STAT 230

Lecture 1

Introduction to Statistics

Naima Hammoud

June 3, 2019

What is it useful for?

- Surveys and polls: for example, we can conduct polls to see who people are voting for in an election. Then, using the data from the polls, we can make statements about who is more likely to win the election.

- Understanding the relation between two variables: for example, in a given neighborhood, how is the price of a house related to the number of rooms in the house?

- Inferring causal relations: for example, does cigarette smoking cause cancer?

Lecture Objectives

- Vocabulary

- Single Variables
  - Distribution
  - Frequency distribution
  - Relative frequency distribution
  - Histogram
  - Stem-and-Leaf Displays

- Properties of a Distribution
  - Modality
  - Skew
  - Center: mean, median, trimmed mean
  - Variability: Range, interquartile range, standard deviation
  - Boxplots

Vocabulary

Population vs. Sample

A population is the entire pool from which you would like to draw information

- If you want to know the dog food brand that dog owners in the US prefer, then the population consists of every dog owner in the US.

- It may be really difficult to ask every dog owner in the US what dog food they prefer (more than 45 million dog owners in the US).

- Instead of interviewing the entire population of 45 million dog owners, we can choose a **representative sample** from which we can **infer** characteristics about the whole population.

Why only a sample?

Why do we study a small sample of the population? Why not the whole population?

- Some individuals in the population are hard to obtain
- Populations are always moving
- VERY costly!

Why only a sample?

Why do we study a small sample of the population? Why not the whole population?

- Some individuals in the population are hard to obtain
- Populations are always moving
- VERY costly!

Think of it as doing a blood test, or trying a spoon of soup from a pot to check if it needs salt. With the blood test, the nurse will not draw all your blood to perform the test! As for the soup, you wouldn't eat the whole pot to determine whether to add salt or not!

Sampling Bias

- **Convenience Bias**: individuals who are easily accessible are more likely to be included in the sample. (e.g. survey your neighbors)

- **Voluntary Response Bias**: when the sample consists of people who volunteer to respond to a survey because they feel strongly about the issue. (e.g. online polls)

- **Nonresponse Bias**: when a nonrandom sample of a randomly sampled group does not respond to a survey. (e.g. illegal immigrants)

- **Undercoverage Bias**: when some nonrandom group of the population is left out. (e.g. opinion poll of random digit dialing of landlines in the US misses 40% of Americans who do not own landlines)

Landon vs. FDR

The Literary Digest

NEW YORK OCTOBER 31, 1936

Topics of the day

LANDON, 1,293,669; ROOSEVELT, 972,897

Final Returns in The Digest's Poll of Ten Million Voters

Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the

lican National Committee purchased THE LITERARY DIGEST?" And all types and varieties, including: "Have the Jews purchased

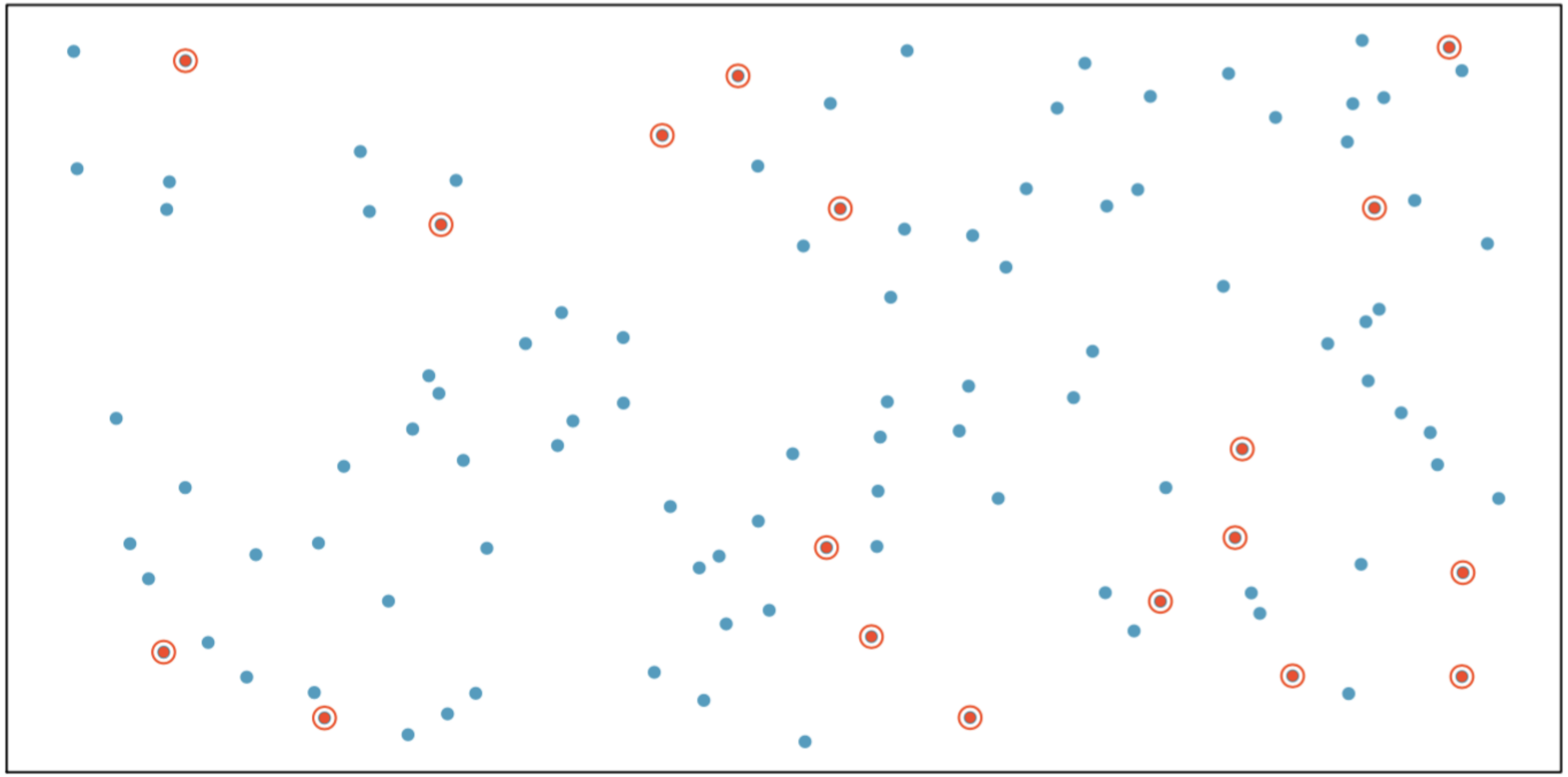
returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.'

"We never make any claims before election but we respectfully refer you to the

In 1936, the American Literary Digest magazine collected over two million surveys and predicted that the Republican nominee, Alf Landon, would beat Franklin Roosevelt 62% to 38%. The exact opposite happened!

Sampling Methods

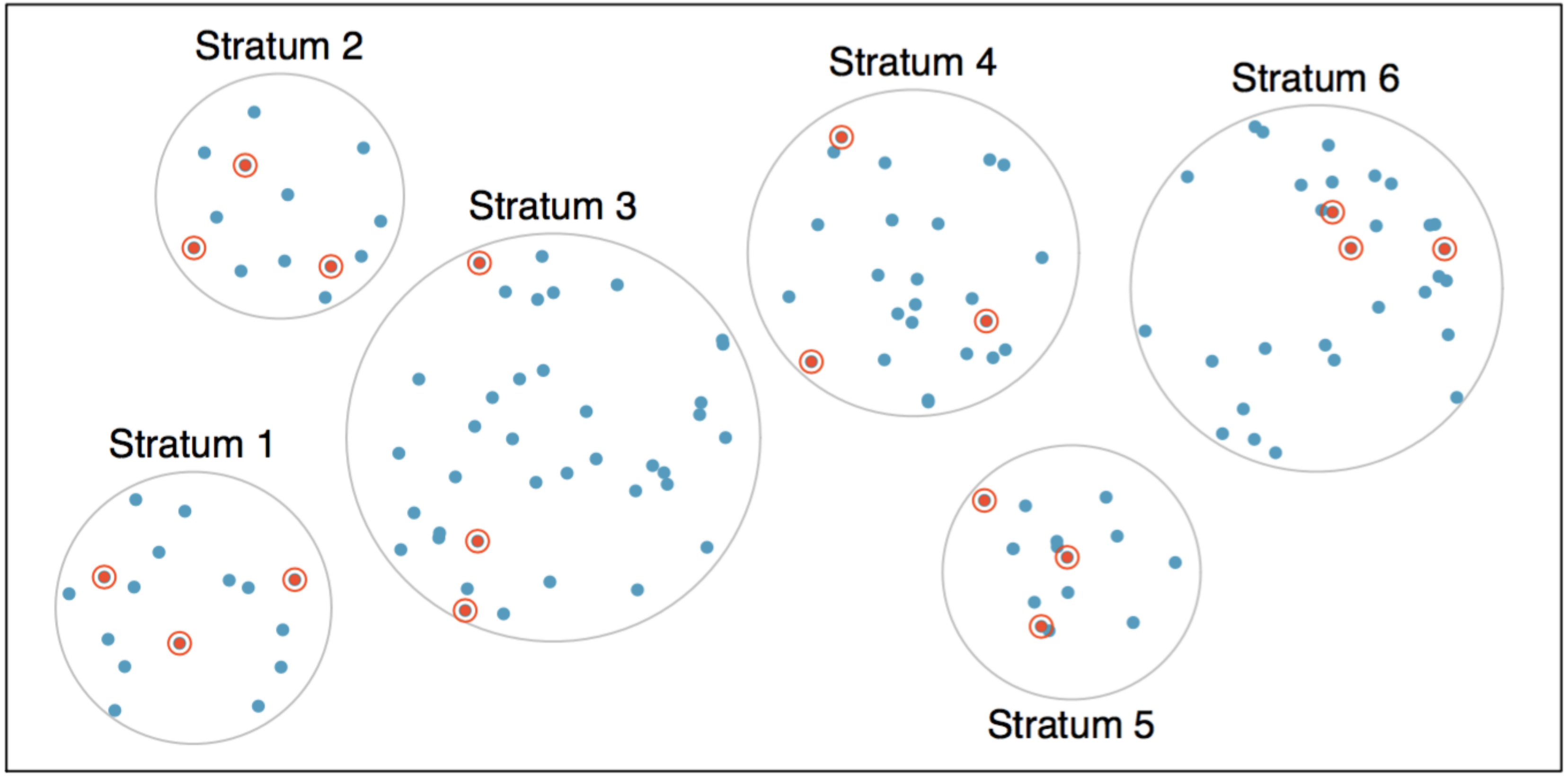
- Simple Random Sample (SRS): randomly select from the population where each individual is equally likely to be selected



Diez, Barr, Çetinkaya-Rundel, OpenIntro Statistics (*3rd Ed.*)

Sampling Methods

- Stratified Sampling:
  - divide population into homogeneous groups called strata, then randomly sample from within each stratum.
  - For example, divide population in male/female, and then randomly sample from each group (if we want male and female to be equally represented).



OpenIntro Statistics (*3rd Ed.*)

Variables

Variables are used to study a certain characteristic or trait of a population

- numerical
  - continuous
    - Your height
  - discrete
    - Number of books you own
- categorical
  - regular
    - Male
    - Female
  - ordinal
    - Very satisfied
    - Satisfied
    - Neutral
    - Dissatisfied
    - Very dissatisfied

HIV in Swaziland

year	Estimated HIV Prevalence% - (Ages 15-49)
1982	0.011
1990	2.3
1991	3.2
1992	4.4
1993	6.1
1994	8.1
1995	10.6
1996	13.3
1997	16
1998	18.5
1999	20.6
2000	22.3
2001	23.6
2002	24.5
2003	25.1
2004	25.5
2005	25.6
2006	25.7
2007	25.8
2008	25.9
2009	25.8
2010	25.9
2011	26

data from www.gapminder.com

HIV in Swaziland

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

year	Estimated HIV Prevalence% - (Ages 15-49)
1982	0.011
1990	2.3
1991	3.2
1992	4.4
1993	6.1
1994	8.1
1995	10.6
1996	13.3
1997	16
1998	18.5
1999	20.6
2000	22.3
2001	23.6
2002	24.5
2003	25.1
2004	25.5
2005	25.6
2006	25.7
2007	25.8
2008	25.9
2009	25.8
2010	25.9
2011	26

Sheet1

data from www.gapminder.com

HIV in Swaziland

year	Estimated HIV Prevalence% - (Ages 15-49)
1982	0.011
1990	2.3
1991	3.2
1992	4.4
1993	6.1
1994	8.1
1995	10.6
1996	13.3
1997	16
1998	18.5
1999	20.6
2000	22.3
2001	23.6
2002	24.5
2003	25.1
2004	25.5
2005	25.6
2006	25.7
2007	25.8
2008	25.9
2009	25.8
2010	25.9
2011	26

year is a categorical variable

data from www.gapminder.com

HIV in Swaziland

year	Estimated HIV Prevalence% - (Ages 15-49)
1982	0.011
1990	2.3
1991	3.2
1992	4.4
1993	6.1
1994	8.1
1995	10.6
1996	13.3
1997	16
1998	18.5
1999	20.6
2000	22.3
2001	23.6
2002	24.5
2003	25.1
2004	25.5
2005	25.6
2006	25.7
2007	25.8
2008	25.9
2009	25.8
2010	25.9
2011	26

year is a categorical variable

data from www.gapminder.com

HIV in Swaziland

year	Estimated HIV Prevalence% - (Ages 15-49)
1982	0.011
1990	2.3
1991	3.2
1992	4.4
1993	6.1
1994	8.1
1995	10.6
1996	13.3
1997	16
1998	18.5
1999	20.6
2000	22.3
2001	23.6
2002	24.5
2003	25.1
2004	25.5
2005	25.6
2006	25.7
2007	25.8
2008	25.9
2009	25.8
2010	25.9
2011	26

year is a categorical variable

HIV % is a numerical variable

data from www.gapminder.com

HIV in Swaziland

year	Estimated HIV Prevalence% - (Ages 15-49)
1982	0.011
1990	2.3
1991	3.2
1992	4.4
1993	6.1
1994	8.1
1995	10.6
1996	13.3
1997	16
1998	18.5
1999	20.6
2000	22.3
2001	23.6
2002	24.5
2003	25.1
2004	25.5
2005	25.6
2006	25.7
2007	25.8
2008	25.9
2009	25.8
2010	25.9
2011	26

