Elementary Statistics – Exploratory Data Analysis (EDA)

Dr. Ab Mosca (they/them)

Slides based off slides courtesy of Kaitlyn Cook (https://www.smith.edu/people/kaitlyn-cook)

Plan for Today

For a single variable:

- Descriptive statistics
- Summary visualizations

Warm Up

Study Designs

- ***Observational***: Researchers observe both the explanatory and response variables without interfering in how the data arises
- ***Experiments***: Researchers intervene and assign treatments (the explanatory variable) to each participant in the study (ideally randomly)

Variables

- ***Response variable:*** the measured outcome of interest.
- ***Explanatory variable:*** a variable that potentials explains or predicts changes in the response.
- ***Confounds*** are associated with the explanatory and response variables, and ***obscure true relationships***.

A researcher observes that towns with higher numbers of doctors also report higher numbers of crimes. They conclude that doctors must commit crimes at higher rate than the general population does.

a) What type of study is this?
b) As stated by the researcher, what are the response and explanatory variables?
c) Is there a confound? If so, what?

Warm Up

Study Designs

- ***Observational***: Researchers observe both the explanatory and response variables without interfering in how the data arises
- ***Experiments***: Researchers intervene and assign treatments (the explanatory variable) to each participant in the study (ideally randomly)

Variables

- ***Response variable:*** the measured outcome of interest.
- ***Explanatory variable:*** a variable that potentials explains or predicts changes in the response.
- ***Confounds*** are associated with the explanatory and response variables, and ***obscure true relationships***.

A researcher believes that light levels might effect how students perform on exams. They randomly assign students to a treatment: fluorescent overhead lighting, yellow overhead lighting, and no overhead lighting (only desk lamps), and have the students take an exam in that lighting. Note: light levels might have different effects on people who wear glasses and people who don't. Some students wear glasses and others do not.

a) What type of study is this?
b) As stated by the researcher, what are the response and explanatory variables?
c) Is there a confound? If so, what?

Big Picture

Thus far we've focused on understanding where our data come from:

Do those data represent a random sample from our target population? Was the explanatory variable randomly allocated?

This in turn determines *what* sorts of conclusions we can draw and to *whom* we can generalize those results:

		Assignment of Explanatory Variable		
		Random allocation of explanatory variable	Individual decides explanatory variable (non-random)	
Selection of Observational Units from the Population	Random sample	The observational units are randomly selected from the population; then the explanatory variable (treatment) is randomly assigned.	The observational units are randomly selected from the population, but the value of the explanatory variable is not randomly assigned by the researcher.	Conclusions generalize directly to the population.
	Other sampling method (non-random)	The observational units are observed (somehow!) and then randomly allocated to the levels of the explanatory variable.	The observational units are observed (somehow!) and the value of the explanatory variable is not randomly assigned by the researcher.	Conclusions might not be generalizable because of volunteer bias.
		Significant conclusions are considered to be cause and effect.	Significant conclusions must be framed with possible confounding variables.	

Big Picture

Thus far we've focused on understanding where our data come from:

Do those data represent a random sample from our target population? Was the explanatory variable randomly allocated?

This in turn determines *what* sorts of conclusions we can draw and to *whom* we can generalize those results:

		Assignment of Explanatory Variable		
		Random allocation of explanatory variable	Individual decides explanatory variable (non-random)	
Selection of Observational Units from the Population	Random sample	The observational units are randomly selected from the population; then the explanatory variable (treatment) is randomly assigned.	The observational units are randomly selected from the population, but the value of the explanatory variable is not randomly assigned by the researcher.	Conclusions generalize directly to the population.
	Other sampling method (non-random)	The observational units are observed (somehow!) and then randomly allocated to the levels of the explanatory variable.	The observational units are observed (somehow!) and the value of the explanatory variable is not randomly assigned by the researcher.	Conclusions might not be generalizable because of volunteer bias.
		Significant conclusions are considered to be cause and effect.	Significant conclusions must be framed with possible confounding variables.	

Big Picture

Thus far we've focused on understanding where our data come from:

Do those data represent a random sample from our target population? Was the explanatory variable randomly allocated?

This in turn determines *what* sorts of conclusions we can draw and to *whom* we can generalize those results:

		Assignment of Explanatory Variable		
		Random allocation of explanatory variable	Individual decides explanatory variable (non-random)	
Selection of Observational Units from the Population	Random sample	The observational units are randomly selected from the population; then the explanatory variable (treatment) is randomly assigned.	The observational units are randomly selected from the population, but the value of the explanatory variable is not randomly assigned by the researcher.	Conclusions generalize directly to the population.
	Other sampling method (non-random)	The observational units are observed (somehow!) and then randomly allocated to the levels of the explanatory variable.	The observational units are observed (somehow!) and the value of the explanatory variable is not randomly assigned by the researcher.	Conclusions might not be generalizable because of volunteer bias.
		Significant conclusions are considered to be cause and effect.	Significant conclusions must be framed with possible confounding variables.	

Big Picture

Thus far we've focused on understanding where our data come from:

> Do those data represent a random sample from our target population? Was the explanatory variable randomly allocated?

This in turn determines *what* sorts of conclusions we can draw and to *whom* we can generalize those results:

		Assignment of Explanatory Variable		
		Random allocation of explanatory variable	Individual decides explanatory variable (non-random)	
Selection of Observational Units from the Population	Random sample	The observational units are randomly selected from the population; then the explanatory variable (treatment) is randomly assigned.	The observational units are randomly selected from the population, but the value of the explanatory variable is not randomly assigned by the researcher.	Conclusions generalize directly to the population.
	Other sampling method (non-random)	The observational units are observed (somehow!) and then randomly allocated to the levels of the explanatory variable.	The observational units are observed (somehow!) and the value of the explanatory variable is not randomly assigned by the researcher.	Conclusions might not be generalizable because of volunteer bias.
		Significant conclusions are considered to be cause and effect.	Significant conclusions must be framed with possible confounding variables.	

Big Picture

Thus far we've focused on understanding where our data come from:

Do those data represent a random sample from our target population? Was the explanatory variable randomly allocated?

This in turn determines *what* sorts of conclusions we can draw and to *whom* we can generalize those results:

		Assignment of Explanatory Variable		
		Random allocation of explanatory variable	Individual decides explanatory variable (non-random)	
Selection of Observational Units from the Population	Random sample	The observational units are randomly selected from the population; then the explanatory variable (treatment) is randomly assigned.	The observational units are randomly selected from the population, but the value of the explanatory variable is not randomly assigned by the researcher.	Conclusions generalize directly to the population.
	Other sampling method (non-random)	The observational units are observed (somehow!) and then randomly allocated to the levels of the explanatory variable.	The observational units are observed (somehow!) and the value of the explanatory variable is not randomly assigned by the researcher.	Conclusions might not be generalizable because of volunteer bias.
		Significant conclusions are considered to be cause and effect.	Significant conclusions must be framed with possible confounding variables.	

