Data Analytics with R

Sumit Mishra

Institute for Financial Management and Research, Sri City

Summarizing Data

04 November 2020

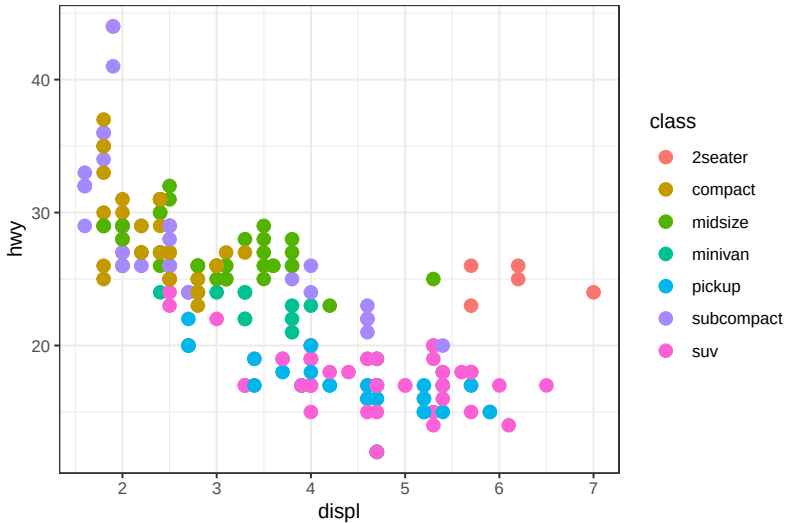
Agenda

- Examining numerical data
- Examining categorical data
- Case study: Racial discrimination in job-market.
- Material: Textbook chapter 02.

Examining Numerical Data

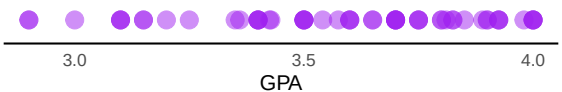
Scatterplot for paired data

Scatterplots are useful for visualizing the relationship between two numerical variables.

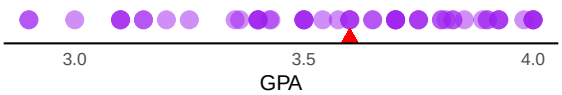


Dot plot

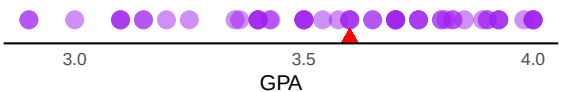
These are useful to visualize one numerical variable. Darker colors represent greater concentration of observations.



Dot plot with mean

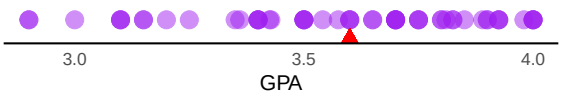


Dot plot with mean



- The mean represents the center of the distribution.
- The mean GPA is

Dot plot with mean



- The mean represents the center of the distribution.
- The mean GPA is 3.59.

Exercise

Construct dot plot for the variable `interest_rate` from the dataset `loan50`.

Exercise

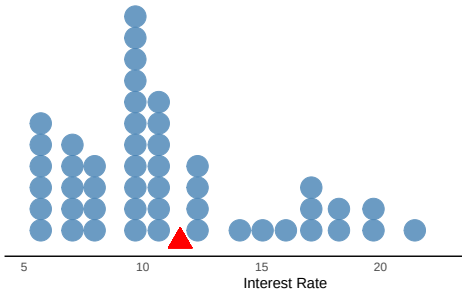
Construct dot plot for the variable `interest_rate` from the dataset `loan50`.

```
loan50 %>%
ggplot(aes(x =interest_rate)) +
geom_dotplot(fill = "steelblue", color = "steelblue",
             alpha = 0.8, binwidth = 0.9) +
guides(y = "none") +
labs(y = NULL, x = "Interest Rate") +
scale_x_continuous(breaks = c(5, 10, 15, 20, 25)) +
theme_minimal() +
geom_point(aes(x = mean(interest_rate), y =0),
           shape = 17, size = 6, color = "red") +
theme(axis.line.x = element_line(),
      panel.grid = element_blank())
```

Exercise

Construct dot plot for the variable `interest_rate` from the dataset `loan50`.

```
loan50 %>%
ggplot(aes(x =interest_rate)) +
geom_dotplot(fill = "steelblue", color = "steelblue",
             alpha = 0.8, binwidth = 0.9) +
guides(y = "none") +
labs(y = NULL, x = "Interest Rate") +
scale_x_continuous(breaks = c(5, 10, 15, 20, 25)) +
theme_minimal() +
geom_point(aes(x = mean(interest_rate), y =0),
           shape = 17, size = 6, color = "red") +
theme(axis.line.x = element_line(),
      panel.grid = element_blank())
```



Mean

The sample mean, denoted as $\bar{x}$, can be calculated as $\bar{x} = \frac{x_1+x_2+....+x_n}{n}$

The population mean, denoted by μ, is an elusive quantity.