- Start analyzing data graphically (find patterns)

data from www.gapminder.com

HIV in Swaziland

year	Estimated HIV Prevalence% - (Ages 15-49)
1982	0.011
1990	2.3
1991	3.2
1992	4.4
1993	6.1
1994	8.1
1995	10.6
1996	13.3
1997	16
1998	18.5
1999	20.6
2000	22.3
2001	23.6
2002	24.5
2003	25.1
2004	25.5
2005	25.6
2006	25.7
2007	25.8
2008	25.9
2009	25.8
2010	25.9
2011	26

- Start analyzing data graphically (find patterns)
- After that analyze data numerically (what do those patterns mean?)

data from www.gapminder.com

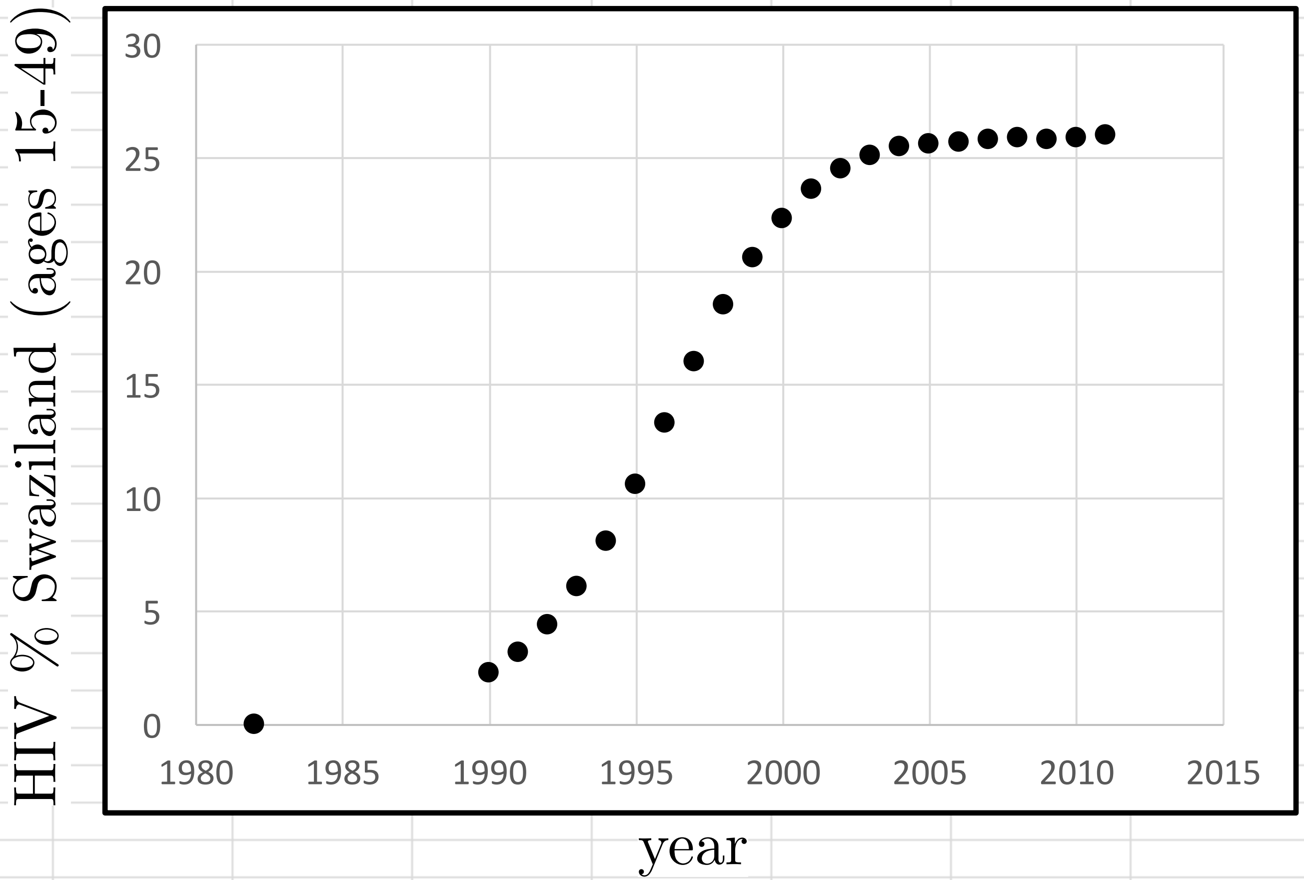
HIV in Swaziland

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

nh56@nyu.edu

Comments Share

year	Estimated HIV Prevalence% - (Ages 15-49)
1982	0.011
1990	2.3
1991	3.2
1992	4.4
1993	6.1
1994	8.1
1995	10.6
1996	13.3
1997	16
1998	18.5
1999	20.6
2000	22.3
2001	23.6
2002	24.5
2003	25.1
2004	25.5
2005	25.6
2006	25.7
2007	25.8
2008	25.9
2009	25.8
2010	25.9
2011	26



Sheet1

data from www.gapminder.com

Single Variables

Distribution

The distribution of a certain variable gives information about the values this variable takes

Distribution

The distribution of a certain variable gives information about the values this variable takes

Country	Life Expectancy (2016)
Afghanistan	52.72
Andorra	84.8
Bermuda	78.6
Cameroon	59.7
Canada	81.7
China	76.5
France	81.9
Hong Kong	83.9
North Korea	72.3
South Korea	81.1
Lesotho	48.86
Swaziland	53.88
UK	81.1
US	79.1

data from www.gapminder.com

Distribution

The distribution of a certain variable gives information about the values this variable takes

Country	Life Expectancy (2016)
Afghanistan	52.72
Andorra	84.8
Bermuda	78.6
Cameroon	59.7
Canada	81.7
China	76.5
France	81.9
Hong Kong	83.9
North Korea	72.3
South Korea	81.1
Lesotho	48.86
Swaziland	53.88
UK	81.1
US	79.1

For example, life expectancy in 2016 in countries around the world ranges from **48.86** (Lesotho) to **84.8** (Andorra)

data from www.gapminder.com

Frequency Distribution

A frequency distribution classifies data on a single variable into non-overlapping intervals and records how many times data values are in each interval children per woman (2015)

	A	B
1	Total fertility rate	2015
2	Abkhazia	
3	Afghanistan	4.47
4	Akrotiri and Dhekelia	
5	Albania	1.78
6	Algeria	2.71
7	American Samoa	
8	Andorra	
9	Angola	5.65
10	Anguilla	
11	Antigua and Barbuda	2.06
12	Argentina	2.15
13	Armenia	1.41
14	Aruba	1.66
15	Australia	1.88
16	Austria	1.5
17	Azerbaijan	1.89
18	Bahamas	1.88
19	Bahrain	2.03
20	Bangladesh	2.12
21	Barbados	1.86
22	Belarus	1.51
23	Belgium	1.86
24	Belize	2.6
25	Benin	4.69
26	Bermuda	

Sheet1

data from www.gapminder.com

Frequency Distribution

A frequency distribution classifies data on a single variable into non-overlapping intervals and records how many times data values are in each interval

We count how many countries have less than two children per woman, how many have between 2 and 3 children per woman, etc.

Frequency Distribution

A frequency distribution classifies data on a single variable into non-overlapping intervals and records how many times data values are in each interval

We count how many countries have less than two children per woman, how many have between 2 and 3 children per woman, etc.

children per woman	1-2	2-3	3-4	4-5	5-6	6-7	7-8
frequency	77	61	23	22	12	3	1

Frequency Distribution

A frequency distribution classifies data on a single variable into non-overlapping intervals and records how many times data values are in each interval

We count how many countries have less than two children per woman, how many have between 2 and 3 children per woman, etc.

children per woman	1-2	2-3	3-4	4-5	5-6	6-7	7-8
frequency	77	61	23	22	12	3	1

number of countries

Relative Frequency Distribution

A relative frequency distribution classifies data on a single variable into non-overlapping intervals and records what % of data values are in each interval

Relative Frequency Distribution

A relative frequency distribution classifies data on a single variable into non-overlapping intervals and records what % of data values are in each interval

We calculate the percentage (or probability) of countries that have less than two children per woman, those that have between 2 and 3 children per woman, etc.

Relative Frequency Distribution

A relative frequency distribution classifies data on a single variable into non-overlapping intervals and records what % of data values are in each interval

We calculate the percentage (or probability) of countries that have less than two children per woman, those that have between 2 and 3 children per woman, etc.

children per woman	1-2	2-3	3-4	4-5	5-6	6-7	7-8
frequency	77	61	23	22	12	3	1

Relative Frequency Distribution

A relative frequency distribution classifies data on a single variable into non-overlapping intervals and records what % of data values are in each interval

We calculate the percentage (or probability) of countries that have less than two children per woman, those that have between 2 and 3 children per woman, etc.

children per woman	1-2	2-3	3-4	4-5	5-6	6-7	7-8
frequency	77	61	23	22	12	3	1

total of 199 countries

Relative Frequency Distribution

A relative frequency distribution classifies data on a single variable into non-overlapping intervals and records what % of data values are in each interval

We calculate the percentage (or probability) of countries that have less than two children per woman, those that have between 2 and 3 children per woman, etc.

children per woman	1-2	2-3	3-4	4-5	5-6	6-7	7-8
frequency	77	61	23	22	12	3	1
Relative frequency	77/199 0.387	61/199 0.306	23/199 0.116	22/199 0.111	12/199 0.06	3/199 0.015	1/199 0.005

Relative Frequency Distribution

A relative frequency distribution classifies data on a single variable into non-overlapping intervals and records what % of data values are in each interval

We calculate the percentage (or probability) of countries that have less than two children per woman, those that have between 2 and 3 children per woman, etc.

children per woman	1-2	2-3	3-4	4-5	5-6	6-7	7-8
frequency	77	61	23	22	12	3	1
Relative frequency	77/199 0.387	61/199 0.306	23/199 0.116	22/199 0.111	12/199 0.06	3/199 0.015	1/199 0.005

NOTE: $0.387 + 0.306 + 0.116 + 0.111 + 0.06 + 0.015 + 0.005 = 1$

Histogram

A histogram is a graphical representation of a frequency distribution for a single numerical variable

Histogram

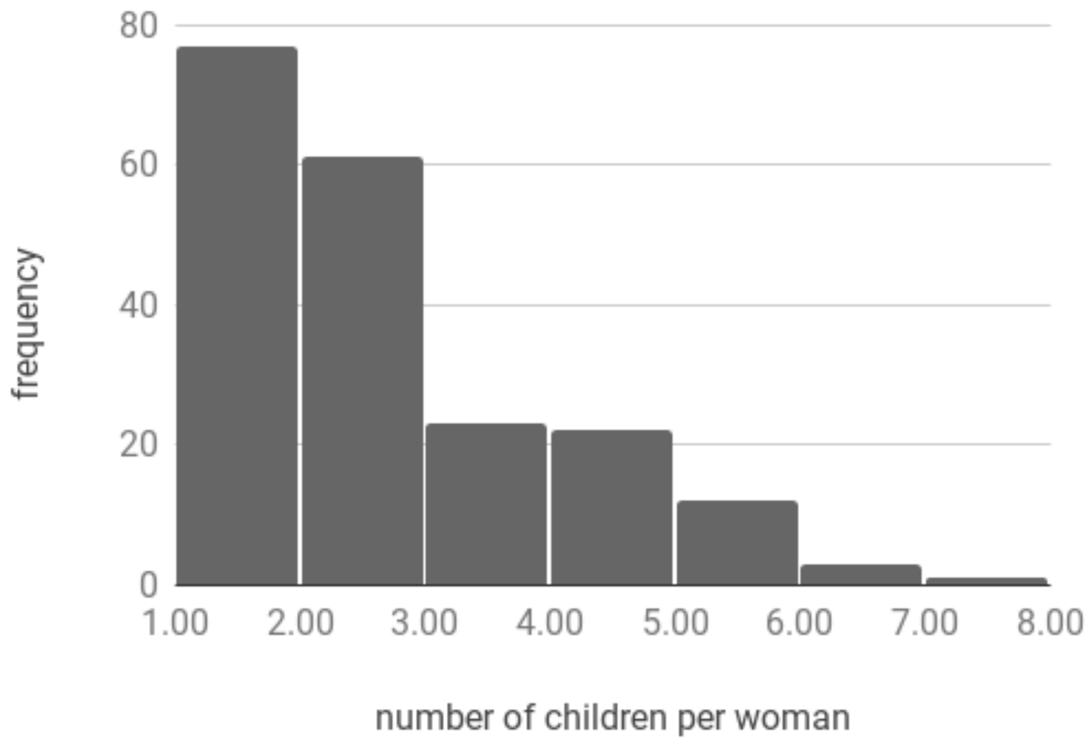
A histogram is a graphical representation of a frequency distribution for a single numerical variable

children per woman	1-2	2-3	3-4	4-5	5-6	6-7	7-8
frequency	77	61	23	22	12	3	1

Histogram

A histogram is a graphical representation of a frequency distribution for a single numerical variable

children per woman	1-2	2-3	3-4	4-5	5-6	6-7	7-8
frequency	77	61	23	22	12	3	1



Histogram

Stem and Leaf Plots

If the entries in a dataset have two or more digits, we can create a stem-and-leaf display by choosing a **stem** which is made of one or more leading digits, and **leaves**, which consist of the remaining trailing digits.

Stem and Leaf Plots

If the entries in a dataset have two or more digits, we can create a stem-and-leaf display by choosing a **stem** which is made of one or more leading digits, and **leaves**, which consist of the remaining trailing digits.

Example: Consider the dataset of 20 exam scores (out of 100)

61, 63, 68, 72, 75, 75, 77, 78, 79, 79, 82, 83, 86, 87, 87, 89, 90, 91, 92, 93

Stem and Leaf Plots

If the entries in a dataset have two or more digits, we can create a stem-and-leaf display by choosing a **stem** which is made of one or more leading digits, and **leaves**, which consist of the remaining trailing digits.

Example: Consider the dataset of 20 exam scores (out of 100)

61, 63, 68, 72, 75, 75, 77, 78, 79, 79, 82, 83, 86, 87, 87, 89, 90, 91, 92, 93

Choose stem=tens digit, leaf=ones digit

Stem and Leaf Plots

If the entries in a dataset have two or more digits, we can create a stem-and-leaf display by choosing a **stem** which is made of one or more leading digits, and **leaves**, which consist of the remaining trailing digits.