Big Picture

Thus far we've focused on understanding where our data come from:

Do those data represent a random sample from our target population? Was the explanatory variable randomly allocated?

This in turn determines *what* sorts of conclusions we can draw and to *whom* we can generalize those results:

		Assignment of Explanatory Variable			
		Random allocation of explanatory variable	Individual decides explanatory variable (non-random)		
Selection of Obser[illegible] Units [illegible] the Population	Random sample	The observational units are randomly selected from the population; then the explanatory [illegible]	The observational units are randomly selected from the population, but the value of the explanatory variable is not [illegible]	→	Conclusions generalize directly to [illegible]
	[illegible] sampling method (non-random)	The observational units are observed (somehow!) and then randomly allocated to the levels of the explanatory variable.	[illegible] observed (somehow!) and the value of the explanatory variable is not randomly assigned by the researcher.	→	Conclusions might not be generalizable because of volunteer bias.
		↓	↓		
		Significant conclusions are considered to be cause and effect.	Significant conclusions must be framed with possible confounding variables.		

But once we have our data. . . then what?

IMDB Movie Dataset

For the next few classes, we'll be working with data on movies in order to understand how the attributes of these movies associate with box office gross!

Data on 3,039 movies made in the US between 1929 and 2016, scraped from IMDB

	movie_title	title_year	budget	log_budget	gross	log_gross	country	language
1	Avatar	2009	237000000	19.28357	760505847	20.44949	USA	English
2	Pirates of the Caribbean: At World's End	2007	300000000	19.51929	309404152	19.55016	USA	English
4	The Dark Knight Rises	2012	250000000	19.33697	448130642	19.92060	USA	English
6	John Carter	2012	263700000	19.39032	73058679	18.10677	USA	English
7	Spider-Man 3	2007	258000000	19.36847	336530303	19.63420	USA	English
8	Tangled	2010	260000000	19.37619	200807262	19.11786	USA	English
9	Avengers: Age of Ultron	2015	250000000	19.33697	458991599	19.94454	USA	English
11	Batman v Superman: Dawn of Justice	2016	250000000	19.33697	330249062	19.61536	USA	English
12	Superman Returns	2006	209000000	19.15784	200069408	19.11417	USA	English
14	Pirates of the Caribbean: Dead Man's Chest	2006	225000000	19.23161	423032628	19.86296	USA	English
15	The Lone Ranger	2013	215000000	19.18615	89289910	18.30740	USA	English
16	Man of Steel	2013	225000000	19.23161	291021565	19.48891	USA	English
17	The Chronicles of Narnia: Prince Caspian	2008	225000000	19.23161	141614023	18.76862	USA	English
18	The Avengers	2012	220000000	19.20914	623279547	20.25051	USA	English

Data are from Kaggle.com and are available at https://github.com/kaitlyncook/data-sets

Exploratory Data Analysis

Exploratory data analysis (EDA) refers to the practice of reducing and summarizing data in ways that:

- Help us make sense of the information that we have
- Help to inform our understanding of our research question

You can think of EDA as the data version of tl;dr.

Exploratory Data Analysis

Exploratory data analysis (EDA) refers to the practice of reducing and summarizing data in ways that:

- Help us make sense of the information that we have
- Help to inform our understanding of our research question

You can think of EDA as the data version of tl;dr.

Graphical Summaries (Data Visualizations)

A visual representation of how our data are *distributed* across the observations in our sample

Numeric Summaries (Summary Statistics)

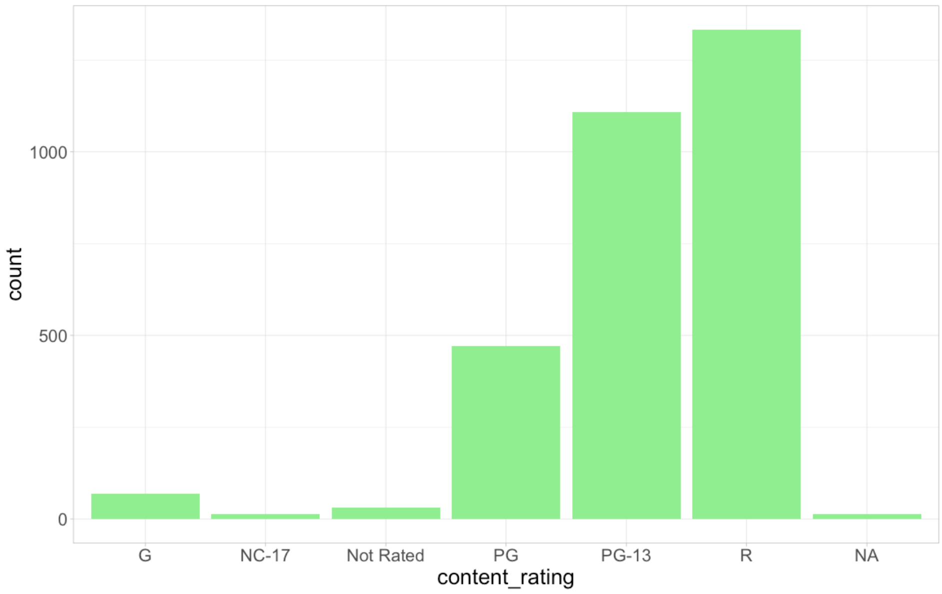
A single number or set of numbers that captures important features of that distribution, such as its *center* and *spread*

EDA for Categorical Variables: **Bar Plots**

The empirical distribution of a categorical variable is comprised of:

- The possible levels or values of the categorical variable
- The (relative) frequency of those levels in the observed data

One method of visualizing this distribution is through a bar plot:



Bar plot showing frequency of MPAA ratings.

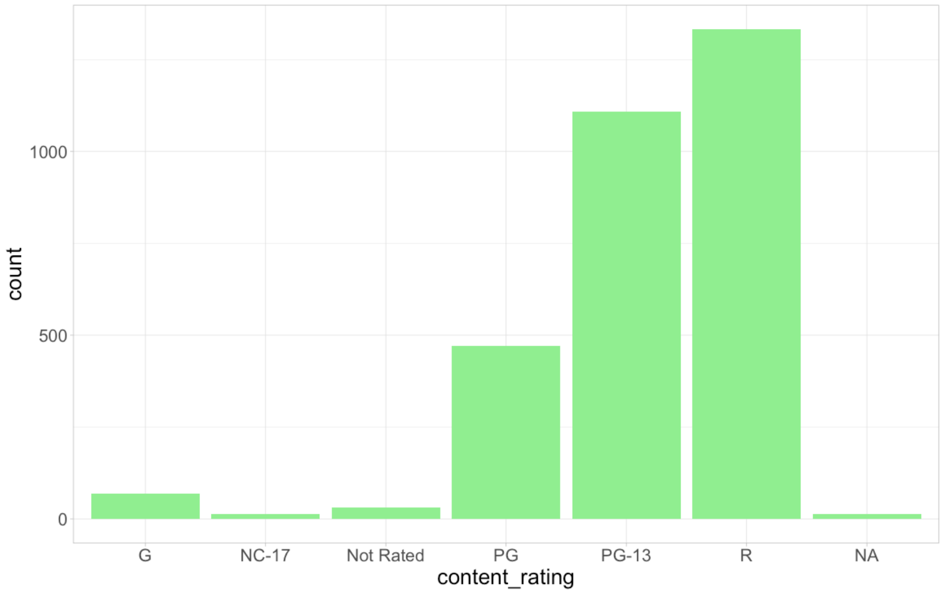
Bar plot showing relative frequency of MPAA ratings.