We assume that $\bar{x}$ (an estimate of the population mean) is a good way to represent μ.

Histogram

Interest Rate	5.0% - 7.5%	7.5% - 10.0%	10.0% - 12.5%	12.5% - 15.0%	···	25.0% - 27.5
Count	11	15	8	4	···	1

Figure 1: Counts for the binned `interest_rate` data.

Histogram

Interest Rate	5.0% - 7.5%	7.5% - 10.0%	10.0% - 12.5%	12.5% - 15.0%	···	25.0% - 27.5
Count	11	15	8	4	···	1

Figure 1: Counts for the binned `interest_rate` data.

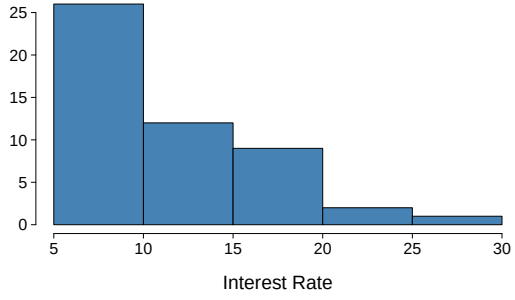
```
hist(loan50$interest_rate,
       col = "steelblue", main =
NULL,
        xlab = "Interest Rate",
ylab = NULL)

```

Histogram

Interest Rate	5.0% - 7.5%	7.5% - 10.0%	10.0% - 12.5%	12.5% - 15.0%	···	25.0% - 27.5
Count	11	15	8	4	···	1

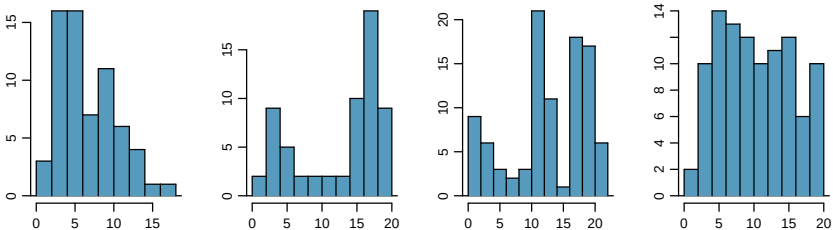
Figure 1: Counts for the binned `interest_rate` data.

```
1  hist(loan50$interest_rate,
2       col = "steelblue", main =
   NULL,
3       xlab = "Interest Rate",
   ylab = NULL)
4
```



Shape of a distribution: Modality

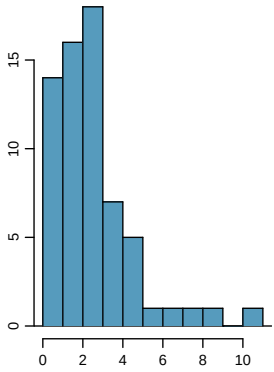
Does the histogram have a single prominent peak (*unimodal*), several prominent peaks (*bimodal/multimodal*), or no apparent peaks (*uniform*)?

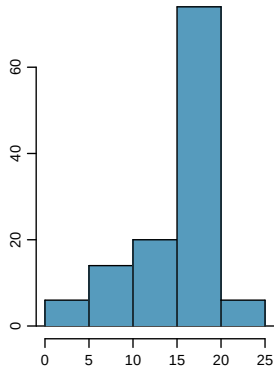


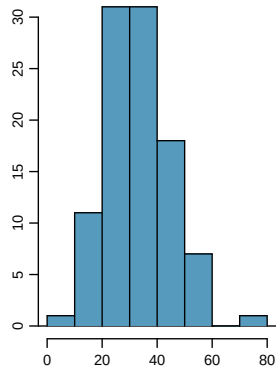
Note: In order to determine modality, step back and imagine a smooth curve over the histogram – imagine that the bars are wooden blocks and you drop a limp spaghetti over them, the shape the spaghetti would take could be viewed as a smooth curve.

Shape of a distribution: skewness

Is the histogram *right skewed*, *left skewed*, or *symmetric*?