Example: Consider the dataset of 20 exam scores (out of 100)

61, 63, 68, 72, 75, 75, 77, 78, 79, 79, 82, 83, 86, 87, 87, 89, 90, 91, 92, 93

6	
7	
8	
9	

Stem and Leaf Plots

If the entries in a dataset have two or more digits, we can create a stem-and-leaf display by choosing a **stem** which is made of one or more leading digits, and **leaves**, which consist of the remaining trailing digits.

Example: Consider the dataset of 20 exam scores (out of 100)

61, 63, 68, 72, 75, 75, 77, 78, 79, 79, 82, 83, 86, 87, 87, 89, 90, 91, 92, 93

6	1 3 8
7	2 5 5 7 8 9 9
8	2 3 6 7 7 9
9	0 1 2 3

Stem and Leaf Plots

If the entries in a dataset have two or more digits, we can create a stem-and-leaf display by choosing a **stem** which is made of one or more leading digits, and **leaves**, which consist of the remaining trailing digits.

Example: Consider the dataset of 20 exam scores (out of 100)

61, 63, 68, 72, 75, 75, 77, 78, 79, 79, 82, 83, 86, 87, 87, 89, 90, 91, 92, 93

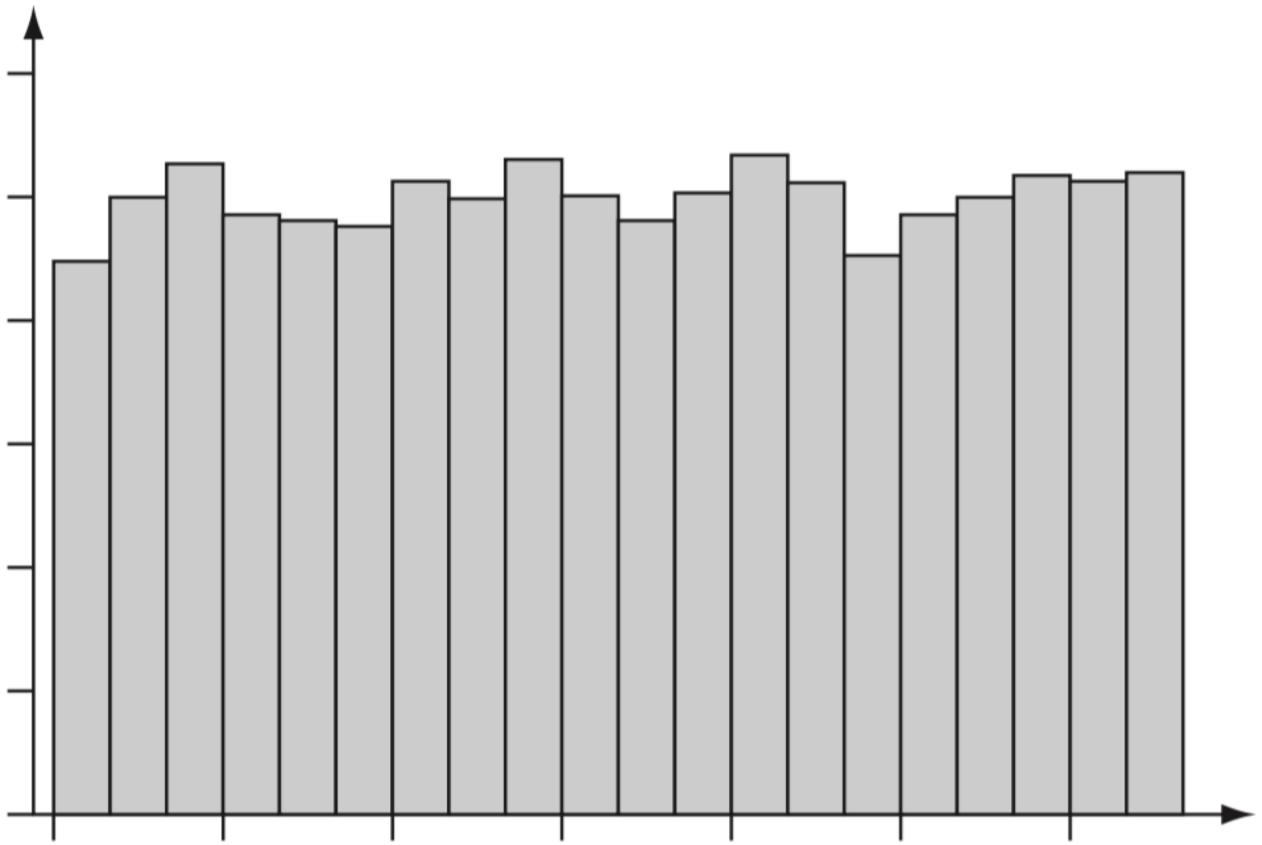
6	1 3 8
7	2 5 5 7 8 9 9
8	2 3 6 7 7 9
9	0 1 2 3

This display conveys information about typical values, the spread about a typical value, the shape of the distribution, and outliers

Properties of a distribution

Modality

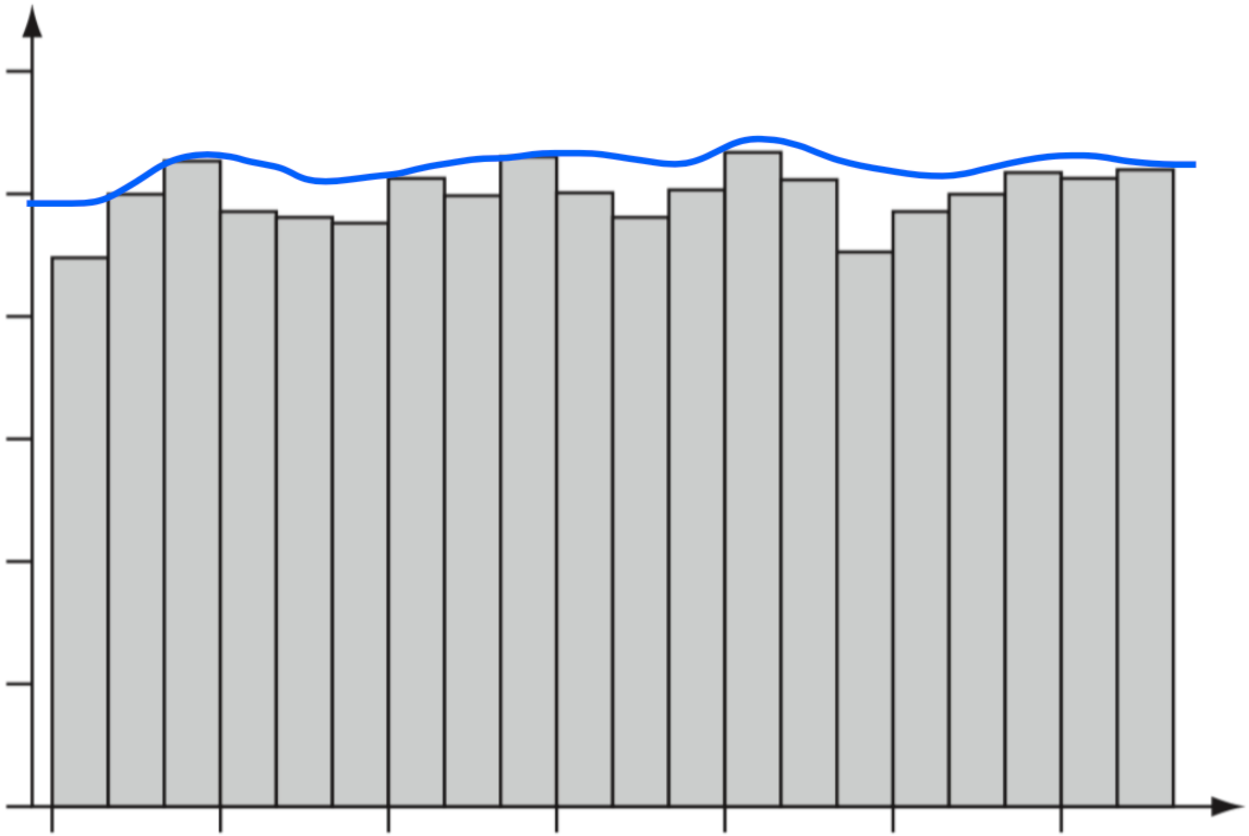
uniform



Devore, *Probability and Statistics 9th Edition* (2016)

Modality

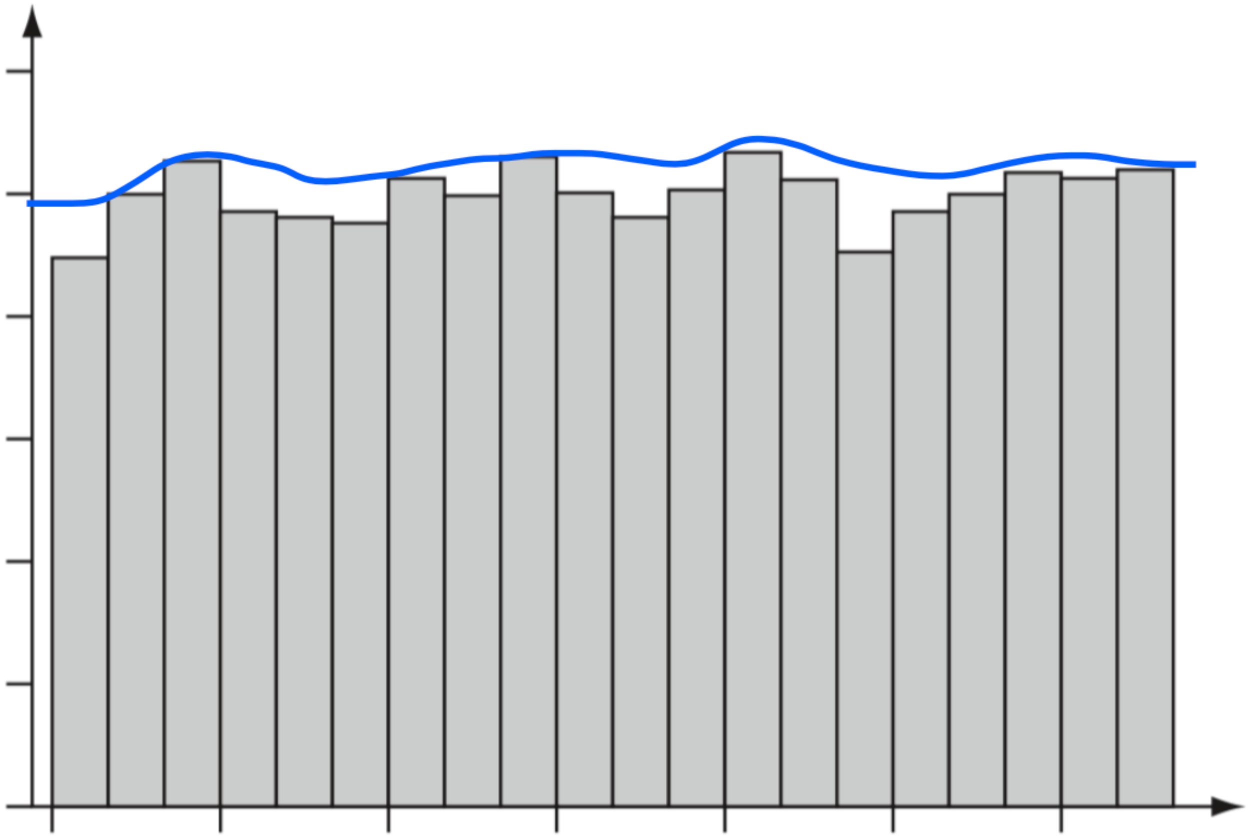
uniform

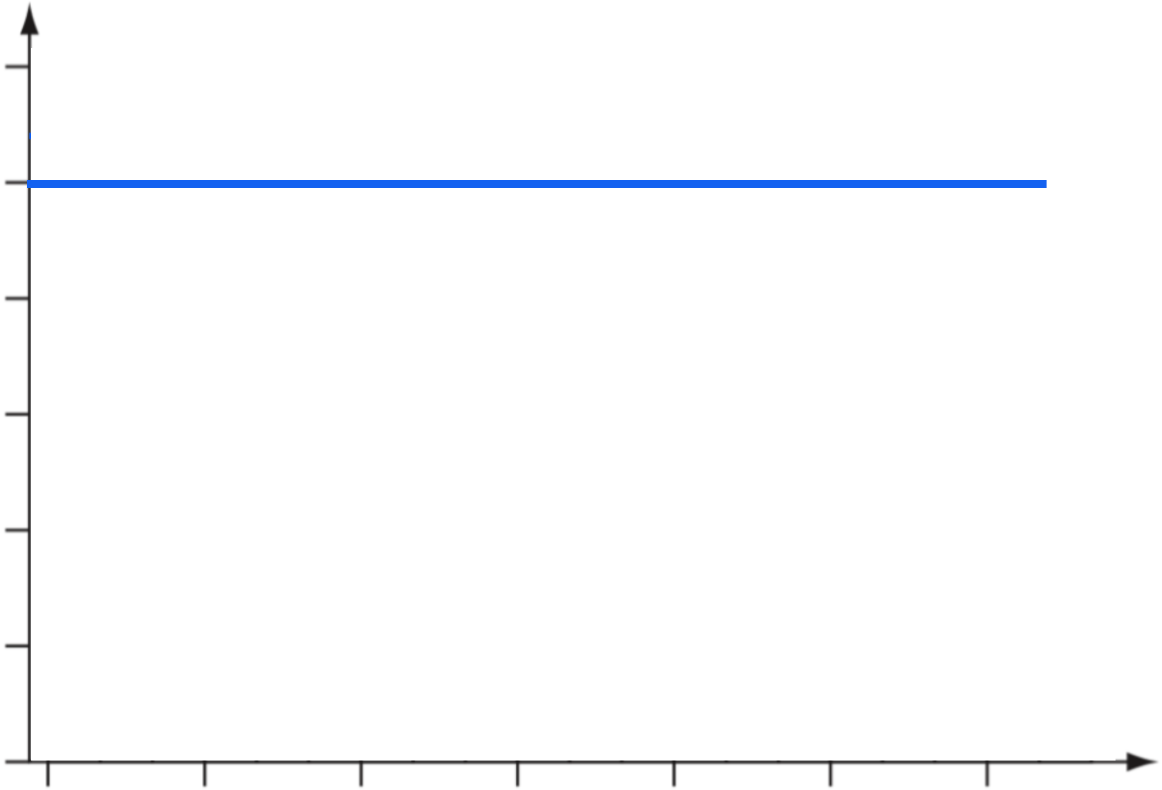


Devore, *Probability and Statistics 9th Edition* (2016)

