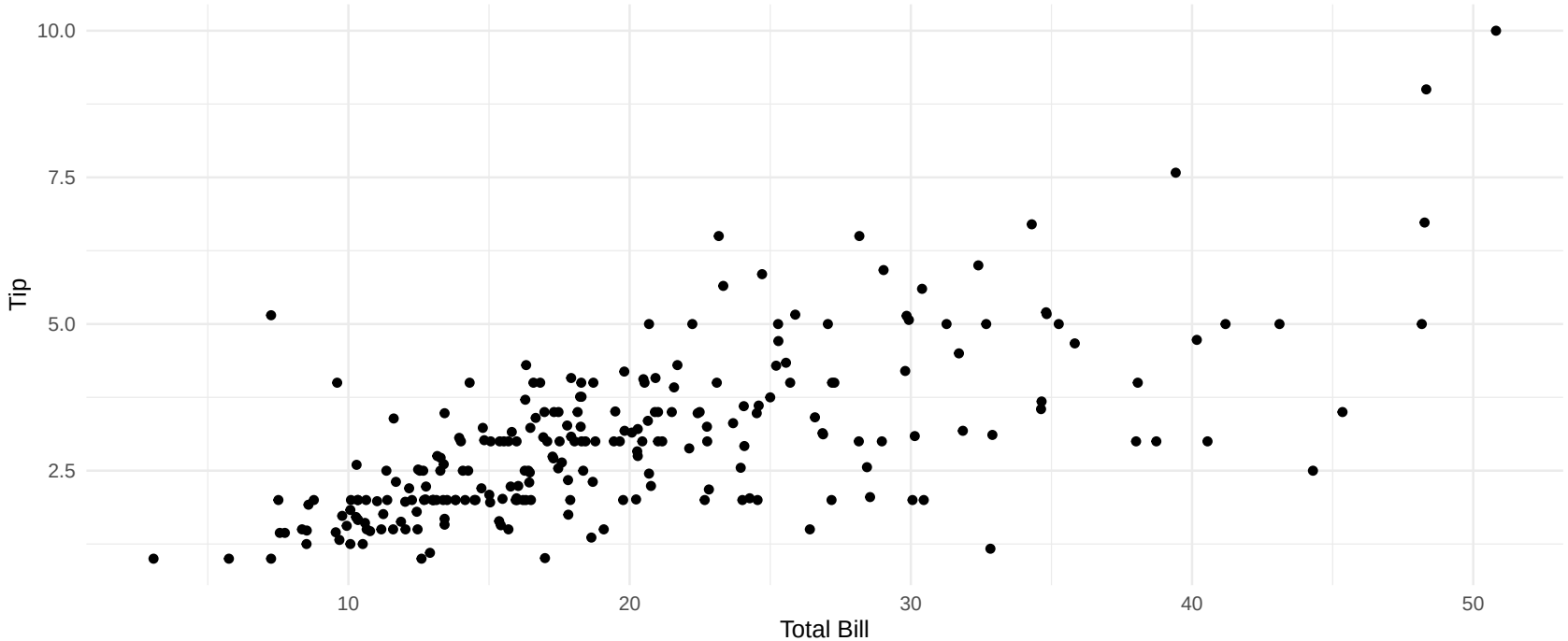
Side-by-side univariate graphs are used to graphically depict relationships between one categorical and one numeric variable



In the "Tips" dataset, do tip and time of day appear associated? Yes, tips tend to be larger at dinner

Bivariate Graphs (two numeric variables)