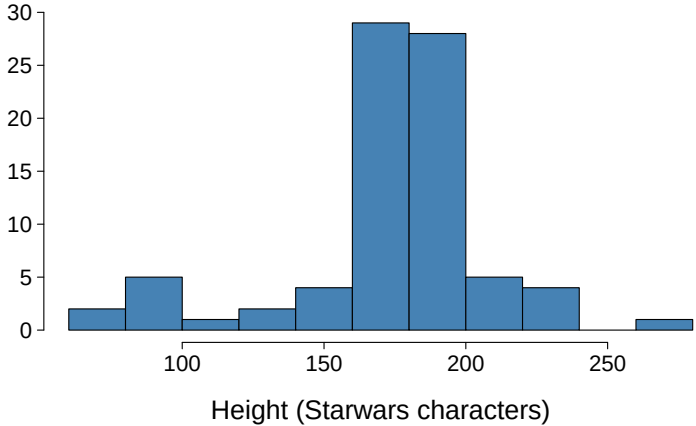




Note: Histograms are said to be skewed to the side of the long tail.

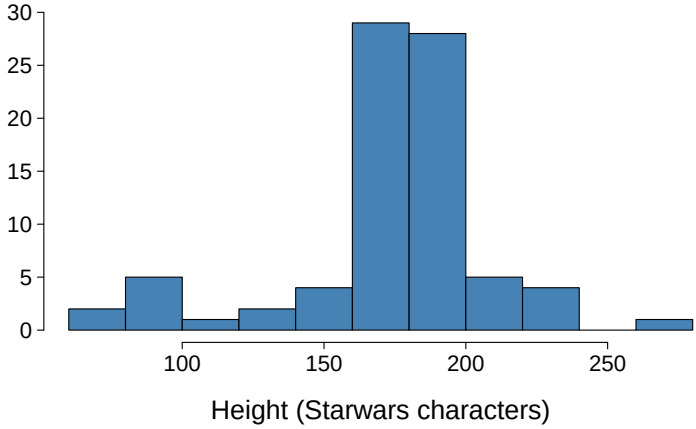
Shape of a distribution: unusual observations

Are there any unusual observations or potential *outliers*?



Shape of a distribution: unusual observations

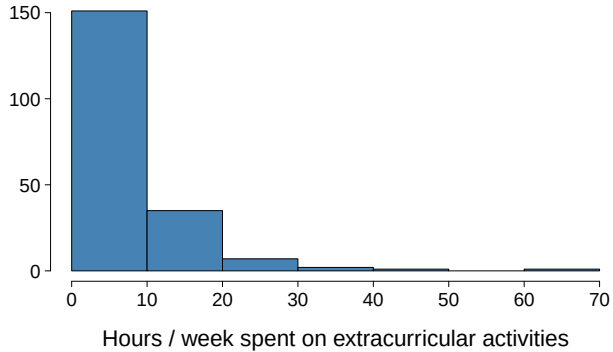
Are there any unusual observations or potential *outliers*?



Plot the variable `mass` from `starwars` dataset, and look for outliers.

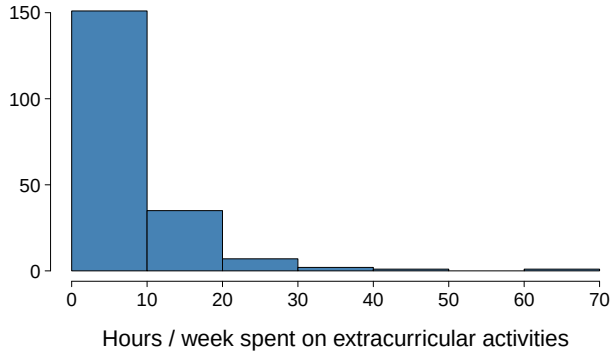
Extracurricular activities

How would you describe the shape of the distribution of hours per week students spend on extracurricular activities?



Extracurricular activities

How would you describe the shape of the distribution of hours per week students spend on extracurricular activities?



Unimodal and right skewed, with an outlier at 60 hours/week.