Modality

uniform

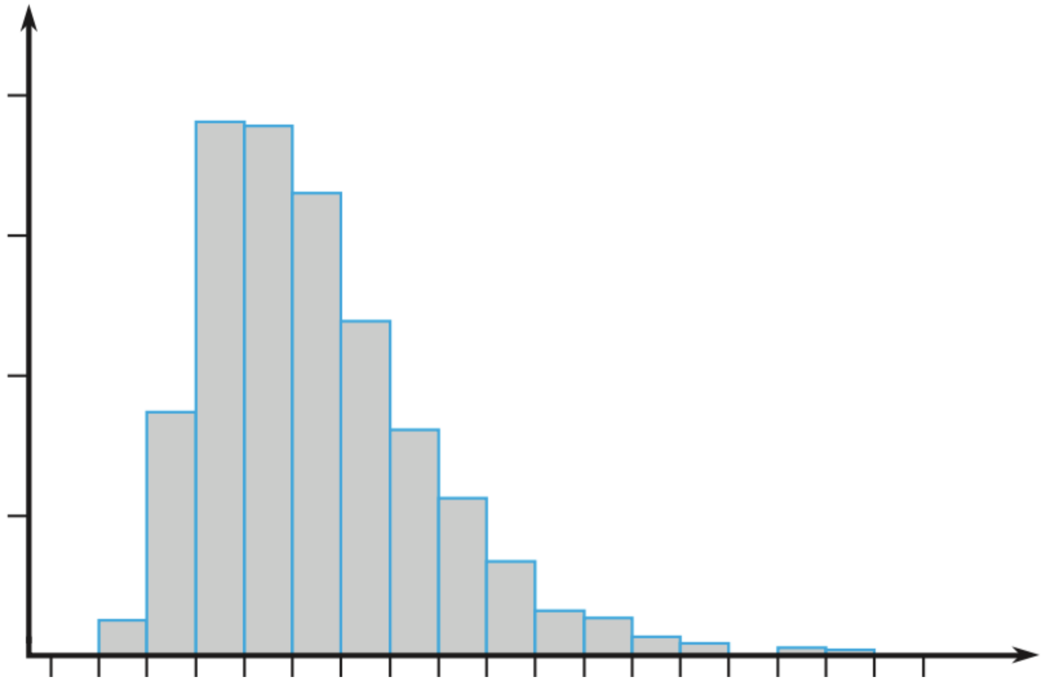


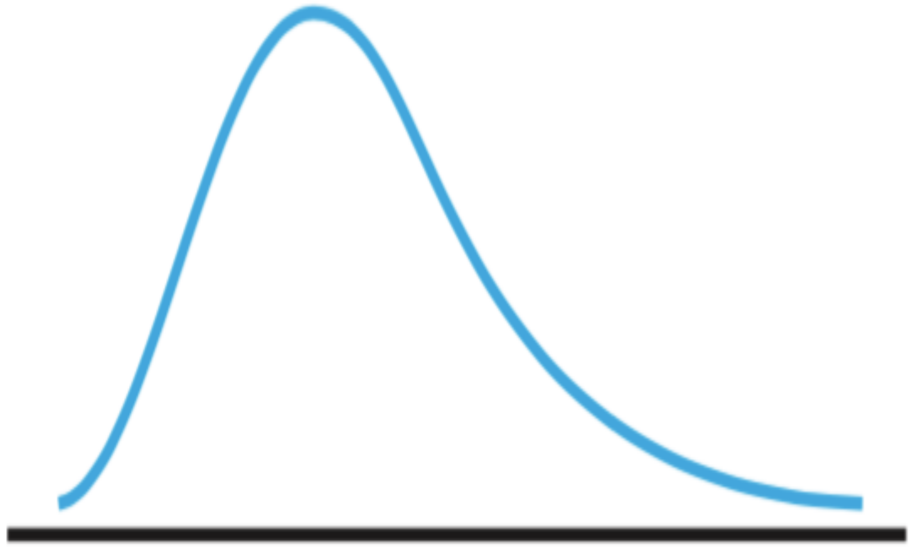


Devore, *Probability and Statistics 9th Edition* (2016)

Modality

unimodal

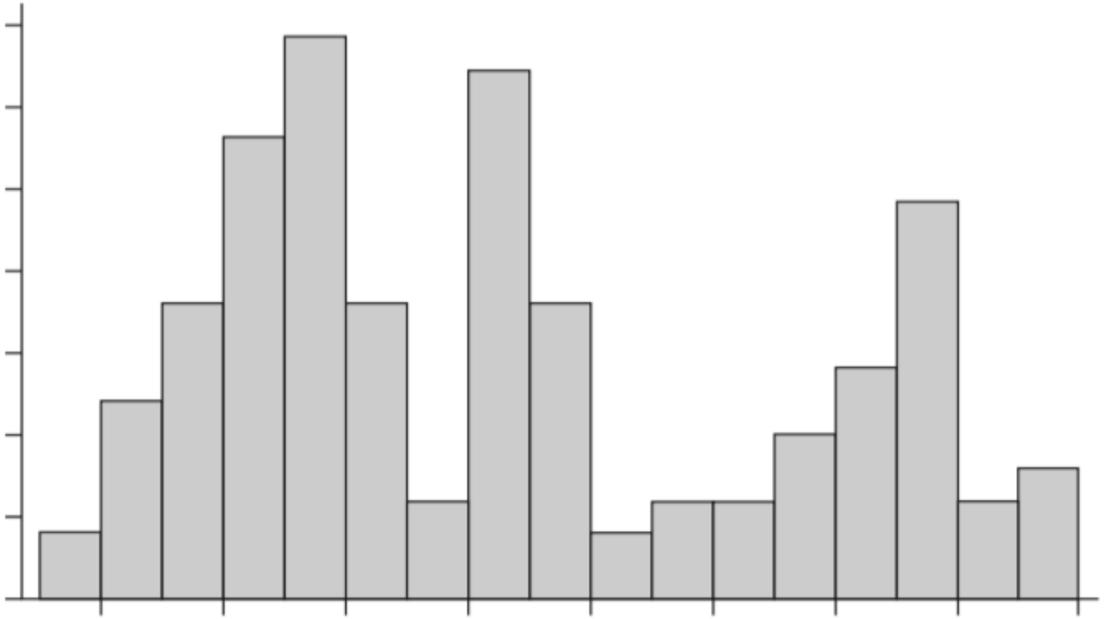


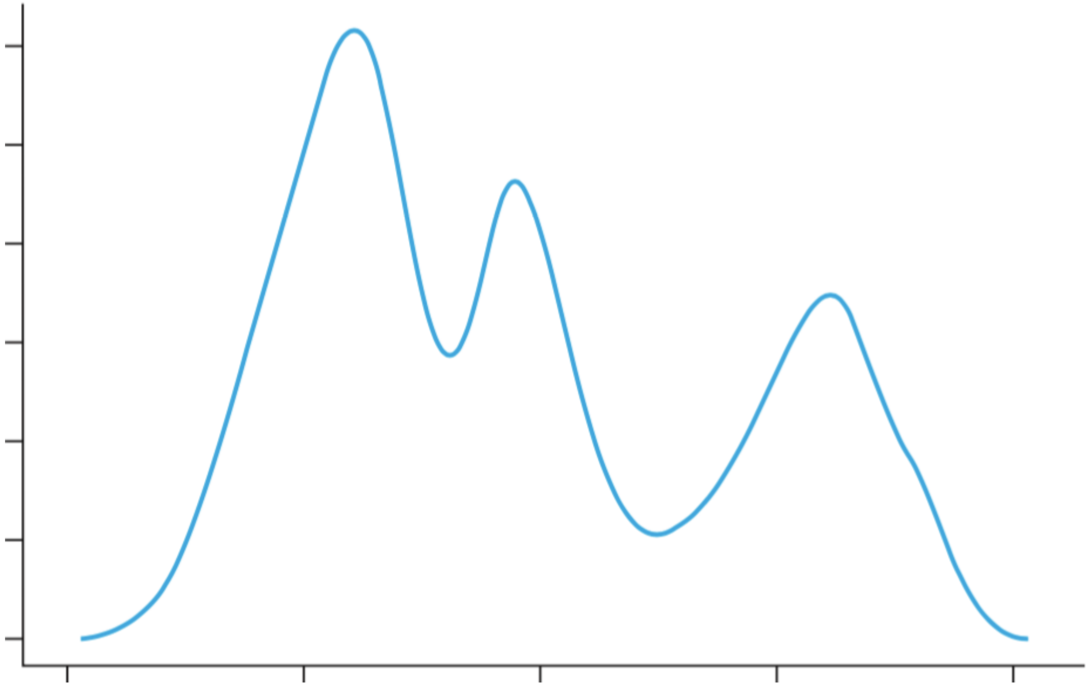


Devore, *Probability and Statistics 9th Edition* (2016)

Modality

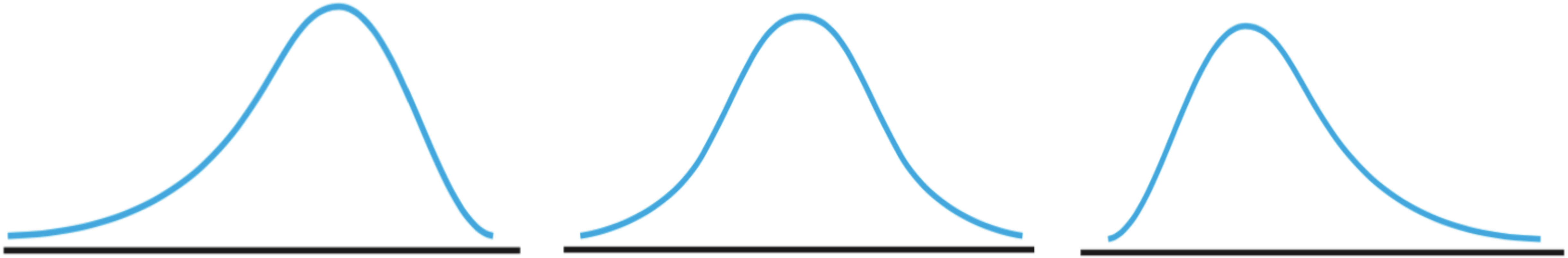
multi-modal





Devore, *Probability and Statistics 9th Edition* (2016)

Skew



left skewed

symmetric

no skew

right skewed

Measures of center

- **Mean**: arithmetic average of (for a population denoted μ and for a sample denoted $\bar{x}$)

- **Median**: midpoint of the data

Measures of center

9 scores on a homework (maximum score is 50):

38, 45, 35, 47, 42, 41, 50, 42, 39

$$\text{mean} : \bar{x} = \frac{38 + 45 + 35 + 47 + 42 + 41 + 50 + 42 + 39}{9} = 42.1$$

$$\text{median} : 35,\ 38,\ 39,\ 41,\ 42,\ 42,\ 45,\ 47,\ 50$$

Measures of center

10 scores on a homework (maximum score is 50):

35, 38, 39, 40, 41, 42, 42, 45, 47, 50

$$\text{median} : \tilde{x} = \frac{41 + 42}{2} = 41.5$$

Measures of center

Suppose someone makes a data entry error, and instead of entering 50, they enter 500

35, 38, 39, 40, 41, 42, 42, 45, 47, 500

$$\bar{x} = \frac{35 + 38 + 39 + 40 + 41 + 42 + 42 + 45 + 47 + 500}{10} = 86.9$$

$$\tilde{x} = \frac{41 + 42}{2} = 41.5$$

So the mean is very sensitive to an outlier, whereas the median isn't at all.

This type of extreme behavior is usually undesirable, which is why we would like some sort of compromise between the two.

Measures of center

Trimmed mean: remove x% of the smallest and largest parts of the data, e.g. a 10% trimmed mean is computed by eliminating the smallest and largest 10% of the data, and then taking the average of what remains

Example: Consider the following dataset (example 1.16 in your book) and compute the 10% trimmed mean

2.0	2.4	2.5	2.6	2.6	2.7	2.7	2.8	3.0	3.1	3.2	3.3	3.3
3.4	3.4	3.6	3.6	3.6	3.6	3.7	4.4	4.6	4.7	4.8	5.3	10.1

There are 26 data entries here; 10% of 26 is 2.6, so we have to remove the first "2.6 data points" and the last "2.6 data points". Since we can't really do that, we find the mean by removing two elements, then three elements, and then interpolate

Measures of center

Trimmed mean: remove x% of the smallest and largest parts of the data, e.g. a 10% trimmed mean is computed by eliminating the smallest and largest 10% of the data, and then taking the average of what remains

Example: Consider the following dataset (example 1.16 in your book) and compute the 10% trimmed mean

~~2.0~~	~~2.4~~	2.5	2.6	2.6	2.7	2.7	2.8	3.0	3.1	3.2	3.3	3.3
3.4	3.4	3.6	3.6	3.6	3.6	3.7	4.4	4.6	4.7	4.8	~~5.3~~	~~10.1~~

2 data points is (2/26)*100 = 7.7% trimming

$$\bar{x}_{tr(7.7)} = \frac{2.5 + 2.6 + 2.6 + \cdots + 4.7 + 4.8}{22} = 3.42$$

Measures of center

Trimmed mean: remove x% of the smallest and largest parts of the data, e.g. a 10% trimmed mean is computed by eliminating the smallest and largest 10% of the data, and then taking the average of what remains

Example: Consider the following dataset (example 1.16 in your book) and compute the 10% trimmed mean

~~2.0~~	~~2.4~~	~~2.5~~	2.6	2.6	2.7	2.7	2.8	3.0	3.1	3.2	3.3	3.3
3.4	3.4	3.6	3.6	3.6	3.6	3.7	4.4	4.6	4.7	~~4.8~~	~~5.3~~	~~10.1~~