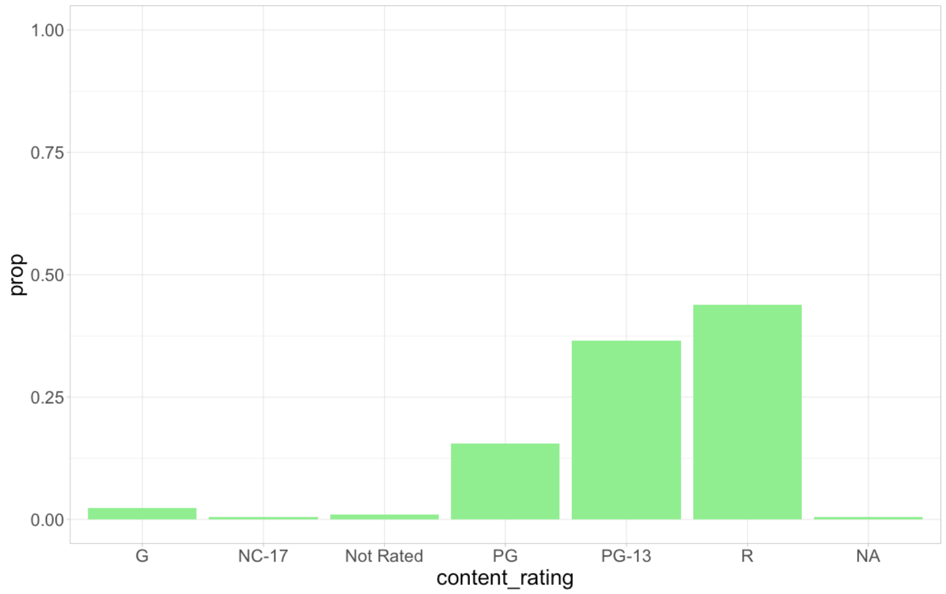
EDA for Categorical Variables: **Bar Plots**

The empirical distribution of a categorical variable is comprised of:

- The possible levels or values of the categorical variable
- The (relative) frequency of those levels in the observed data

One method of visualizing this distribution is through a bar plot:





Example

Work with 2-3 other people to visualize the distribution of years (first year, sophomore, junior, senior) in this class.

EDA for Categorical Variables: **Summary Statistics**

We can present this same information numerically using a frequency table, which displays both:

- the number of movies (n) that obtained each rating
- the relative frequency of (prop) those ratings

content_rating	n	prop
Not Rated	21	0.0069767
G	66	0.0219269
PG	471	0.1564784
PG-13	1108	0.3681063
R	1331	0.4421927
NC-17	13	0.0043189

EDA for Categorical Variables: **Summary Statistics**

We can present this same information numerically using a frequency table, which displays both:

- the number of movies (n) that obtained each rating
- the relative frequency of (prop) those ratings

content_rating	n	prop
Not Rated	21	0.0069767
G	66	0.0219269
PG	471	0.1564784
PG-13	1108	0.3681063
R	1331	0.4421927

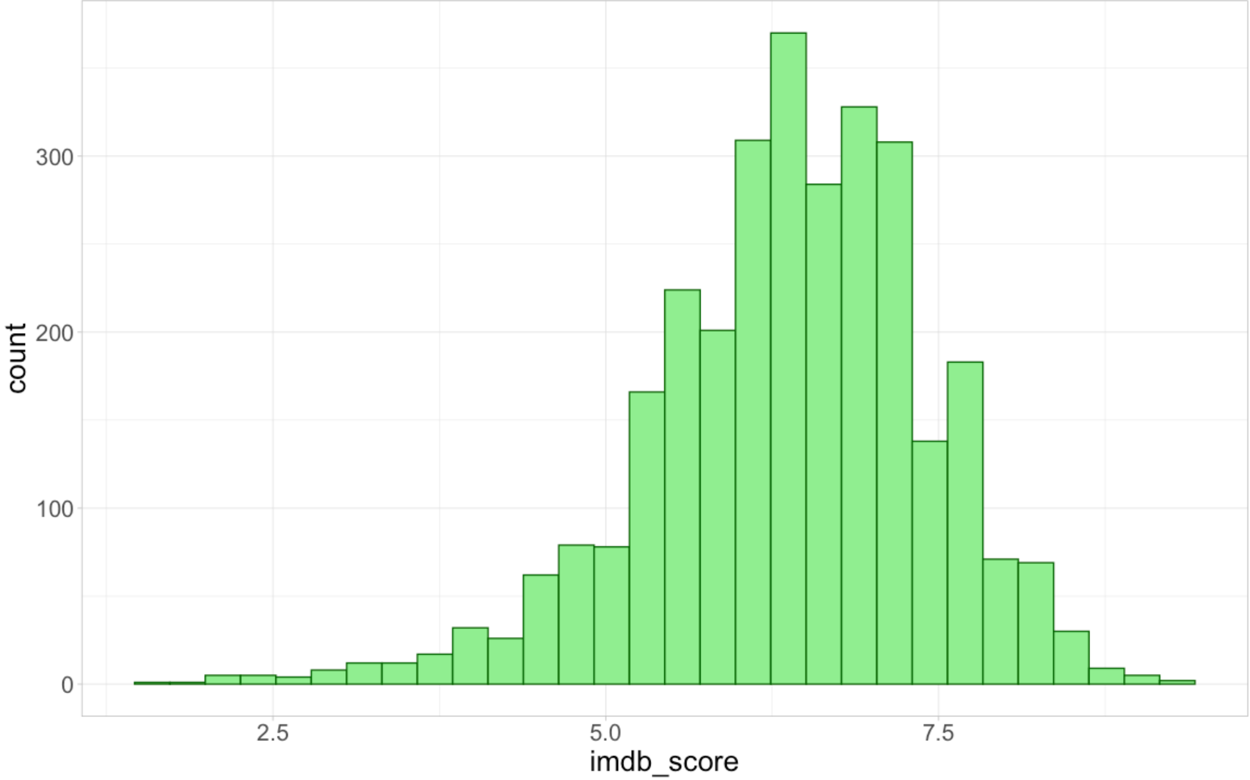
Example

Work with 2-3 other people to represent the distribution of years (first year, sophomore, junior, senior) in this class with a frequency table.