Commonly observed shapes of distributions

- modality

Commonly observed shapes of distributions

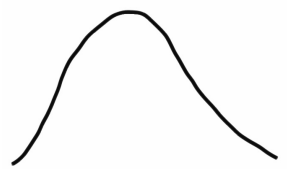
- modality

unimodal

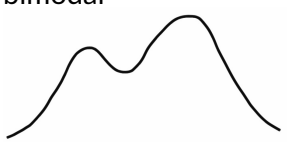
Commonly observed shapes of distributions

- modality

unimodal



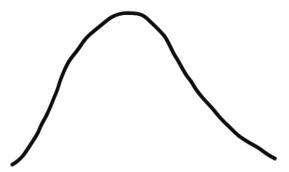
bimodal



Commonly observed shapes of distributions

- modality

unimodal



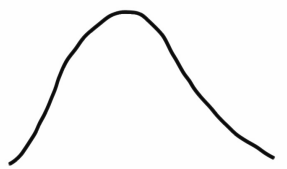
bimodal

multimodal

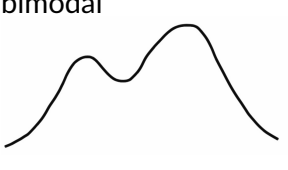
Commonly observed shapes of distributions

- modality

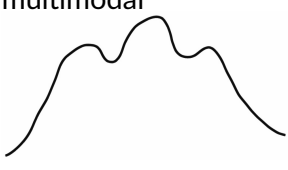
unimodal



bimodal



multimodal

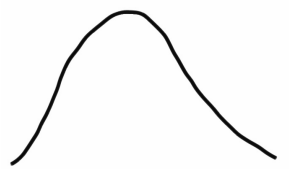


uniform

Commonly observed shapes of distributions

- modality

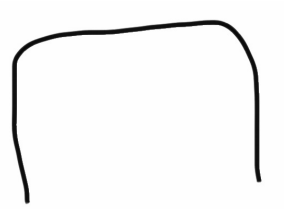
unimodal



bimodal

multimodal

uniform

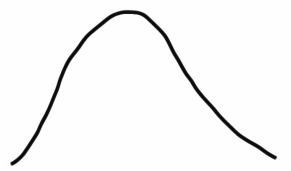


- skewness

Commonly observed shapes of distributions

- modality

unimodal



bimodal

multimodal

uniform

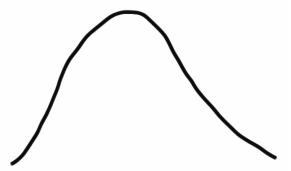
- skewness

right skew

Commonly observed shapes of distributions

- modality

unimodal



bimodal

multimodal

uniform

- skewness

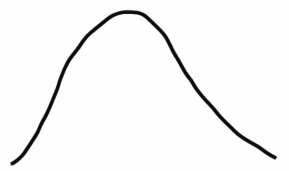
right skew

left skew

Commonly observed shapes of distributions

- modality

unimodal



bimodal

multimodal

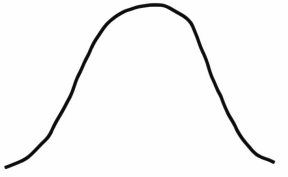
uniform

- skewness

right skew

left skew

symmetric



Variance

The variance, denoted as $V(x)$, can be calculated as $V(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

Variance

The variance, denoted as $V(x)$, can be calculated as $V(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

Compute the variance for the variable `height` from the dataset `starwars` with and without using the canned function.

Variance

The variance, denoted as $V(x)$, can be calculated as $V(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

Compute the variance for the variable `height` from the dataset `starwars` with and without using the canned function.

- We square up the deviations so that only the distance (and not the direction) matters.
- Larger deviations (irrespective of the direction) get greater weightage.

Standard Deviation

The standard deviation is the square root of the variance $V(x)$, and has the same units as the data

$s = \sqrt{V(x)}$

Median

- The *median* is the value that splits the data in half when ordered in ascending order.

$$0, 1, 2, 3, 4$$

- If there are an even number of observations, then the median is the average of the two values in the middle.

$$0, 1, \underline{2, 3}, 4, 5 \rightarrow \frac{2+3}{2} = 2.5$$

- Since the median is the midpoint of the data, 50% of the values are below it. Hence, it is also the 50^{th} *percentile*.

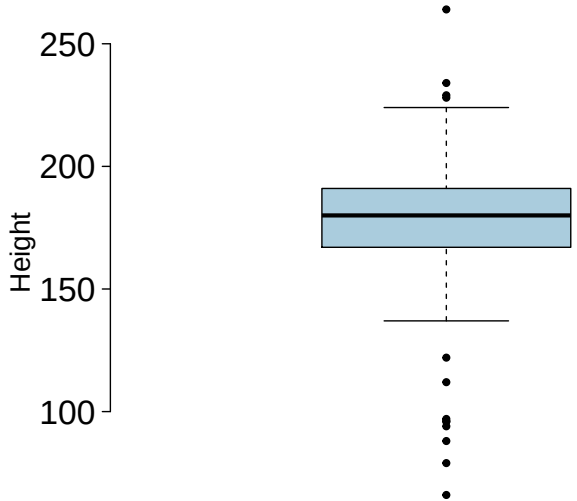
Q1, Q3, and IQR

- The 25^{th} percentile is also called the first quartile, *Q1*.
- The 50^{th} percentile is also called the median.
- The 75^{th} percentile is also called the third quartile, *Q3*.
- Between Q1 and Q3 is the middle 50% of the data. The range these data span is called the *interquartile range*, or the *IQR*.