3 data points is (3/26)*100 = 11.5% trimming

$$\bar{x}_{tr(11.5)} = \frac{2.6 + 2.6 + \cdots + 4.7}{20} = 3.39$$

Measures of center

Trimmed mean: remove x% of the smallest and largest parts of the data, e.g. a 10% trimmed mean is computed by eliminating the smallest and largest 10% of the data, and then taking the average of what remains

Example: Consider the following dataset (example 1.16 in your book) and compute the 10% trimmed mean

2.0	2.4	2.5	2.6	2.6	2.7	2.7	2.8	3.0	3.1	3.2	3.3	3.3
3.4	3.4	3.6	3.6	3.6	3.6	3.7	4.4	4.6	4.7	4.8	5.3	10.1

To get a 10% trimmed mean $\longrightarrow$ linear interpolation between 7.7% and 11.5%

$$\frac{\bar{x}_{tr(11.5)} - \bar{x}_{tr(7.7)}}{11.5 - 7.7} = \frac{\bar{x}_{tr(11.5)} - \bar{x}_{tr(10)}}{11.5 - 10}$$

Measures of center

Trimmed mean: remove x% of the smallest and largest parts of the data, e.g. a 10% trimmed mean is computed by eliminating the smallest and largest 10% of the data, and then taking the average of what remains

Example: Consider the following dataset (example 1.16 in your book) and compute the 10% trimmed mean

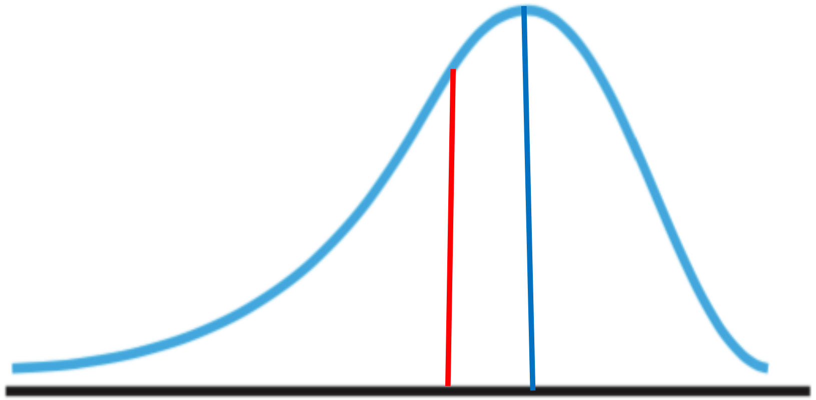
2.0	2.4	2.5	2.6	2.6	2.7	2.7	2.8	3.0	3.1	3.2	3.3	3.3
3.4	3.4	3.6	3.6	3.6	3.6	3.7	4.4	4.6	4.7	4.8	5.3	10.1

To get a 10% trimmed mean $\longrightarrow$ linear interpolation between 7.7% and 11.5%

$$\bar{x}_{tr(10)} = \frac{(10 - 7.7)(3.39) + (11.5 - 10)(3.42)}{11.5 - 7.7} = 3.40$$

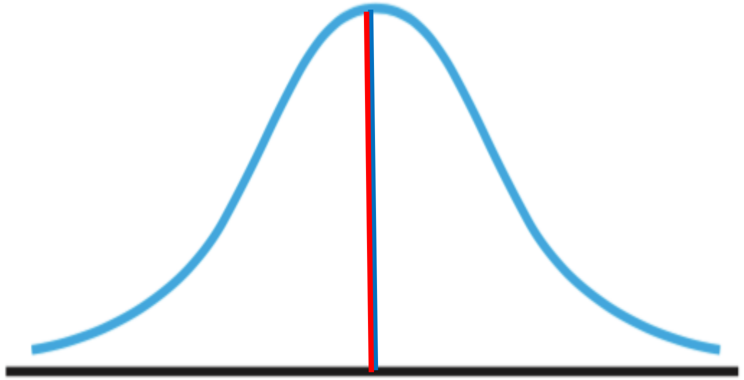
Mean, median and skew

left skewed



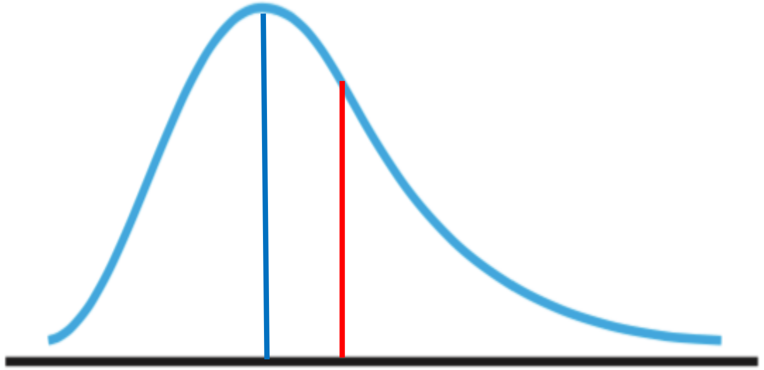
mean < median

symmetric



mean ~ median

right skewed



mean > median

Distributions

- A **distribution** of a population (or of sampled data from the population) is a representation that shows all the possible data outcomes and their frequency (how often they occur)

Distributions

- A **distribution** of a population (or of sampled data from the population) is a representation that shows all the possible data outcomes and their frequency (how often they occur)

- So far we learned how to measure the center of a distribution: mean and median

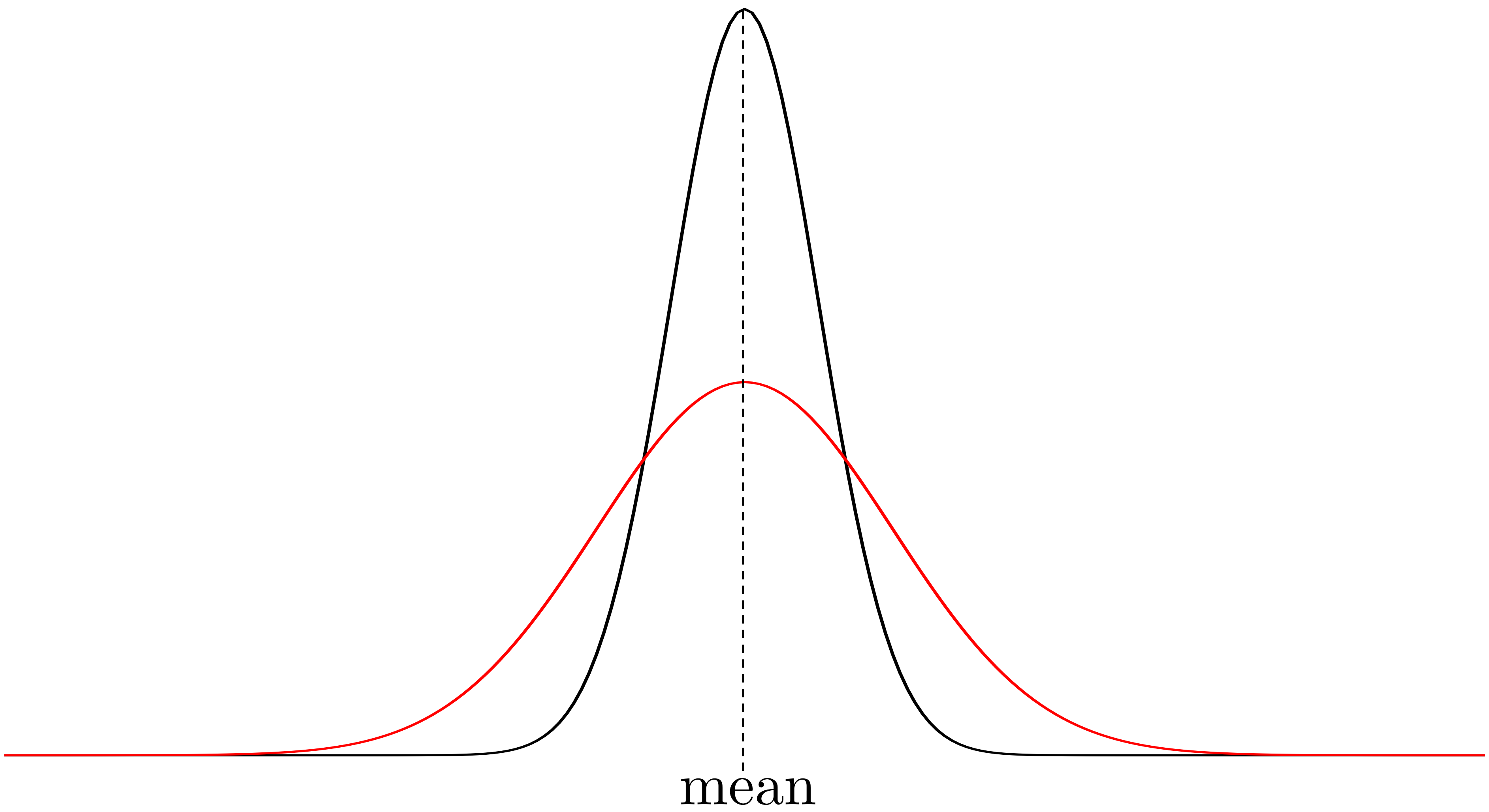
Distributions

- A **distribution** of a population (or of sampled data from the population) is a representation that shows all the possible data outcomes and their frequency (how often they occur)

- So far we learned how to measure the center of a distribution: mean and median

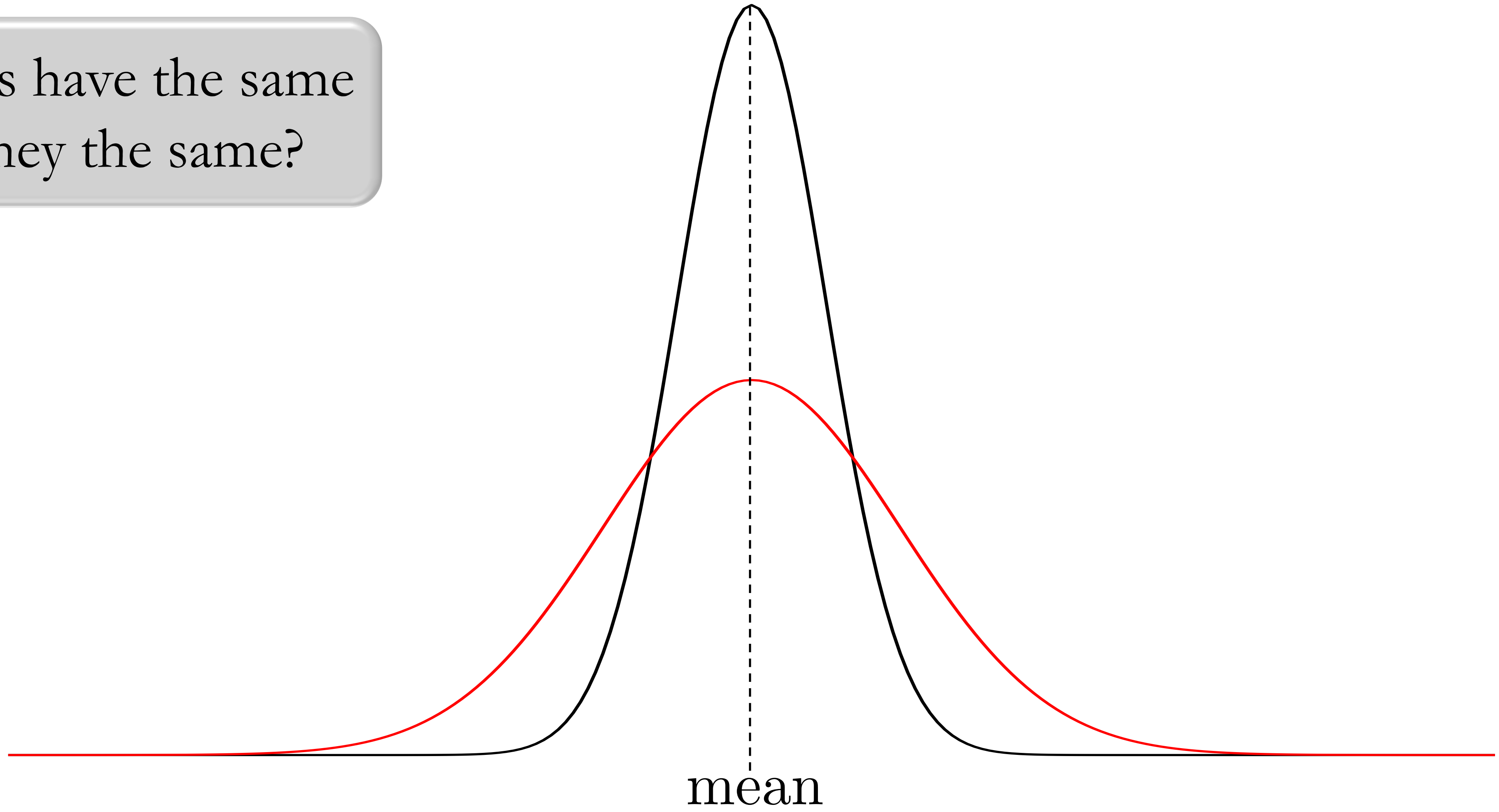
- But the center alone does not give us enough details about the distribution