EDA for Numerical Variables: **Histograms**

When the variable that we're summarizing is numerical, we can instead visualize its distribution using either a histogram or density plot

Histogram: numerical analog of the frequency bar plot



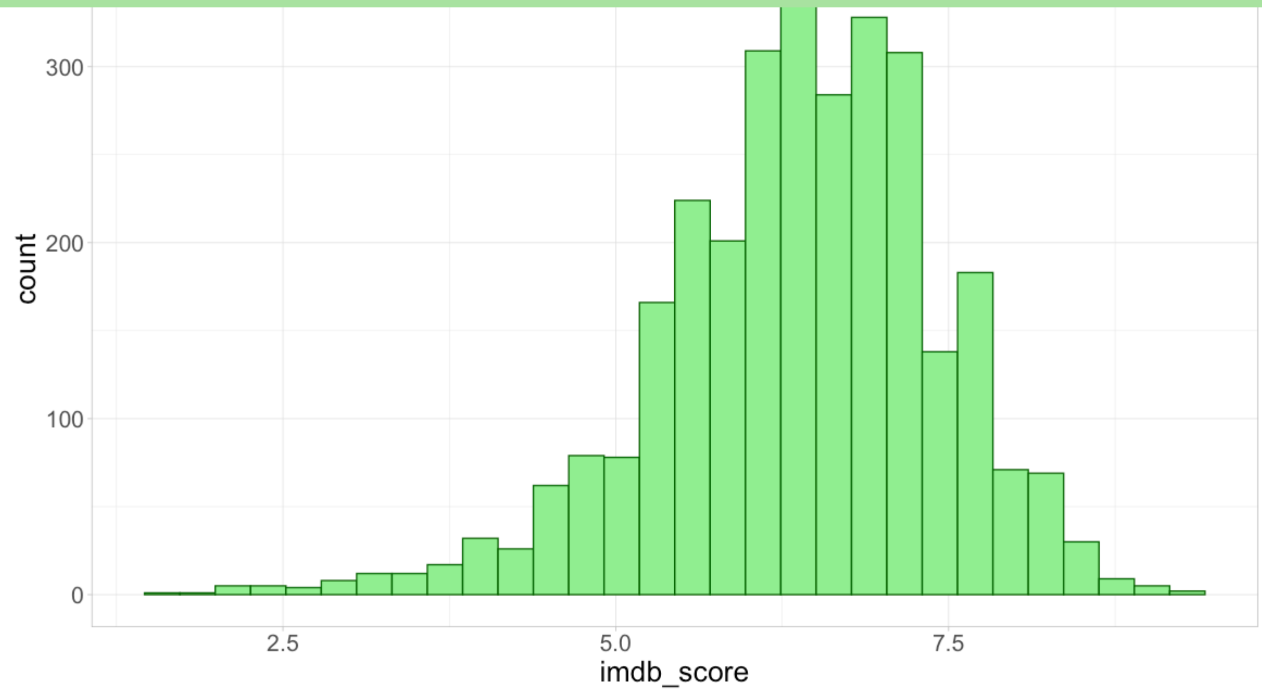
Created by:

- Dividing the range of IMDB ratings (here from 1.6 to 9.3) into intervals (also called "bins") of equal width
- Counting the number of movies whose IMDB rating falls into each bin

EDA for Numerical Variables: **Histograms**

Example

Work with 2-3 other people to visualize the distribution of ages in this class with a histogram.



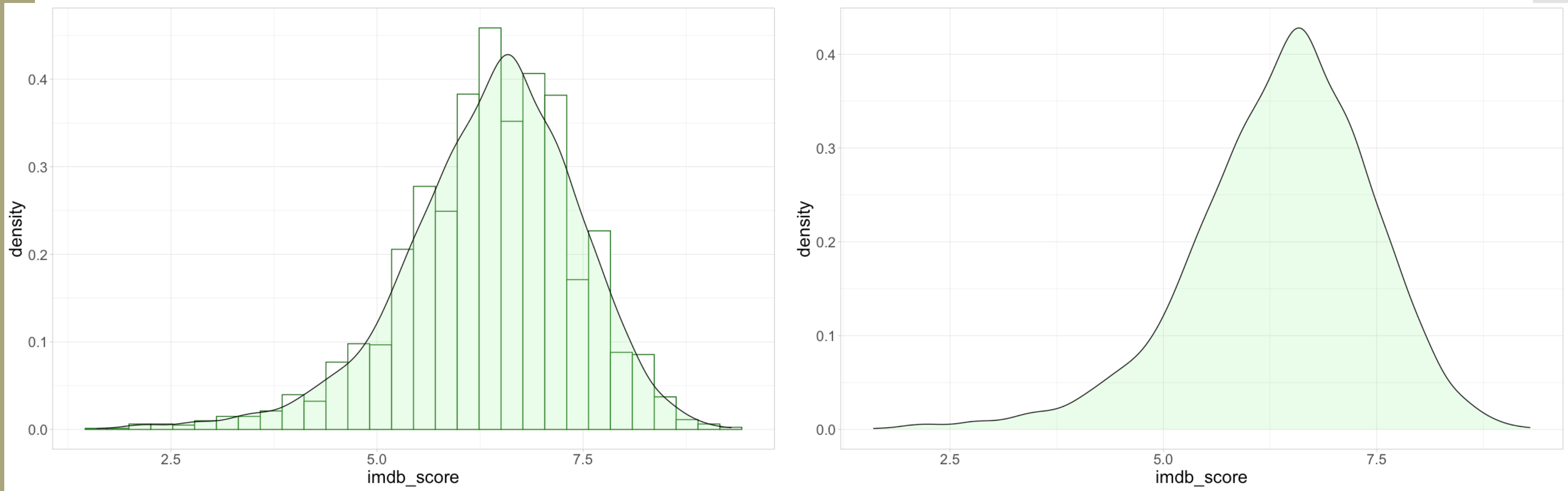
Created by:

- Dividing the range of IMDB ratings (here from 1.6 to 9.3) into intervals (also called “bins”) of equal width
- Counting the number of movies whose IMDB rating falls into each bin

EDA for Numerical Variables: **Density Plots**

When the variable that we're summarizing is numerical, we can instead visualize its distribution using either a histogram or density plot