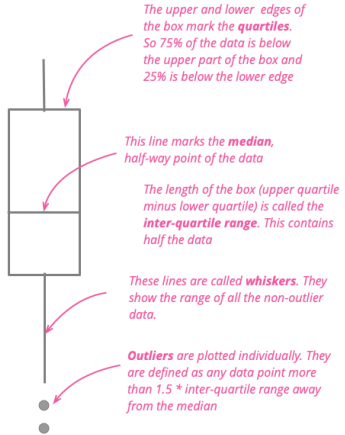
$$IQR = Q3 - Q1$$

Boxplot

The box in a *box plot* represents the middle 50% of the data, and the thick line in the box is the median.



Boxplot Explained



Note: As an aside, this is a nice blog post on comparing groups (I have added the nice boxplot explainer from this piece) that you should read- `https://martinfowler.com/articles/dont-compare-averages.html`

Whiskers and outliers

- *Whiskers* of a box plot can extend up to $1.5 \times IQR$ away from the quartiles.

$$\text{max upper whisker reach} = Q3 + 1.5 \times IQR$$
$$\text{max lower whisker reach} = Q1 - 1.5 \times IQR$$

Whiskers and outliers

- *Whiskers* of a box plot can extend up to $1.5 \times IQR$ away from the quartiles.

$$\text{max upper whisker reach} = Q3 + 1.5 \times IQR$$
$$\text{max lower whisker reach} = Q1 - 1.5 \times IQR$$

$$\text{IQR} : 191 - 167 = 24$$
$$\text{max upper whisker reach} = 191 + 1.5 \times 24 = 227$$
$$\text{max lower whisker reach} = 167 - 1.5 \times 24 = 131$$

Whiskers and outliers

- *Whiskers* of a box plot can extend up to $1.5 \times IQR$ away from the quartiles.

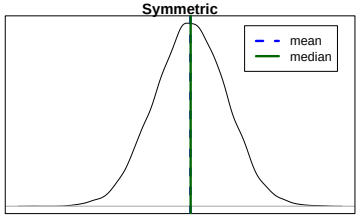
$$\text{max upper whisker reach} = Q3 + 1.5 \times IQR$$
$$\text{max lower whisker reach} = Q1 - 1.5 \times IQR$$

$$\text{IQR} : 191 - 167 = 24$$
$$\text{max upper whisker reach} = 191 + 1.5 \times 24 = 227$$
$$\text{max lower whisker reach} = 167 - 1.5 \times 24 = 131$$

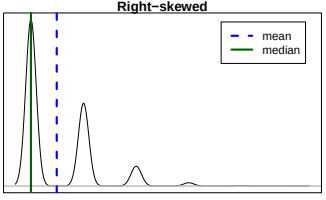
- A potential *outlier* is defined as an observation beyond the maximum reach of the whiskers. It is an observation that appears extreme relative to the rest of the data.

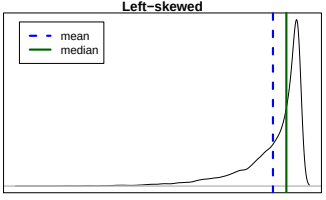
Mean vs. median

- If the distribution is symmetric, center is often defined as the mean: mean $\approx$ median



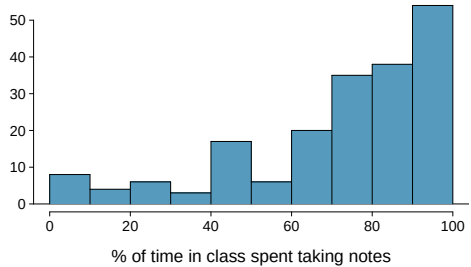
- If the distribution is skewed or has extreme outliers, center is often defined as the median
  - Right-skewed: mean $>$ median
  - Left-skewed: mean $<$ median





Practice

Which is most likely true for the distribution of percentage of time actually spent taking notes in class versus on Facebook, Twitter, etc.?