Distributions



Distributions

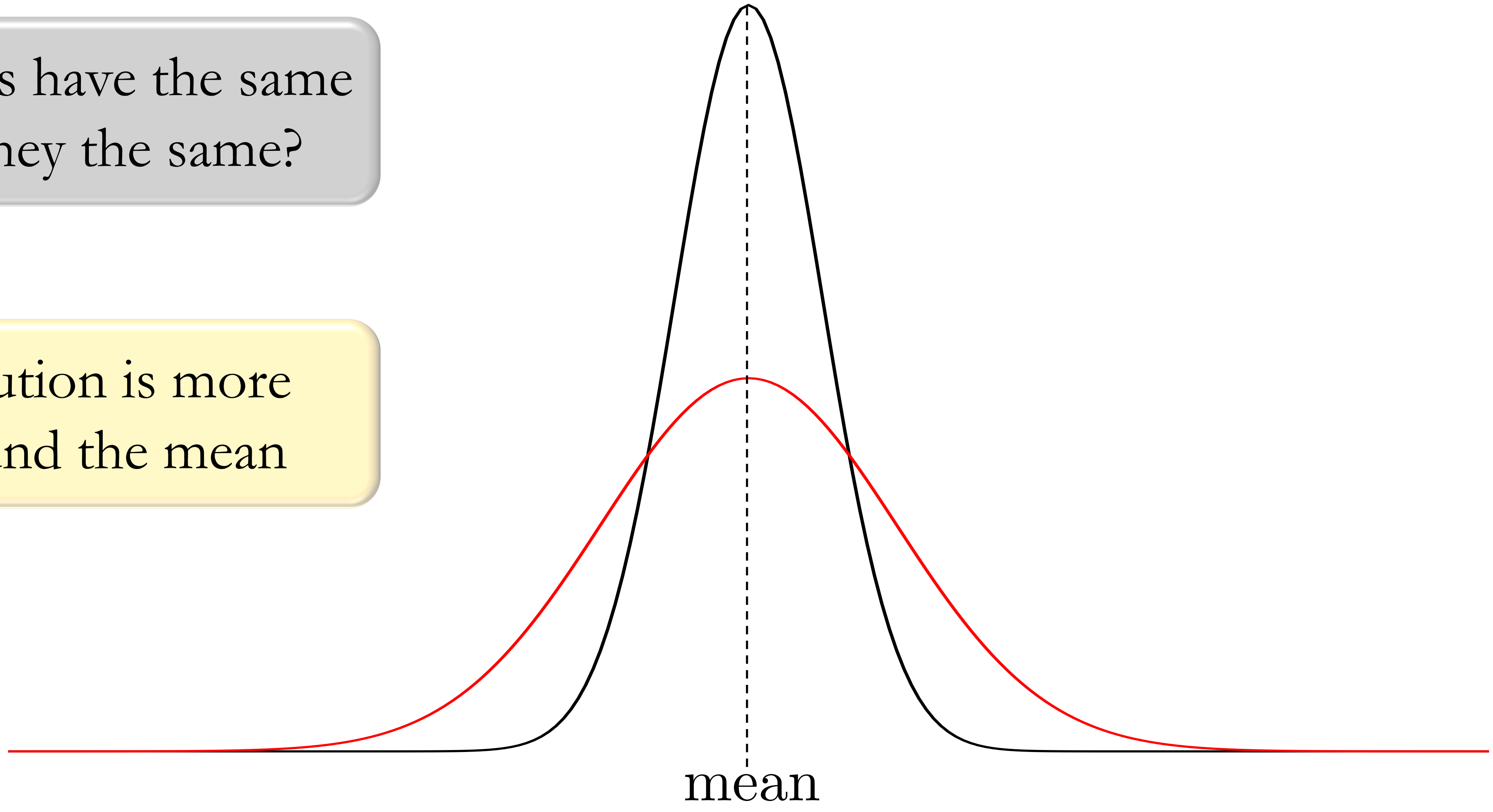
These distributions have the same center, but are they the same?



Distributions

These distributions have the same center, but are they the same?

The **red** distribution is more spread out around the mean



Distributions

- A **distribution** of a population (or of sampled data from the population) is a representation that shows all the possible data outcomes and their frequency (how often they occur)

- So far we learned how to measure the center of a distribution: mean and median

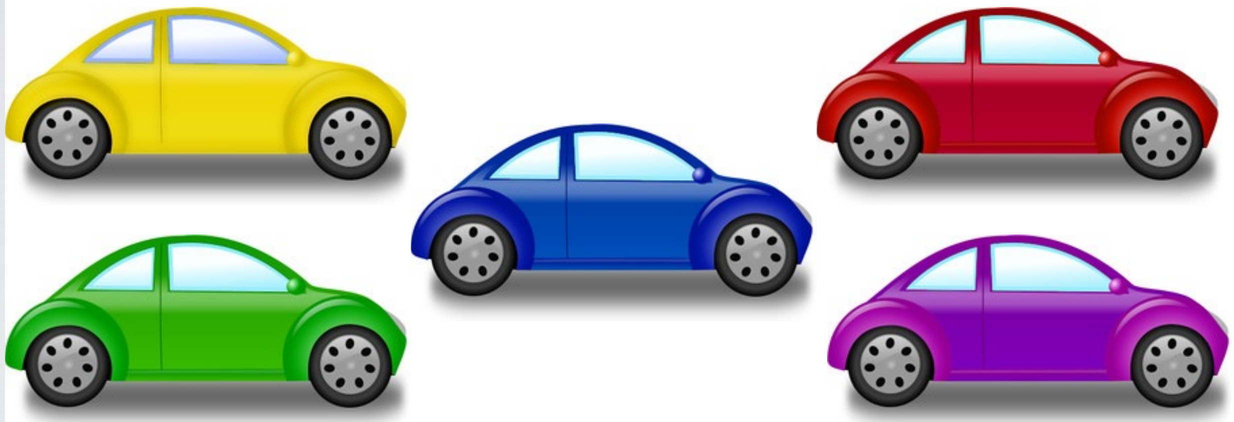
- But the center alone does not give us enough details about the distribution

- We need to measure the variability or **spread** of a distribution

Diversity vs Variability

Which of the following sets of cars has more diversity?

☐ SET 1



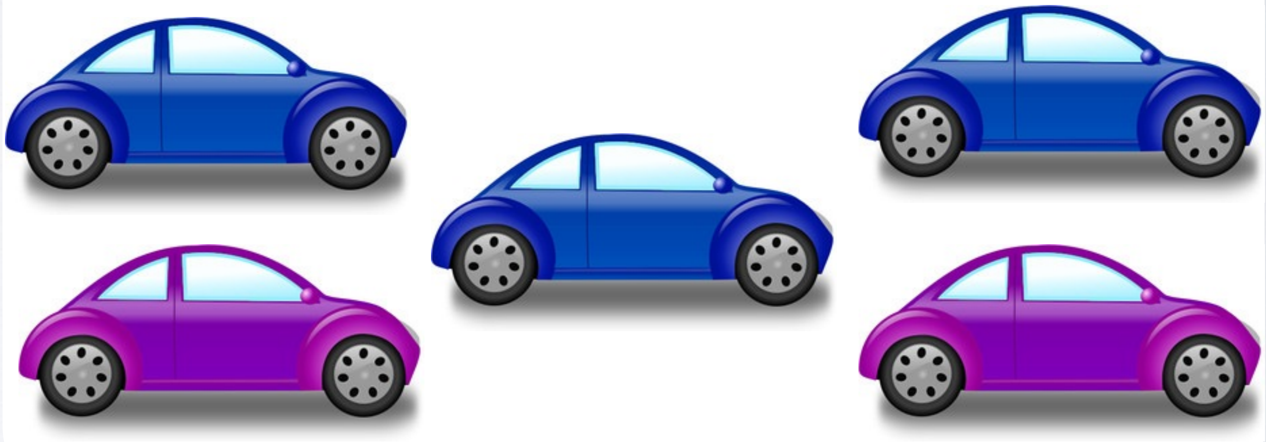
☐ SET 2

Diversity vs Variability

Which of the following sets of cars has more diversity?

☑ SET 1

☐ SET 2

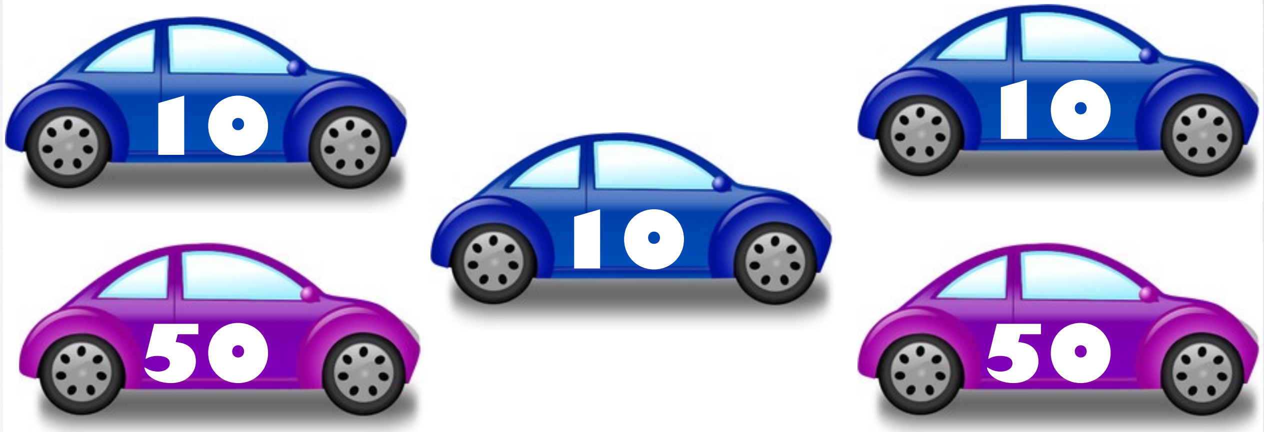


Diversity vs Variability

Which of the following sets of cars has more **variable** mileage?

☐ SET 1

☐ SET 2

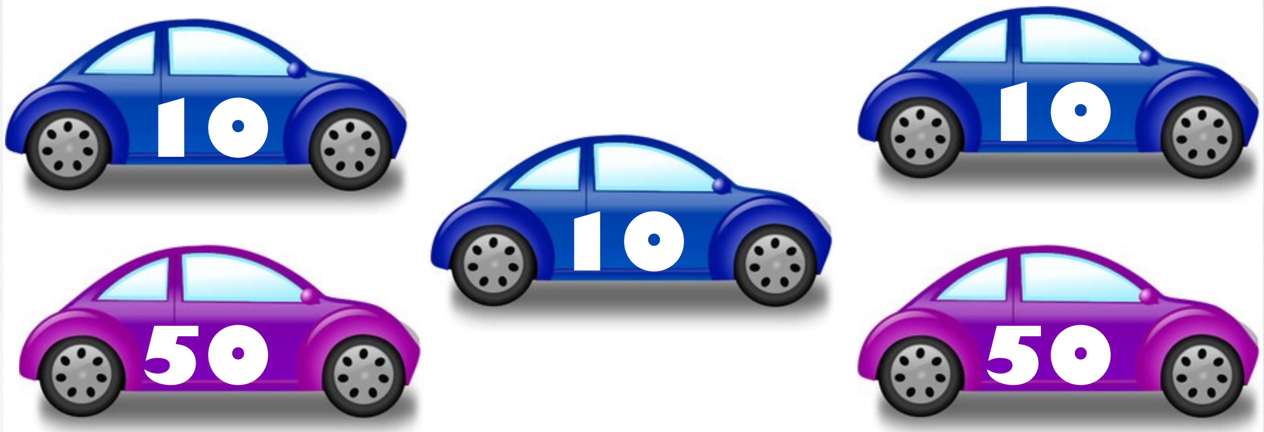


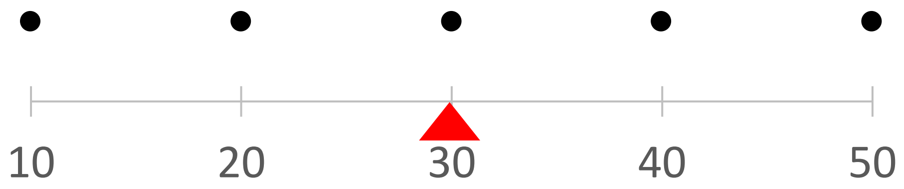
Diversity vs Variability

Which of the following sets of cars has more **variable** mileage?

☐ SET 1

☐ SET 2

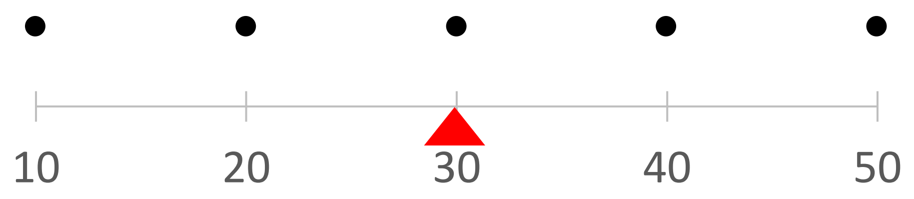




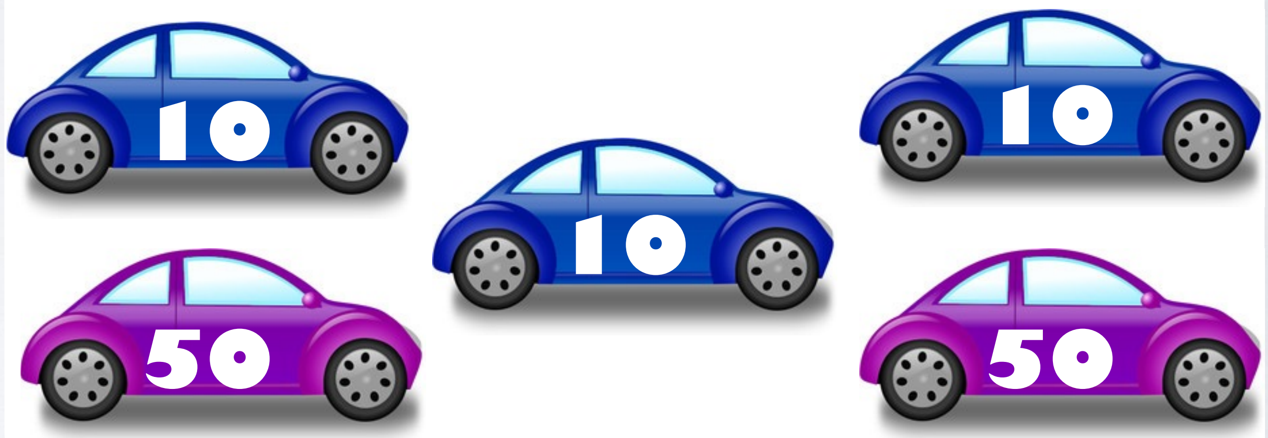
Diversity vs Variability

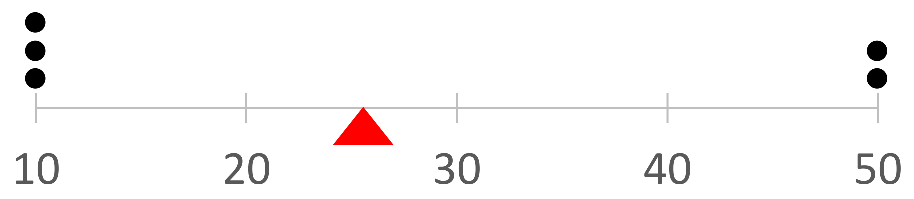
Which of the following sets of cars has more **variable** mileage?