Density plot: numerical analog of relative frequency bar plot

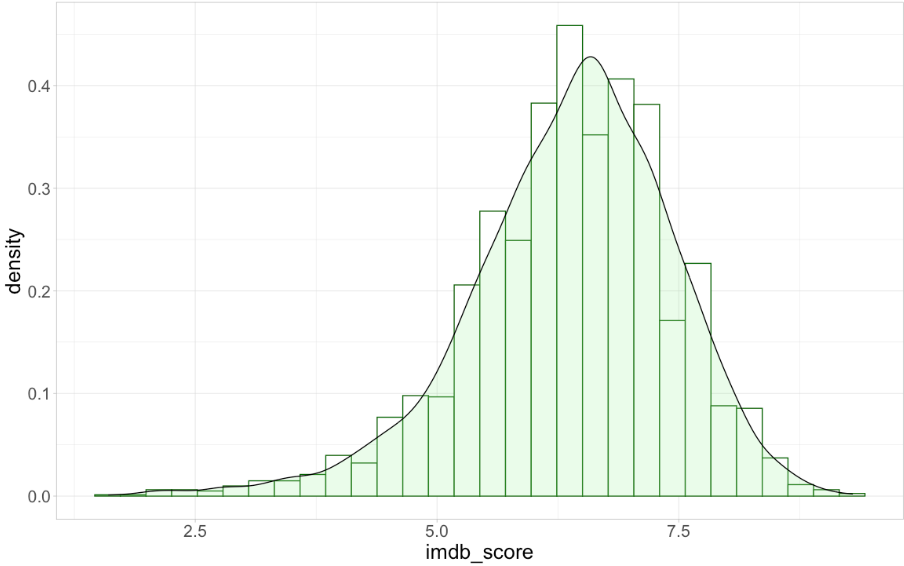


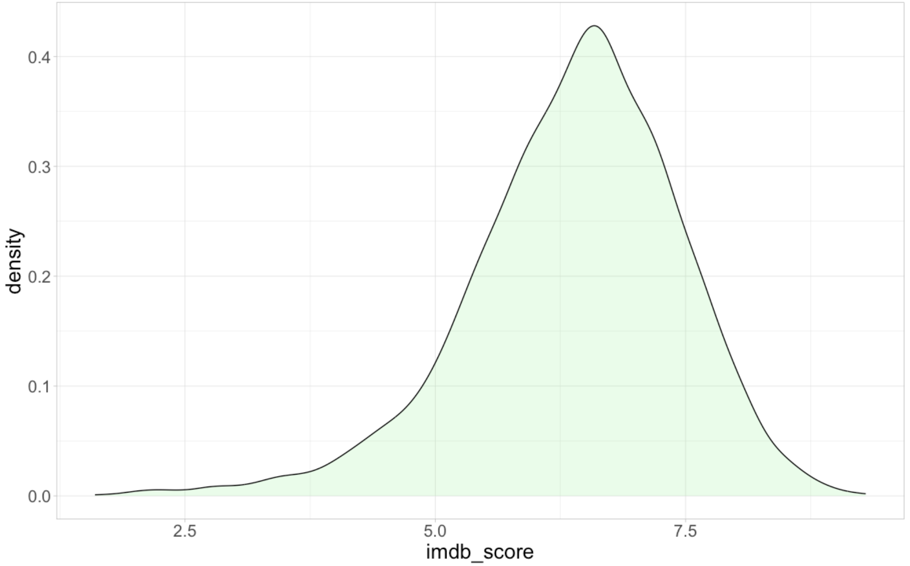
Created by standardizing and smoothing over the corresponding histogram

EDA for Numerical Variables: **Density Plots**

Example

Work with 2-3 other people to visualize the distribution of ages in this class with a density plot.





Created by standardizing and smoothing over the corresponding histogram

Interlude: Describing Distributions

When looking at a variable's distribution, we want to pay attention to and describe the following attributes:

Skewness	Center
Modality	Spread

Interlude: Describing Distributions

When looking at a variable's distribution, we want to pay attention to and describe the following attributes:

- Graphical
  - Skewness
  - Modality
- Summary Statistics
  - Center
  - Spread

Interlude: Describing Distributions

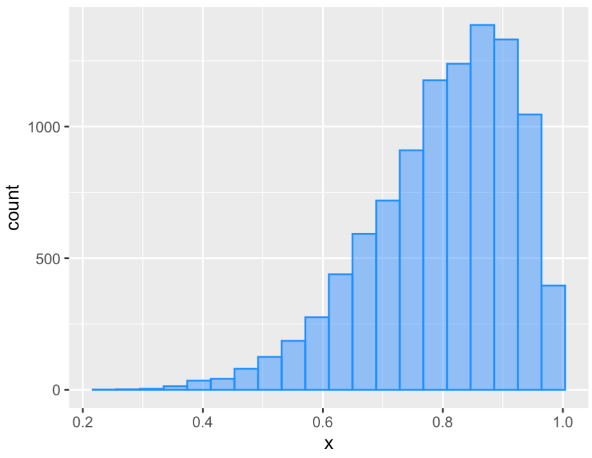
When looking at a variable's distribution, we want to pay attention to and describe the following attributes:

Graphical
- Skewness
- Modality

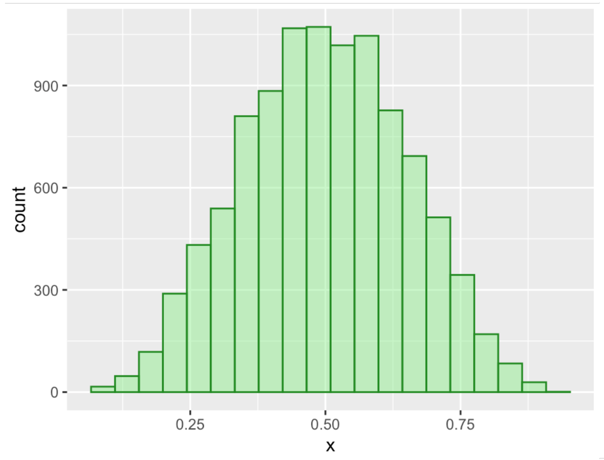
Summary Statistics
- Center
- Spread

Skewness is a measure of (a)symmetry!

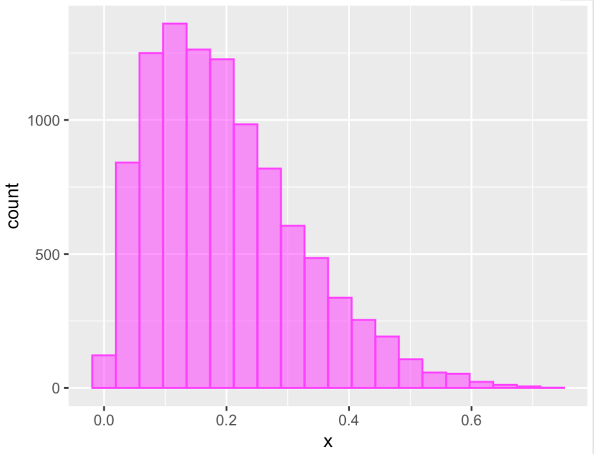
→ Why pay attention to skew? Later in this course we'll see statistical tools that assume our data are (close to) symmetric, and we need to be able to assess whether this assumption is reasonable.



Left ("negative") skewed distribution.



Symmetric distribution.



Right ("positive") skewed distribution.