(a) mean $>$ median

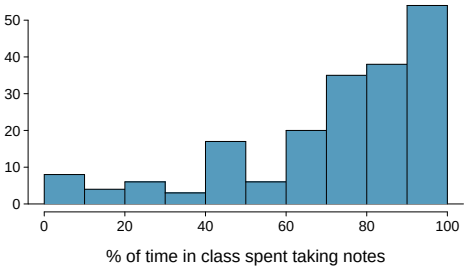
(b) mean $<$ median

(c) mean $\approx$ median

(d) impossible to tell

Practice

Which is most likely true for the distribution of percentage of time actually spent taking notes in class versus on Facebook, Twitter, etc.?



median: 80%
mean: 76%

(a) mean $>$ median

(b) mean $<$ median

(c) mean $\approx$ median

(d) impossible to tell

Extremely Skewed Distribution

When data are extremely skewed, transforming them might make modeling easier. A common transformation is the *log transformation*.

Extremely Skewed Distribution

When data are extremely skewed, transforming them might make modeling easier. A common transformation is the *log transformation*.

The histogram on the left shows the distribution of mass of Star Wars characters. The histogram on the right shows the distribution of log of mass.

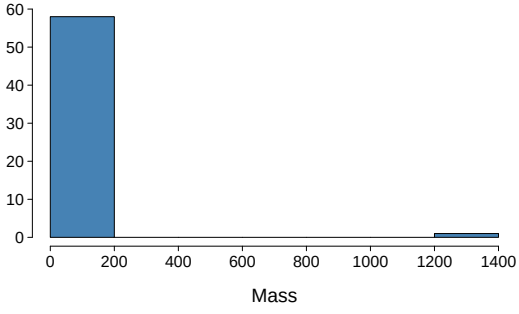


Figure 2: How it started

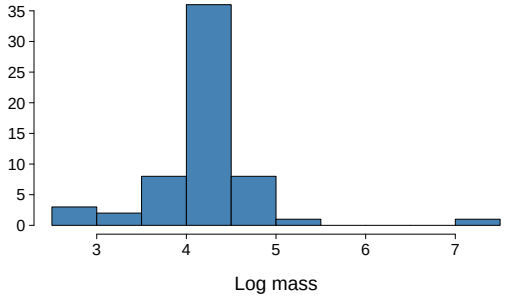


Figure 3: How it's going

Considering Categorical Data

Contingency Table

A table that summarizes data for two categorical variables is called a *contingency table*.

Contingency Table

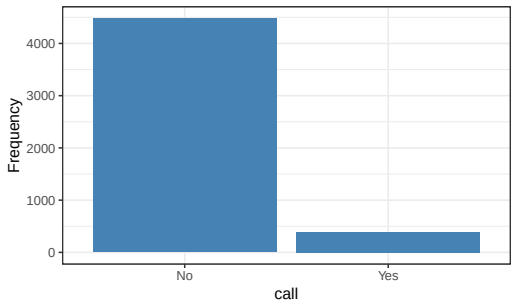
A table that summarizes data for two categorical variables is called a *contingency table*.

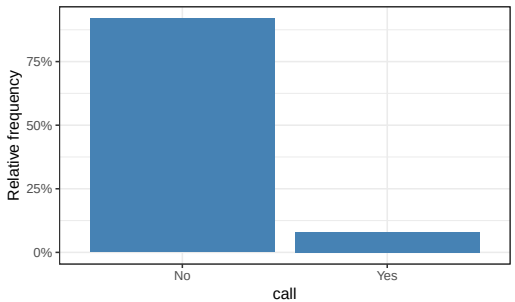
The contingency table below shows the distribution of callback rate and race of applicants.

		Callback		
		No	Yes	Total
Race	Black	2278	157	2435
	White	2200	235	2435
	Total	4478	392	4870

Bar plots

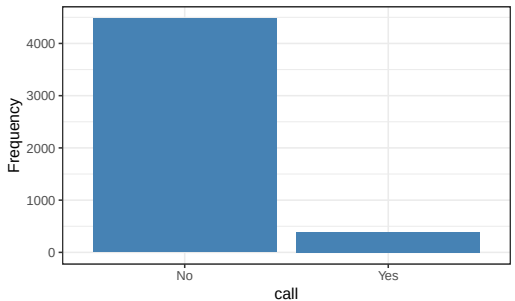
A *bar plot* is a common way to display a single categorical variable. A bar plot where proportions instead of frequencies are shown is called a *relative frequency bar plot*.