☐ SET 1



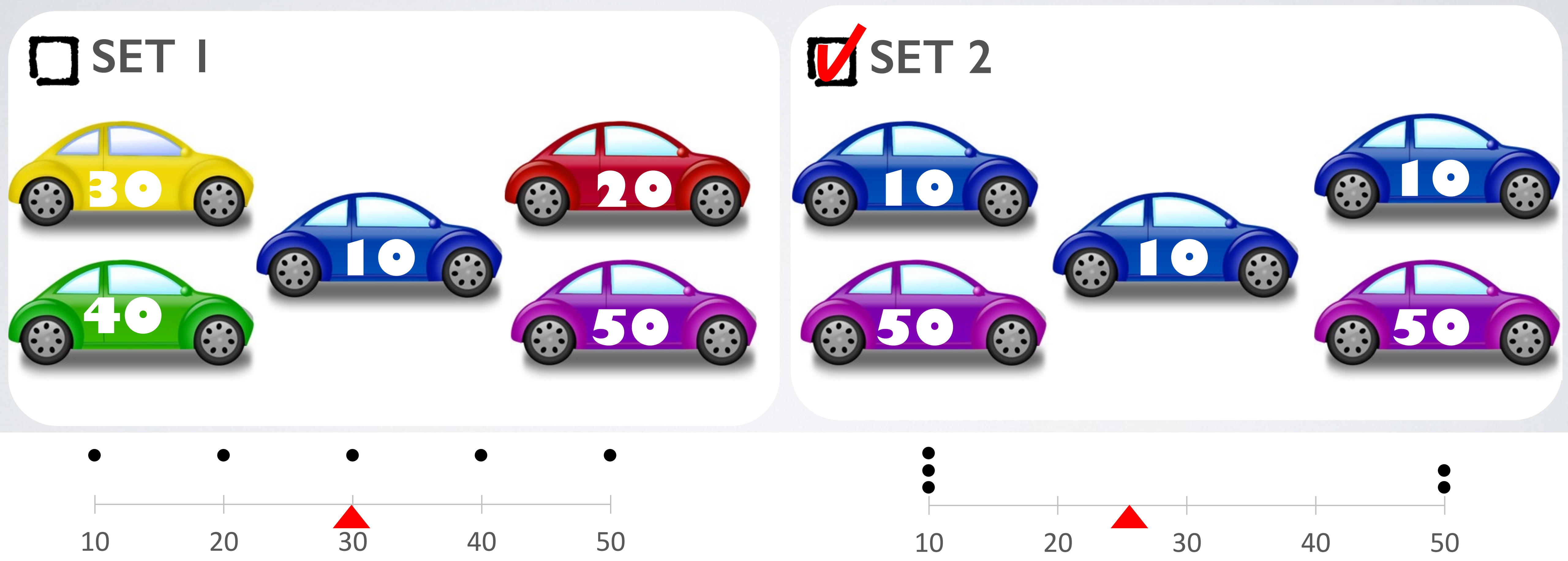
☐ SET 2





Diversity vs Variability

Which of the following sets of cars has more **variable** mileage?



www.coursera.org/learn/probability-intro/lecture/t9Wbk/measures-of-spread

Measuring Variability

Range

- The easiest measure is the range: **max – min**

  (not a very useful measure because it only looks at the extremes)

Interquartile Range

- The easiest measure is the range: **max – min**
  (not a very useful measure because it only looks at the extremes)

- Another measure is the **interquartile range**, where we need to measure the **first quartile** (25^{th} percentile), and the **third quartile** (75^{th} percentile)

Interquartile Range

- The easiest measure is the range: **max – min**

  (not a very useful measure because it only looks at the extremes)

- Another measure is the **interquartile range**, where we need to measure the **first quartile** (25th percentile), and the **third quartile** (75th percentile)

Q_1

Interquartile Range

- The easiest measure is the range: **max – min**
  (not a very useful measure because it only looks at the extremes)

- Another measure is the **interquartile range**, where we need to measure the **first quartile** (25^{th} percentile), and the **third quartile** (75^{th} percentile)

Q_1 Q_3

Interquartile Range

- The easiest measure is the range: **max – min**

  (not a very useful measure because it only looks at the extremes)

- Another measure is the **interquartile range**, where we need to measure the **first quartile** (25th percentile), and the **third quartile** (75th percentile)

Q_1 Q_3

$$\boxed{\text{interquartile range} = Q_3 - Q_1}$$

Example

9 scores on a homework (maximum score is 50):

38, 45, 35, 47, 42, 41, 50, 42, 39

Example

9 scores on a homework (maximum score is 50):

38, 45, 35, 47, 42, 41, 50, 42, 39

median : $35, \ 38, \ 39, \ 41, \ 42, \ 42, \ 45, \ 47, \ 50$

$\tilde{x}$

Example

9 scores on a homework (maximum score is 50):

38, 45, 35, 47, 42, 41, 50, 42, 39

median : 35, 38, 39, 41, 42, 42, 45, 47, 50

$\tilde{x}$

Example

9 scores on a homework (maximum score is 50):

38, 45, 35, 47, 42, 41, 50, 42, 39

median : 35, 38, 39, 41, 42, 42, 45, 47, 50

the median of this is Q_1

$\tilde{x}$

Example

9 scores on a homework (maximum score is 50):

38, 45, 35, 47, 42, 41, 50, 42, 39

median : 35, 38, 39, 41, 42, 42, 45, 47, 50

the median of this is Q_1

$\tilde{x}$

$$Q_1 = 39$$

Example

9 scores on a homework (maximum score is 50):

38, 45, 35, 47, 42, 41, 50, 42, 39

median : 35, 38, 39, 41, 42, 42, 45, 47, 50

the median of this is Q_1

$\tilde{x}$

$$Q_1 = 39$$

Example

9 scores on a homework (maximum score is 50):

38, 45, 35, 47, 42, 41, 50, 42, 39

median : 35, 38, 39, 41, 42, 42, 45, 47, 50

the median of this is Q_1

$\tilde{x}$

the median of this is Q_3

$Q_1 = 39$

Example

9 scores on a homework (maximum score is 50):

38, 45, 35, 47, 42, 41, 50, 42, 39

median : 35, 38, 39, 41, 42, 42, 45, 47, 50

the median of this is Q_1 $\quad$ $\tilde{x}$ $\quad$ the median of this is Q_3

$Q_1 = 39$ $\qquad$ $Q_3 = 45$

Example

9 scores on a homework (maximum score is 50):

38, 45, 35, 47, 42, 41, 50, 42, 39

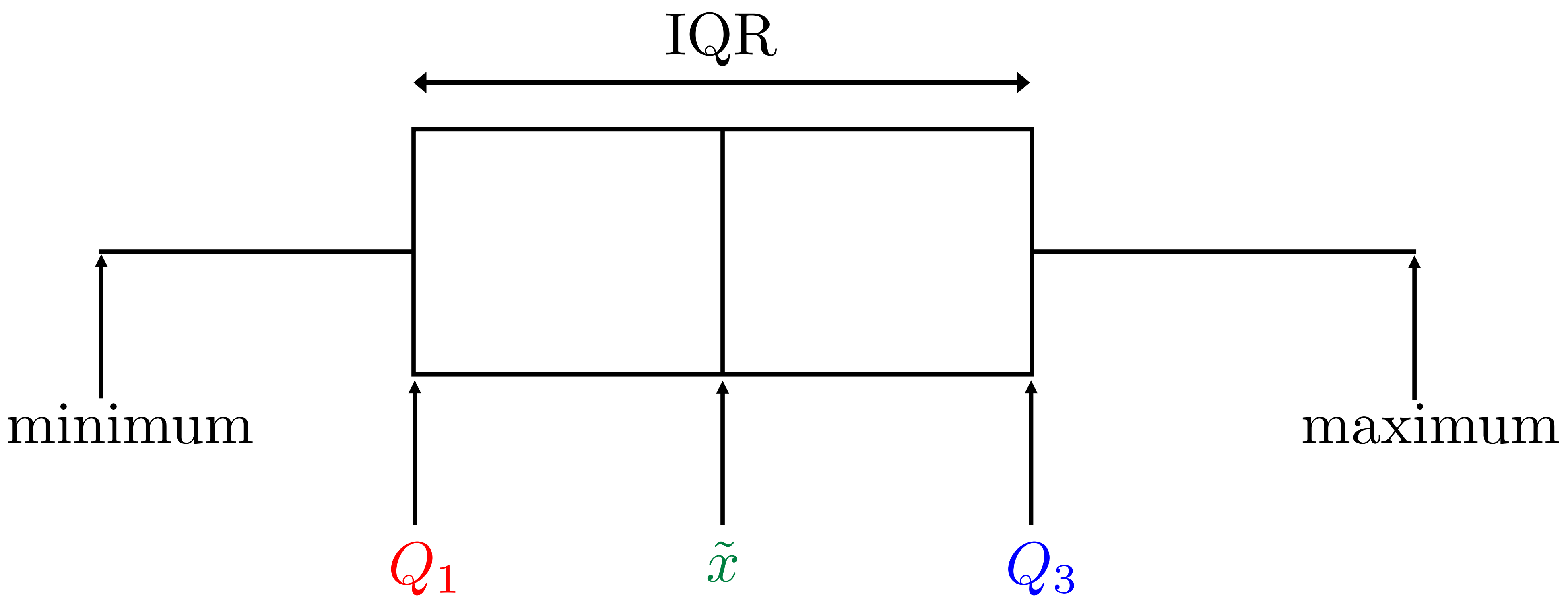
median : 35, 38, 39, 41, 42, 42, 45, 47, 50

the median of this is Q_1 $\tilde{x}$ the median of this is Q_3

$Q_1 = 39$ $Q_3 = 45$

$\text{interquartile range} = Q_3 - Q_1 = 6$

Boxplots



Any observation farther than 1.5 IQR from the closest fourth is an **outlier.**

An outlier is **extreme** if it is more than 3 IQR from the closest fourth, and it is **mild** otherwise.

The Standard Deviation

When you buy stocks or mutual funds, you need to be aware of how to quantify and balance mean gain with the variability or risk of the investment, especially given the volatile years the market has experienced in the past decade. Consider the PIMCO Total Return A (symbol: PTTAX), a fund that invests in intermediate-term fixed-income securities.

Here are its annual total returns for a recent 10-year period:

Calendar Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Return (in percent)	11.56	8.99	9.69	5.07	4.65	2.41	3.51	8.57	4.32	13.33

The Standard Deviation

Calendar Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Return (in percent)	11.56	8.99	9.69	5.07	4.65	2.41	3.51	8.57	4.32	13.33

The Standard Deviation

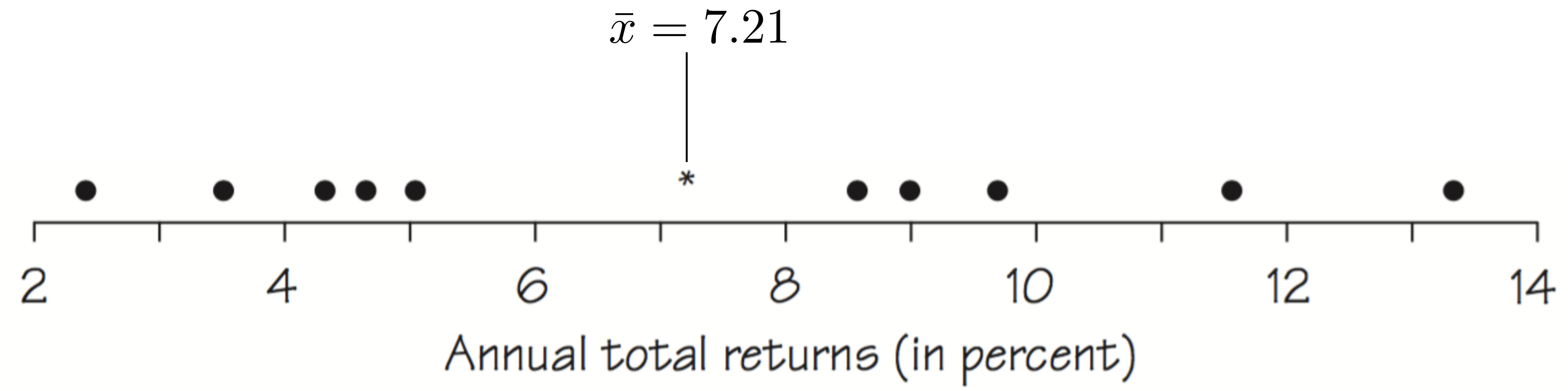
Calendar Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Return (in percent)	11.56	8.99	9.69	5.07	4.65	2.41	3.51	8.57	4.32	13.33

$$\bar{x} = \frac{11.56 + 8.99 + 9.69 + 5.07 + 4.65 + 2.41 + 3.51 + 8.57 + 4.32 + 13.33}{10} = 7.21\%$$

The Standard Deviation

Calendar Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Return (in percent)	11.56	8.99	9.69	5.07	4.65	2.41	3.51	8.57	4.32	13.33

$$\bar{x} = \frac{11.56 + 8.99 + 9.69 + 5.07 + 4.65 + 2.41 + 3.51 + 8.57 + 4.32 + 13.33}{10} = 7.21\%$$

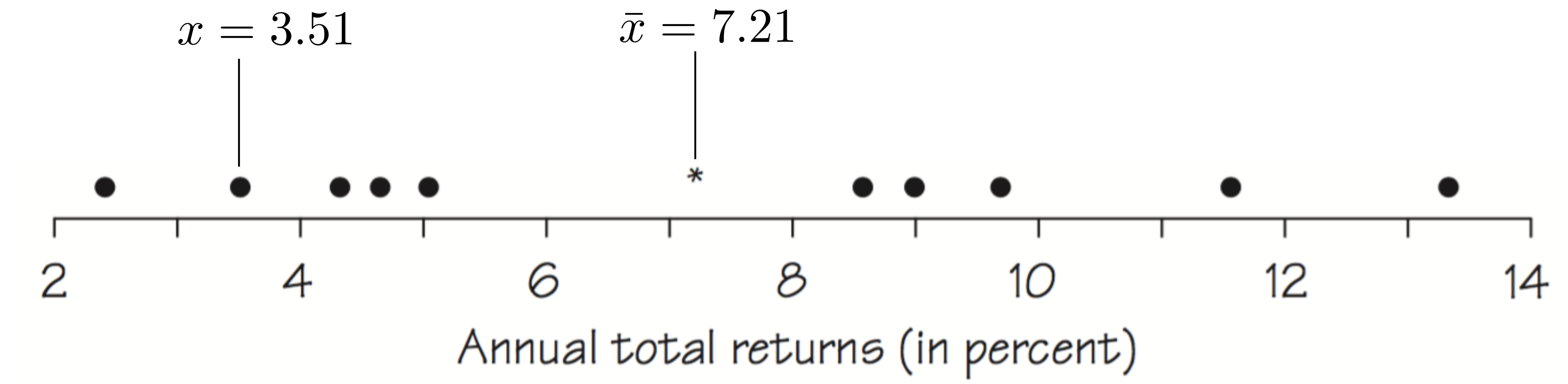


Garfunkel, For All Practical Purposes (9th edition)