Tip: Whatever side the long tail is on is the side of skew

When looking at a variable’s distribution, we want to pay attention to and describe the following attributes:

Graphical: **Skewness**, Modality

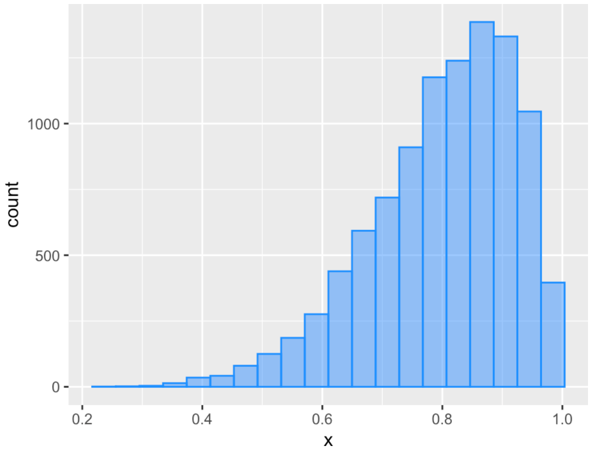
Summary Statistics: Center, Spread

Interlude: Describing Distributions

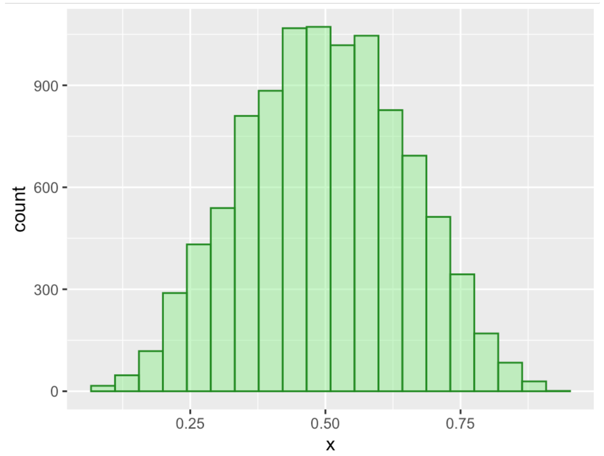
Example

What is the skewness of our age data?

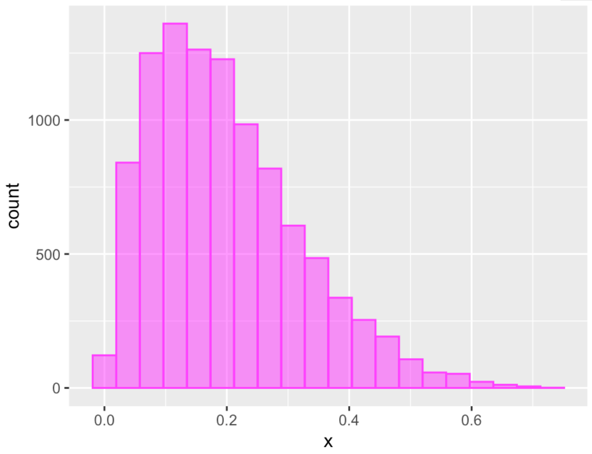
be able to assess whether this assumption is reasonable.



Left (“negative”) skewed distribution.



Symmetric distribution.



Right (“positive”) skewed distribution.

Tip: Whatever side the long tail is on is the side of skew

Interlude: Describing Distributions

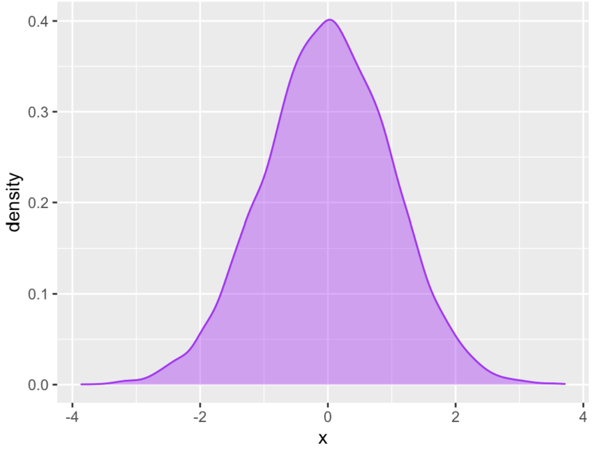
When looking at a variable's distribution, we want to pay attention to and describe the following attributes:

Graphical: Skewness, **Modality**

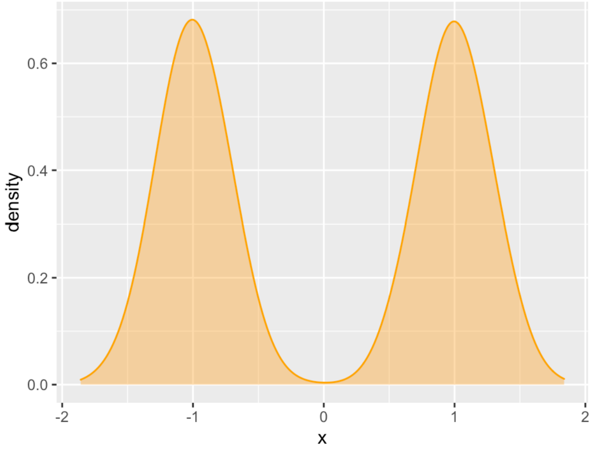
Summary Statistics: Center, Spread

Modality is a measure of how many peaks ("modes") the distribution has

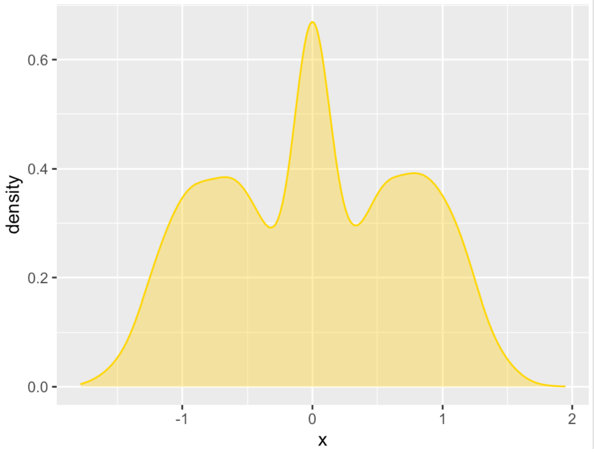
→ Why pay attention to modality? A mode is a value that occurs with high frequency in our data, and it can help to inform our understanding of what values our variable tends to take on



Unimodal distribution.



Bimodal distribution.