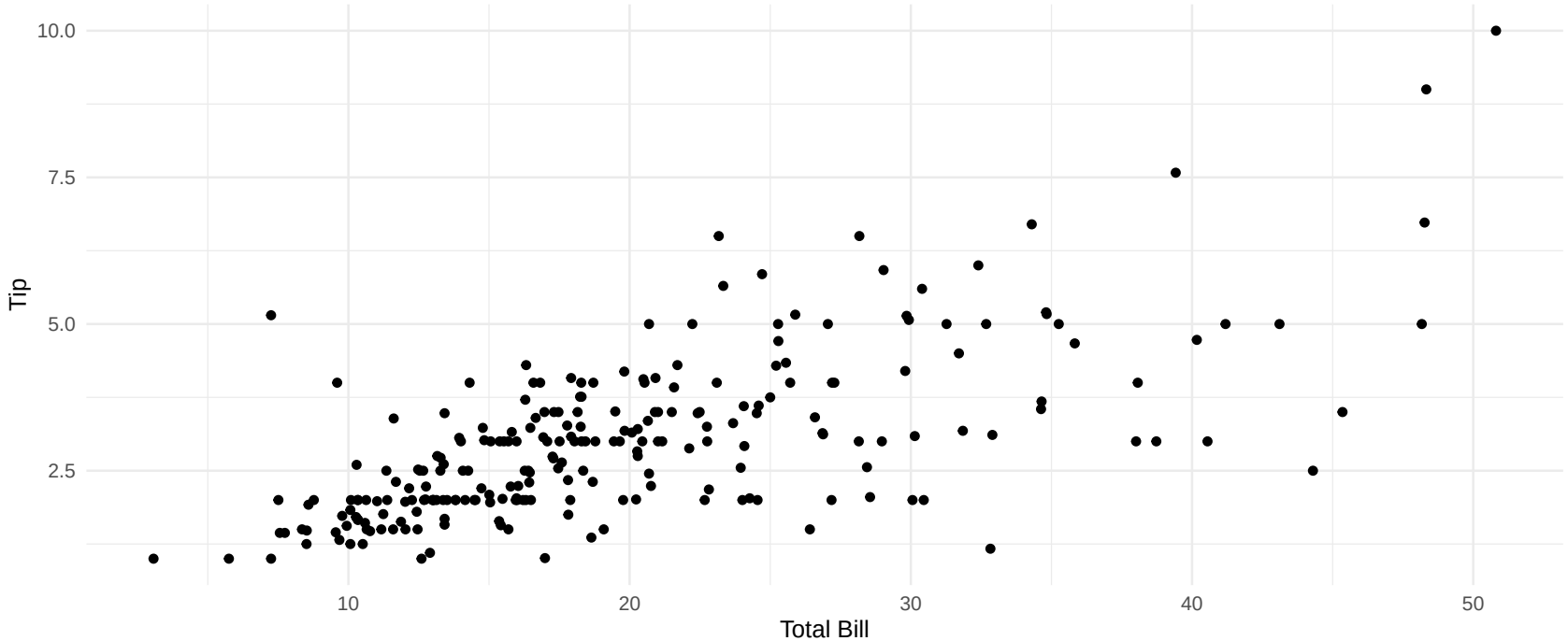
Scatterplots are used to graphically display two numeric variables



In the "Tips" dataset, do total bill and tip appear associated?

Bivariate Graphs (two numeric variables)

Scatterplots are used to graphically display two numeric variables



In the "Tips" dataset, do total bill and tip appear associated? Yes, there is a *moderate*, *positive*, *linear* relationship

Describing Associations (two numeric variables)

Relationships between two numeric variables can be qualitatively expressed in terms of the following factors:

1) **Form** - what type of trend or pattern do the data tend to follow
2) **Strength** - how closely do the data follow that trend or pattern
3) **Direction** - do larger values of X tend to correspond to larger values of Y (positive) or smaller values of Y (negative)

Describing Associations (two numeric variables)

Relationships between two numeric variables can be qualitatively expressed in terms of the following factors:

1) **Form** - what type of trend or pattern do the data tend to follow
2) **Strength** - how closely do the data follow that trend or pattern
3) **Direction** - do larger values of X tend to correspond to larger values of Y (positive) or smaller values of Y (negative)

Note: Directions like "positive" or "negative" only make sense for certain forms of association (ie: linear), for other forms you might choose adjectives like "increasing" or "decreasing", or you might not be able to describe a single direction

Pearson's Correlation Coefficient

Pearson's correlation coefficient summarizes the *strength* of a *linear relationship* between two numeric variables, X and Y:

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

Notes:

1) Pearson's correlation is poorly suited for non-linear associations
2) If no qualifier is given, "correlation" refers to a specific type of association (ie: linear, numeric variables)