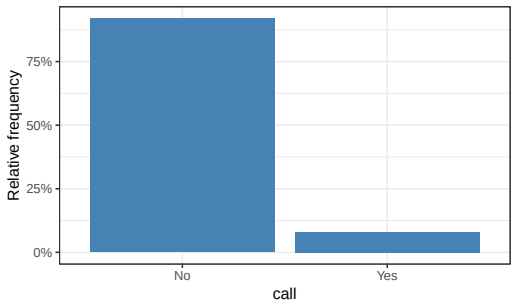




Bar plots

A *bar plot* is a common way to display a single categorical variable. A bar plot where proportions instead of frequencies are shown is called a *relative frequency bar plot*.

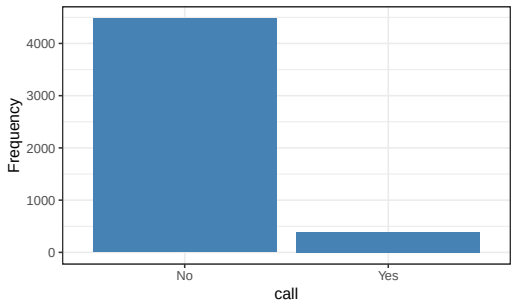


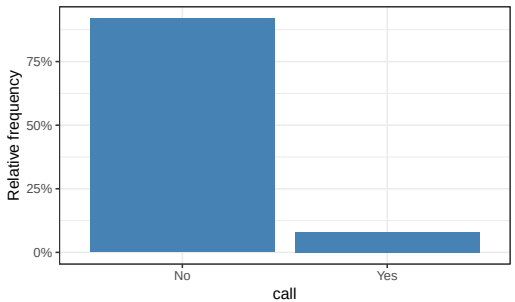


How are bar plots different than histograms?

Bar plots

A *bar plot* is a common way to display a single categorical variable. A bar plot where proportions instead of frequencies are shown is called a *relative frequency bar plot*.





How are bar plots different than histograms?

Bar plots are used for displaying distributions of categorical variables, histograms are used for numerical variables. The x-axis in a histogram is a number line, hence the order of the bars cannot be changed. In a bar plot, the categories can be listed in any order (though some orderings make more sense than others, especially for ordinal variables.)

Choosing the appropriate proportion

Does there appear to be a relationship between race and survival callback rate?

		Callback		
		No	Yes	Total
Race	Black	2278	157	2435
	White	2200	235	2435
	Total	4478	392	4870

Choosing the appropriate proportion

Does there appear to be a relationship between race and survival callback rate?

		Callback		
		No	Yes	Total
Race	Black	2278	157	2435
	White	2200	235	2435
	Total	4478	392	4870

To answer this question we examine the row proportions:

Choosing the appropriate proportion

Does there appear to be a relationship between race and survival callback rate?

		Callback		
		No	Yes	Total
Race	Black	2278	157	2435
	White	2200	235	2435
	Total	4478	392	4870

To answer this question we examine the row proportions:

- % black applicants who got callback: 157 / 2435 $\approx$ 6.4

Choosing the appropriate proportion

Does there appear to be a relationship between race and survival callback rate?

		Callback		
		No	Yes	Total
Race	Black	2278	157	2435
	White	2200	235	2435
	Total	4478	392	4870

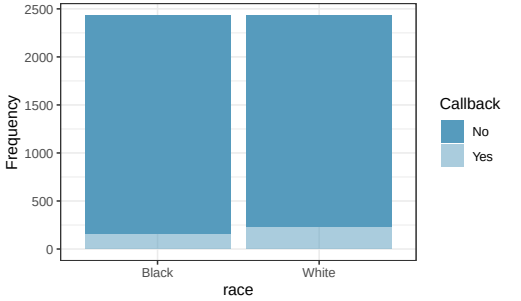
To answer this question we examine the row proportions:

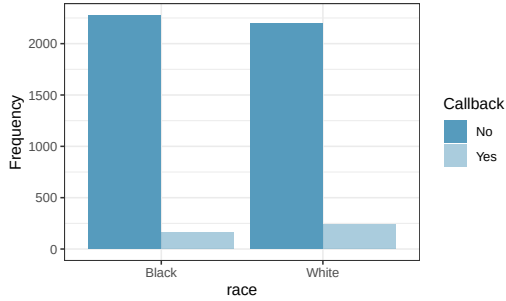
- % black applicants who got callback: 157 / 2435 $\approx$ 6.4
- % white applicants who got callback: 392 / 2435 $\approx$ 9.7