Multimodal distribution

Interlude: Describing Distributions

When looking at a variable's distribution, we want to pay attention to and describe the following attributes:

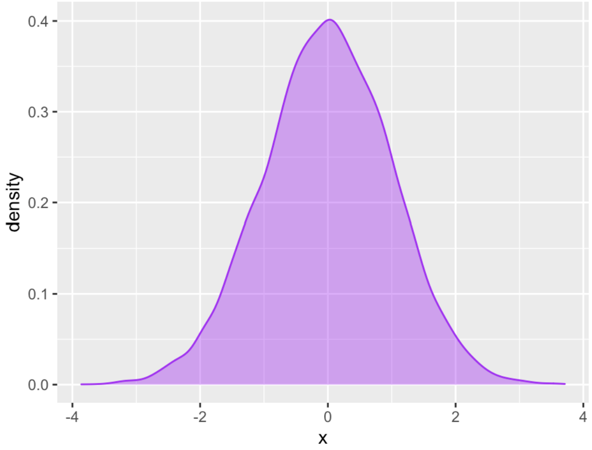
Graphical
- Skewness
- **Modality**

Summary Statistics
- Center
- Spread

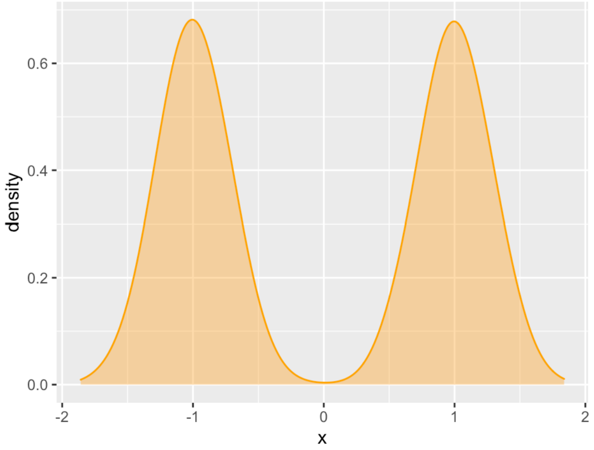
Example

What is the modality of our age data?

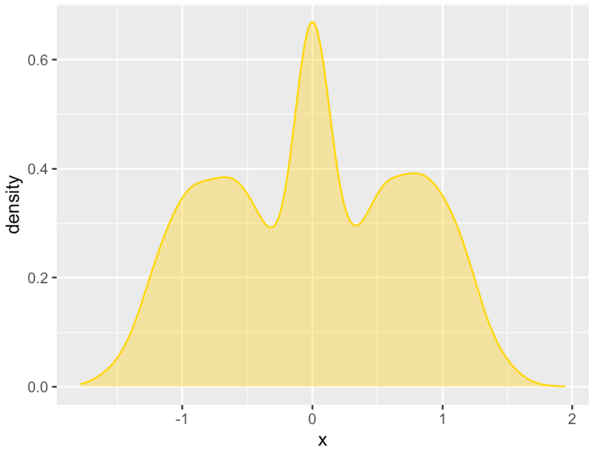
high frequency in our data, and it can help to inform our understanding of what values our variable tends to take on



Unimodal distribution.



Bimodal distribution.



Multimodal distribution

Interlude: Describing Distributions

When looking at a variable's distribution, we want to pay attention to and describe the following attributes:

- Graphical
  - Skewness
  - Modality
- Summary Statistics
  - **Center**
  - Spread

Let $x_1, x_2, x_3, \ldots x_n$ be the observed values of our variable of interest across the n observational units in our dataset.

Measures of central tendency give us a sense of what the typical value of this variable might look like.

Mean: the average value of the variable,

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ \ldots. + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Median: suppose we order the observations from smallest to largest. the median is the value of x_i that falls in the middle (or, if n is even, the average of the two middle values).

$\Rightarrow$ At least half of our data are less than or equal to the median and at least half are greater than or equal to the median

Interlude: Describing Distributions

When looking at a variable's distribution, we want to pay attention to and describe the following attributes:

Graphical
- Skewness
- Modality

Summary Statistics
- **Center**
- Spread

Let $x_1, x_2, x_3, \ldots x_n$ be the observed values of our variable of interest across the n observational units in our dataset.

Mean: the average value of the variable,

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \;\ldots. + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Example

Work with 2-3 other people to find the mean age of students in this class.

Interlude: Describing Distributions

When looking at a variable's distribution, we want to pay attention to and describe the following attributes:

Graphical
- Skewness
- Modality

Summary Statistics
- **Center**
- Spread

Let $x_1, x_2, x_3, \ldots x_n$ be the observed values of our variable of interest across the n observational units in our dataset.

Median: suppose we order the observations from smallest to largest. the median is the value of x_i that falls in the middle (or, if n is even, the average of the two middle values).

Example

Work with 2-3 other people to find the median age of students in this class.

Interlude: Describing Distributions

When looking at a variable's distribution, we want to pay attention to and describe the following attributes:

Graphical
- Skewness
- Modality

Summary Statistics
- Center
- **Spread**

Measures of dispersion give us a sense of how much observation to observation variability there is in a variable.

Interlude: Describing Distributions

When looking at a variable's distribution, we want to pay attention to and describe the following attributes:

Graphical
- Skewness
- Modality

Summary Statistics
- Center
- **Spread**

Measures of dispersion give us a sense of how much observation to observation variability there is in a variable.

Range: the difference between the maximum and minimum values in the dataset

Interquartile Range: the difference between the 75th and 25th percentiles of the data

Variance: (almost) the average squared distance between the observed data for the ith observational unit, x_i, and the sample mean, $\bar{x}$

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n - 1} = \frac{1}{n - 1} \sum_{i-1}^{n} (x_i - \bar{x})^2$$

Standard Deviation: the square root of the variance, $s = \sqrt{s^2}$

Interlude: Describing Distributions

When looking at a variable's distribution, we want to pay attention to and describe the following attributes:

Graphical
- Skewness
- Modality

Summary Statistics
- Center
- **Spread**

Measures of dispersion give us a sense of how much observation to observation variability there is in a variable.

Range: the difference between the maximum and minimum values in the dataset

Example

Work with 2-3 other people to find the range of age.

Interlude: Describing Distributions

When looking at a variable's distribution, we want to pay attention to and describe the following attributes:

- Graphical
  - Skewness
  - Modality
- Summary Statistics
  - Center
  - **Spread**

Measures of dispersion give us a sense of how much observation to observation variability there is in a variable.

Interquartile Range: the difference between the 75th and 25th percentiles of the data

Example

Work with 2-3 other people to find the interquartile range of age.

Hint: The 25th percentile is the median of the lower half of your data and the 75th percentile is the median of the upper half of your data.

Interlude: Describing Distributions

When looking at a variable's distribution, we want to pay attention to and describe the following attributes:

- Graphical
  - Skewness
  - Modality
- Summary Statistics
  - Center
  - **Spread**

Measures of dispersion give us a sense of how much observation to observation variability there is in a variable.

Variance: (almost) the average squared distance between the observed data for the ith observational unit, x_i, and the sample mean, $\bar{x}$

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n - 1} = \frac{1}{n - 1}\sum_{i - 1}^{n}(x_i - \bar{x})^2$$

Example

Work with 2-3 other people to find variance of age.

Interlude: Describing Distributions

When looking at a variable's distribution, we want to pay attention to and describe the following attributes:

Graphical
- Skewness
- Modality

Summary Statistics
- Center
- **Spread**

Measures of dispersion give us a sense of how much observation to observation variability there is in a variable.