Correlation

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

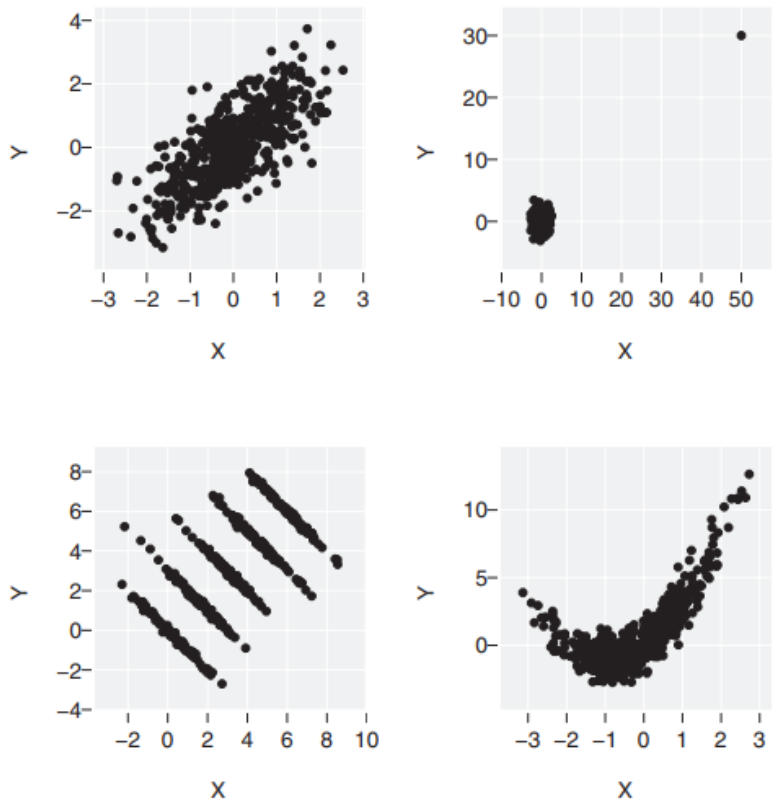


Fig. 6.1. Studying dependence between X and Y. All four pairs of variables have correlation approximately equal to 0.7, but they all have very different patterns. Only the top left plot shows two variables matching a dependence modeled by correlation.

Robust and Non-linear Correlations

- Nonlinear correlations (ie: the strength of quadratic/polynomial patterns) can be found using the `nlcor` package in R (not a topic we'll focus on)
- Spearman and Kendall's rank correlations are robust alternatives to Pearson's correlation available in R's `cor()` function

```
cor(x = tips$total_bill, y = tips$tip, method = "pearson")
```

```
## [1] 0.6757341
```

```
cor(x = tips$total_bill, y = tips$tip, method = "spearman")
```

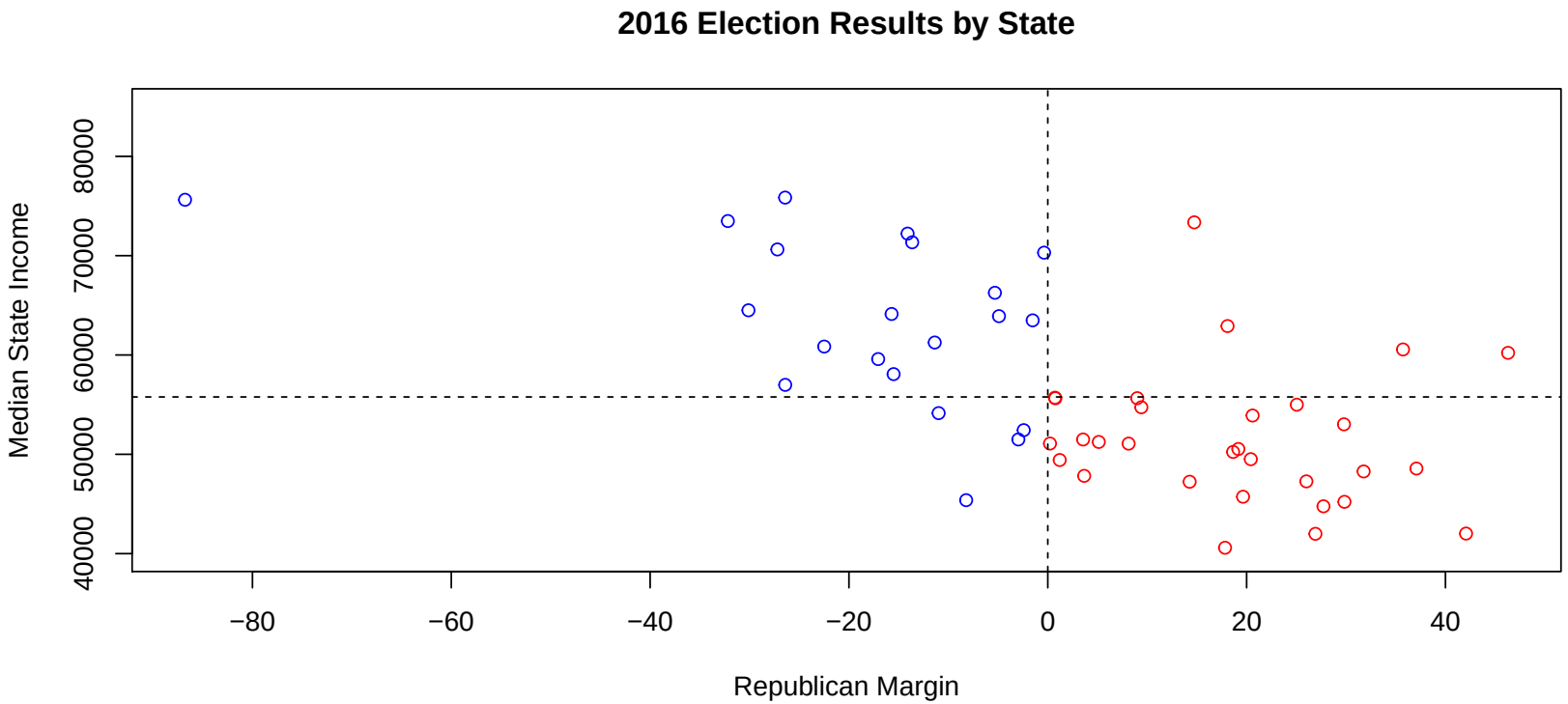
```
## [1] 0.6789681
```

```
cor(x = tips$total_bill, y = tips$tip, method = "kendall")
```

```
## [1] 0.517181
```