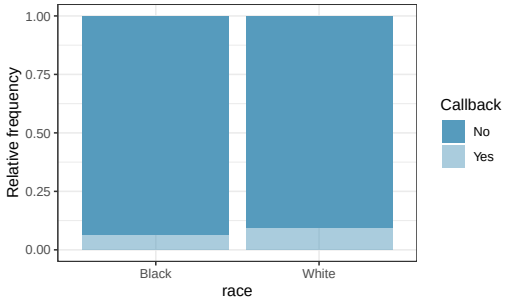
Bar plots with two variables

- *Stacked bar plot:* Graphical display of contingency table information, for counts.
- *Side-by-side bar plot:* Displays the same information by placing bars next to, instead of on top of, each other.
- *Standardized stacked bar plot*: Graphical display of contingency table information, for proportions.

What are the differences between the three visualizations shown below?

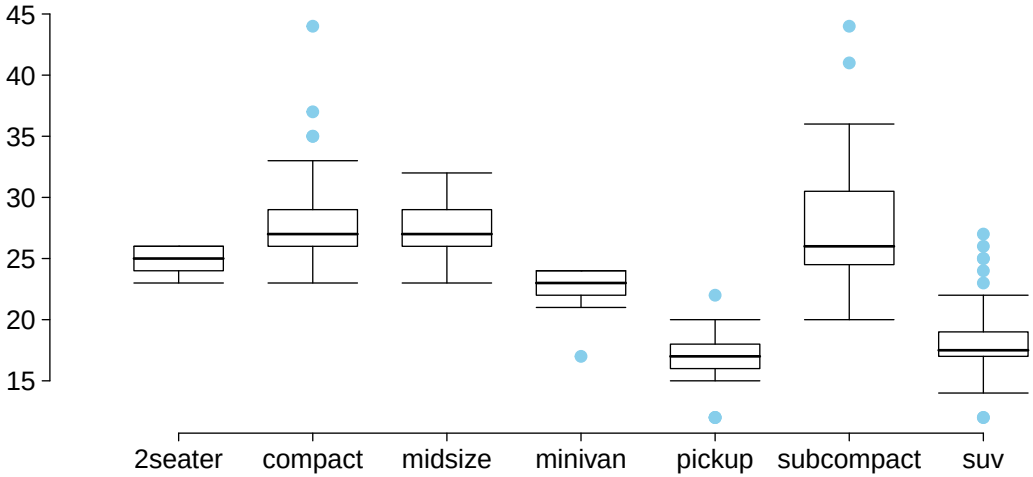






Side-by-side box plots

Does there appear to be a relationship between type of vehicle and miles per gallon?



Case Study

Racial discrimination

As we saw earlier, there is difference of about 3% in callback between CVs with white names and those with African-American names.

Practice

We saw a difference of almost 3% between the proportion of African-American and White CVs getting callbacks. Based on what we know, there are many things to consider.

(a) If we were to repeat the experiment we will definitely see that more African-American names get callback. This was a fluke.

(b) Callback is dependent on race, White-sounding names are more likely to get callback, and hence there is racial discrimination against African American job-seekers.

(c) The difference in the proportions of callback is due to chance, this is not evidence of racial discrimination against African-Americans in labour market.

(d) African-American are less qualified than White job-applicants, and this is why fewer African-American get callback.

Two competing claims

1. “There is nothing going on.”
   Callback and race are *independent*, no racial discrimination, observed difference in proportions is simply due to chance. $\rightarrow$ *Null hypothesis*

Two competing claims

1. “There is nothing going on.”
   Callback and race are *independent*, no racial discrimination, observed difference in proportions is simply due to chance. $\rightarrow$ *Null hypothesis*
2. “There is something going on.”
   Callback and race are *dependent*, there is racial discrimination, observed difference in proportions is not due to chance. $\rightarrow$ *Alternative hypothesis*

Recap: hypothesis testing framework

- We start with a *null hypothesis (H_0)* that represents the status quo.
- We also have an *alternative hypothesis (H_A)* that represents our research question, i.e. what we're testing for.
- We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation (today) or theoretical methods (later in the course).
- If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis. If they do, then we reject the null hypothesis in favor of the alternative.

Simulating the experiment...

... under the assumption of independence, i.e. leave things up to chance.

If results from the simulations based on the *chance model* look like the data, then we can determine that the difference between the proportions of promoted files between White and African-American job-applicants was simply *due to chance* (callback and race are independent).

If the results from the simulations based on the chance model do not look like the data, then we can determine that the difference between the proportions of promoted files between White and African-American job-applicants was not due to chance, but *due to an actual effect of race* (callback and race are dependent).

Simulation Result