Standard Deviation: the square root of the variance, $s = \sqrt{s^2}$

Example

Work with 2-3 other people to find standard deviation of age.

Interlude: Describing Distributions

When looking at a variable's distribution, we want to pay attention to and describe the following attributes:

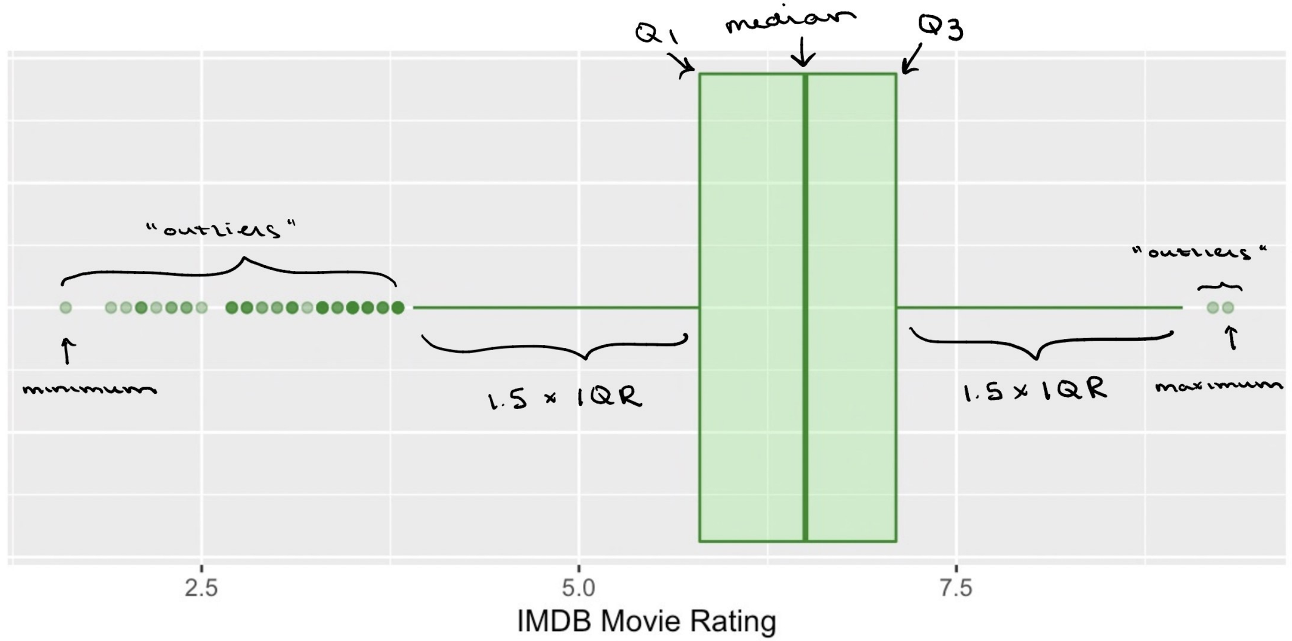
Graphical: Skewness, Modality

Summary Statistics: Center, **Spread**

The following five statistics make up the five-number summary, which captures information about both the center *and* spread of the data:

Minimum | 25th percentile | Median | 75th percentile | Maximum

We can use a box plot to visualize all of these statistics in one go:



Interlude: Describing Distributions

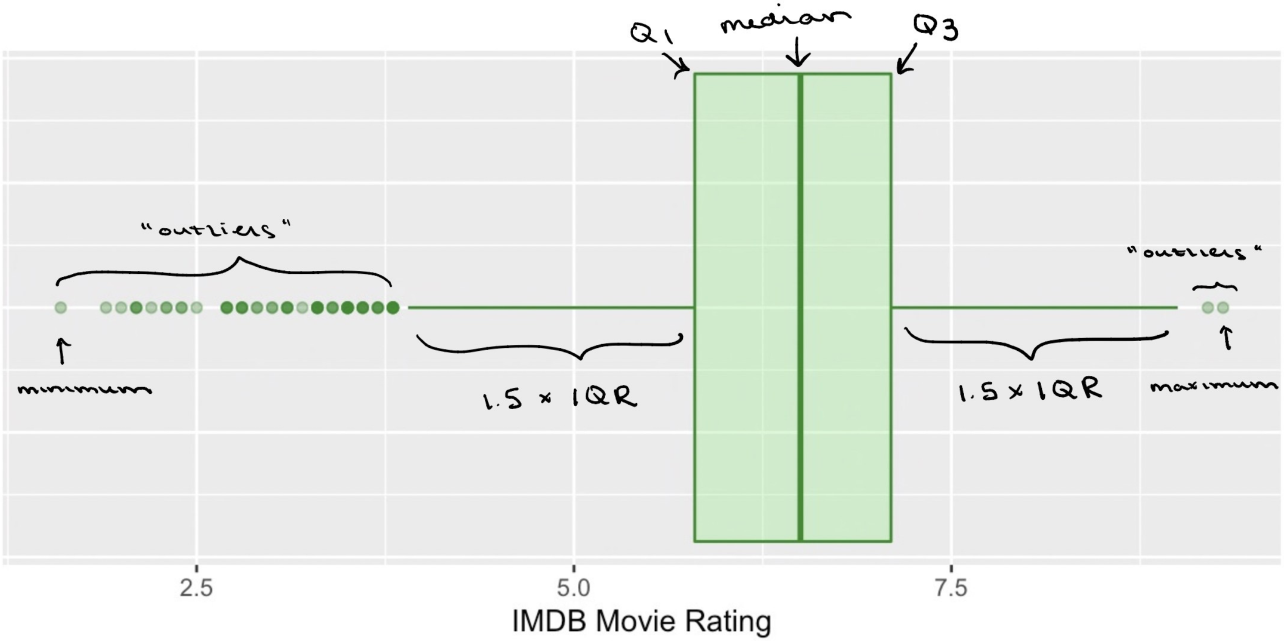
Example

Work with 2-3 other people to visualize the five-number summary of age.

The following five statistics make up the five-number summary, which captures information about both the center *and* spread of the data:

Minimum / 25th percentile (Q1) / Median / 75th percentile (Q3) / Maximum

We can use a box plot to visualize all of these statistics in one go:



EDA Practice

Open movies.csv (under Demos on the course website) in excel or google sheets.

Work with 1-2 other people.

Choose 1 categorical and 1 numerical variable. For each variable, generate the appropriate summary visualizations and summary statistics.

You in some cases, you will need to manipulate the raw data and use formulas. Helpful tips can be found here:

- Excel
  - https://www.princeton.edu/~otorres/Excel/excelstata.htm
  - https://statisticsbyjim.com/basics/descriptive-statistics-excel/
- Google Sheets
  - http://www.comfsm.fm/~dleeling/statistics/text6.html#page-031
  - https://www.groovypost.com/howto/quickly-get-column-statistics-in-google-sheets/