Ecological Correlations

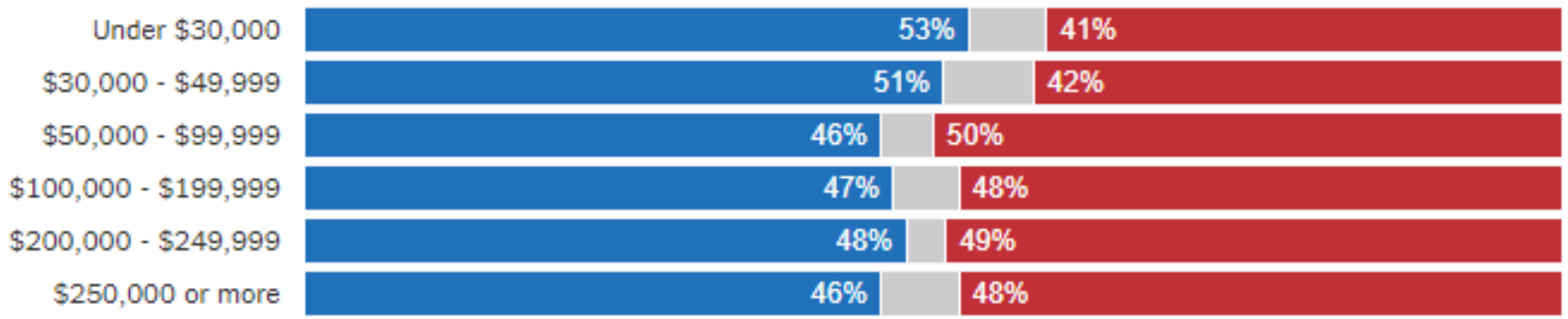
Ecological correlations compare variables at an ecological level (ie: The cases are aggregated data - like countries or states)



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

The Ecological Fallacy

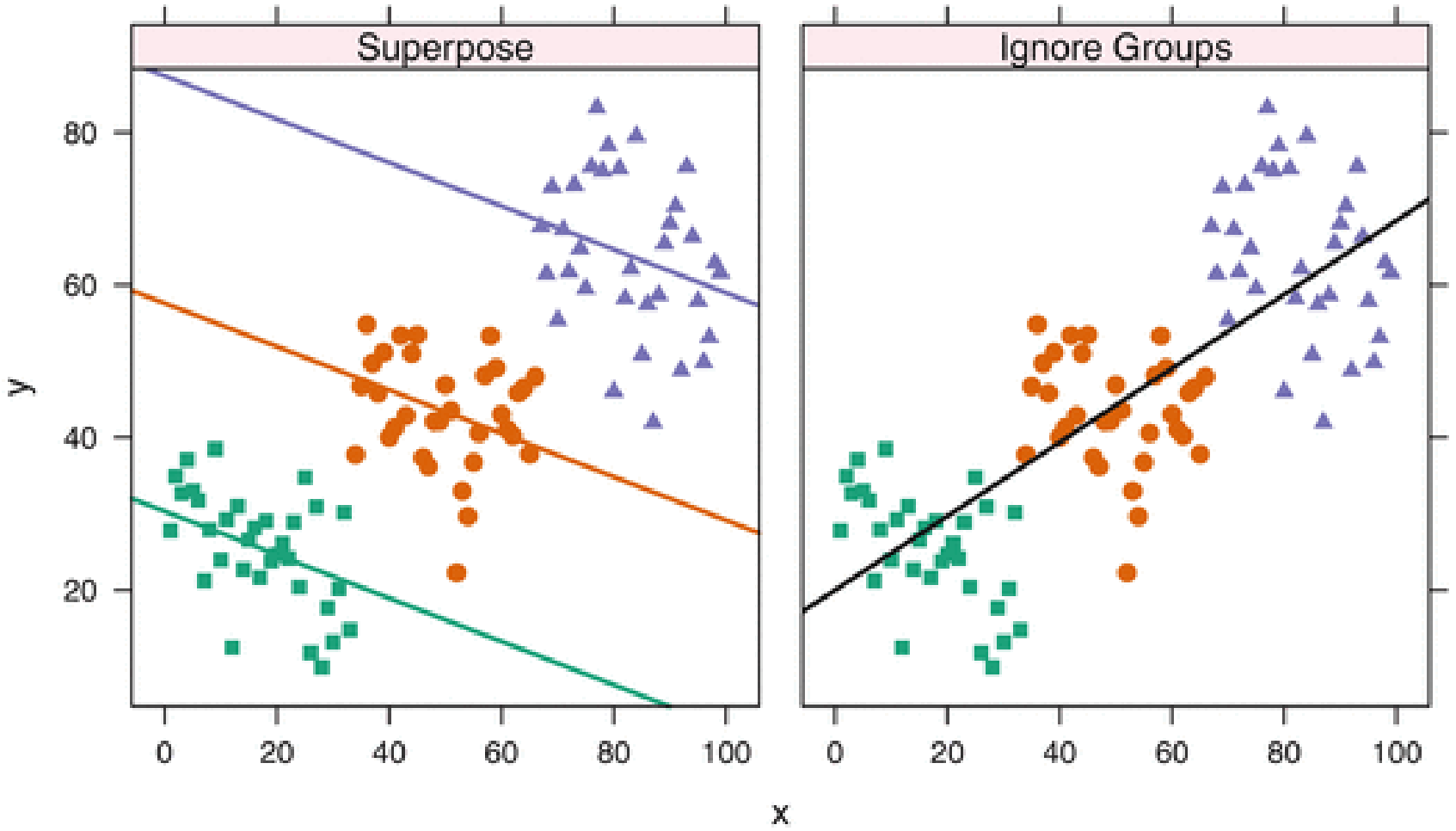
Using 2016 exit polls, conducted by the NY Times (Link), we can get a sense of how party vote and income are related *for individuals*:



- The "reversal" that occurs when considering individuals rather than states 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:



Summary - Measuring Association

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

- Two categorical variables - **differences in proportions**
  - The proportion of tables with exactly 2 patrons is 0.174 higher for lunches than for dinners

Summary - Measuring Association

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

- Two categorical variables - **differences in proportions**
  - The proportion of tables with exactly 2 patrons is 0.174 higher for lunches than for dinners
- One quantitative and one categorical variable - **differences in means**
  - The mean tip is \$1.6 higher for dinners than it is for lunches

Summary - Measuring Association

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

- Two categorical variables - **differences in proportions**
  - The proportion of tables with exactly 2 patrons is 0.174 higher for lunches than for dinners
- One quantitative and one categorical variable - **differences in means**
  - The mean tip is \$1.6 higher for dinners than it is for lunches
- Two quantitative variables - **correlation coefficient**
  - The correlation between tip and total bill is 0.676, suggesting higher bills are associated with higher tips

Summary - Visualizing Association

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

- Two categorical variables - **stacked barcharts** (possibly conditional)
  - Focus on the more populated categories when interpreting
- One categorical variable and one quantitative variable - **side-by-side boxplots**
  - Compare the corresponding quartiles, medians, etc. when interpreting
- Two quantitative variables - **scatterplots**
  - Discuss the form, strength, and direction when interpreting