The Standard Deviation

Calendar Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Return (in percent)	11.56	8.99	9.69	5.07	4.65	2.41	3.51	8.57	4.32	13.33

$$\bar{x} = \frac{11.56 + 8.99 + 9.69 + 5.07 + 4.65 + 2.41 + 3.51 + 8.57 + 4.32 + 13.33}{10} = 7.21\%$$

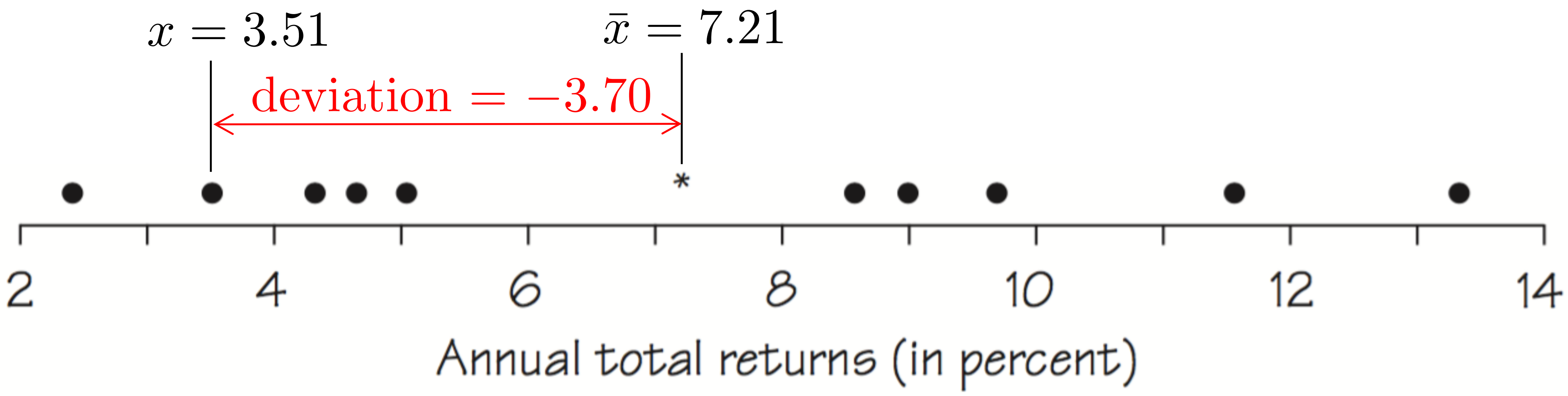


Garfunkel, For All Practical Purposes (9th edition)

The Standard Deviation

Calendar Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Return (in percent)	11.56	8.99	9.69	5.07	4.65	2.41	3.51	8.57	4.32	13.33

$$\bar{x} = \frac{11.56 + 8.99 + 9.69 + 5.07 + 4.65 + 2.41 + 3.51 + 8.57 + 4.32 + 13.33}{10} = 7.21\%$$

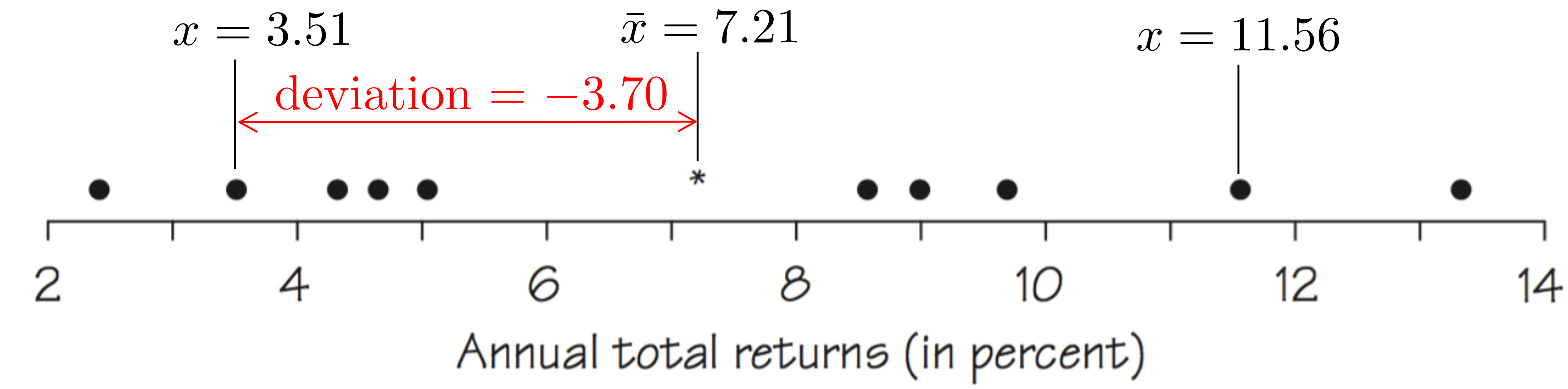


Garfunkel, For All Practical Purposes (9th edition)

The Standard Deviation

Calendar Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Return (in percent)	11.56	8.99	9.69	5.07	4.65	2.41	3.51	8.57	4.32	13.33

$$\bar{x} = \frac{11.56 + 8.99 + 9.69 + 5.07 + 4.65 + 2.41 + 3.51 + 8.57 + 4.32 + 13.33}{10} = 7.21\%$$

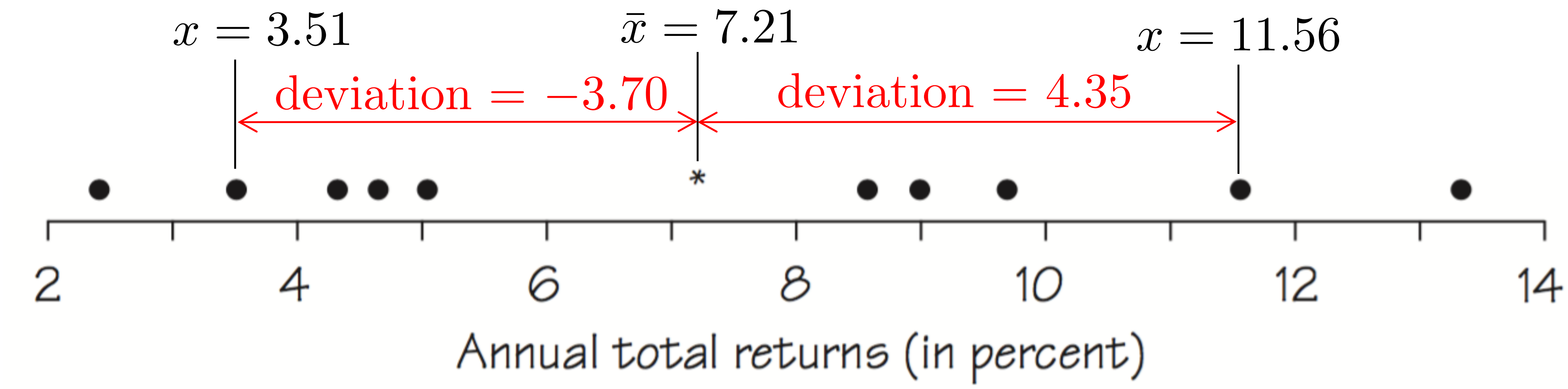


Garfunkel, For All Practical Purposes (9th edition)

The Standard Deviation

Calendar Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Return (in percent)	11.56	8.99	9.69	5.07	4.65	2.41	3.51	8.57	4.32	13.33

$$\bar{x} = \frac{11.56 + 8.99 + 9.69 + 5.07 + 4.65 + 2.41 + 3.51 + 8.57 + 4.32 + 13.33}{10} = 7.21\%$$



Garfunkel, For All Practical Purposes (9th edition)

The Standard Deviation

- We won’t get a useful measure of variability by totaling up all the positive and negative deviations from the mean because they will always sum to zero!

The Standard Deviation

- We won't get a useful measure of variability by totaling up all the positive and negative deviations from the mean because they will always sum to zero!
- Squaring the deviations makes these numbers all positive. Also, squaring weighs large deviations heavily (which is one reason why we don't take absolute value).

The Standard Deviation

- We won't get a useful measure of variability by totaling up all the positive and negative deviations from the mean because they will always sum to zero!
- Squaring the deviations makes these numbers all positive. Also, squaring weighs large deviations heavily (which is one reason why we don't take absolute value).
- A reasonable measure of variability is the average of the squared deviations.

The Standard Deviation

- We won't get a useful measure of variability by totaling up all the positive and negative deviations from the mean because they will always sum to zero!
- Squaring the deviations makes these numbers all positive. Also, squaring weighs large deviations heavily (which is one reason why we don't take absolute value).
- A reasonable measure of variability is the average of the squared deviations.
- The average of the squared deviations is called the **variance**.

The Standard Deviation

- We won't get a useful measure of variability by totaling up all the positive and negative deviations from the mean because they will always sum to zero!
- Squaring the deviations makes these numbers all positive. Also, squaring weighs large deviations heavily (which is one reason why we don't take absolute value).
- A reasonable measure of variability is the average of the squared deviations.
- The average of the squared deviations is called the **variance**.
- We denote the variance of a population by $\boldsymbol{\sigma}^2$, and the variance of a sample by $\boldsymbol{s}^2$.

The Standard Deviation

- We won't get a useful measure of variability by totaling up all the positive and negative deviations from the mean because they will always sum to zero!
- Squaring the deviations makes these numbers all positive. Also, squaring weighs large deviations heavily (which is one reason why we don't take absolute value).
- A reasonable measure of variability is the average of the squared deviations.
- The average of the squared deviations is called the **variance**.
- We denote the variance of a population by $\boldsymbol{\sigma}^2$, and the variance of a sample by $\boldsymbol{s}^2$.

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

The Standard Deviation

- We won't get a useful measure of variability by totaling up all the positive and negative deviations from the mean because they will always sum to zero!
- Squaring the deviations makes these numbers all positive. Also, squaring weighs large deviations heavily (which is one reason why we don't take absolute value).
- A reasonable measure of variability is the average of the squared deviations.
- The average of the squared deviations is called the **variance**.
- We denote the variance of a population by $\boldsymbol{\sigma}^2$, and the variance of a sample by $\boldsymbol{s}^2$.

sum

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

The Standard Deviation

- We won't get a useful measure of variability by totaling up all the positive and negative deviations from the mean because they will always sum to zero!
- Squaring the deviations makes these numbers all positive. Also, squaring weighs large deviations heavily (which is one reason why we don't take absolute value).
- A reasonable measure of variability is the average of the squared deviations.
- The average of the squared deviations is called the **variance**.
- We denote the variance of a population by $\boldsymbol{\sigma}^2$, and the variance of a sample by $\boldsymbol{s}^2$.

sum observation

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

The Standard Deviation

- We won't get a useful measure of variability by totaling up all the positive and negative deviations from the mean because they will always sum to zero!
- Squaring the deviations makes these numbers all positive. Also, squaring weighs large deviations heavily (which is one reason why we don't take absolute value).
- A reasonable measure of variability is the average of the squared deviations.
- The average of the squared deviations is called the **variance**.
- We denote the variance of a population by $\boldsymbol{\sigma}^2$, and the variance of a sample by $\boldsymbol{s}^2$.

sum observation

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

mean

The Standard Deviation

- We won't get a useful measure of variability by totaling up all the positive and negative deviations from the mean because they will always sum to zero!
- Squaring the deviations makes these numbers all positive. Also, squaring weighs large deviations heavily (which is one reason why we don't take absolute value).
- A reasonable measure of variability is the average of the squared deviations.
- The average of the squared deviations is called the **variance**.
- We denote the variance of a population by $\boldsymbol{\sigma}^2$, and the variance of a sample by $\boldsymbol{s}^2$.

sum observation

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

mean

total number of observations

The Standard Deviation

- We won't get a useful measure of variability by totaling up all the positive and negative deviations from the mean because they will always sum to zero!
- Squaring the deviations makes these numbers all positive. Also, squaring weighs large deviations heavily (which is one reason why we don't take absolute value).
- A reasonable measure of variability is the average of the squared deviations.
- The average of the squared deviations is called the **variance**.
- We denote the variance of a population by $\boldsymbol{\sigma}^2$, and the variance of a sample by $\boldsymbol{s}^2$.

sum observation

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

mean

total number of observations

Units of variance are squared of the units in the dataset, so not very useful. A more useful measure is the **standard deviation** which is simply the square root of the variance

Example

The **May 1, 2009, issue of** ***The Montclarian*** reported the following home sale amounts for a sample of homes in Alameda, CA that were sold the previous month (1000s of dollars):

590, 815, 575, 608, 350, 1285, 408, 540, 555, 679

Example

The **May 1, 2009, issue of *The Montclarian*** reported the following home sale amounts for a sample of homes in Alameda, CA that were sold the previous month (1000s of dollars):

590, 815, 575, 608, 350, 1285, 408, 540, 555, 679

$$\bar{x} = \frac{590 + 815 + 575 + 608 + 350 + 1285 + 408 + 540 + 555 + 679}{10} = 640.5$$

Example

The **May 1, 2009, issue of** ***The Montclarian*** reported the following home sale amounts for a sample of homes in Alameda, CA that were sold the previous month (1000s of dollars):

590, 815, 575, 608, 350, 1285, 408, 540, 555, 679

$$\bar{x} = \frac{590 + 815 + 575 + 608 + 350 + 1285 + 408 + 540 + 555 + 679}{10} = 640.5$$

$$s^2 = \frac{(590 - 640.5)^2 + (815 - 640.5)^2 + \cdots + (555 - 640.5)^2 + (679 - 640.5)^2}{9} \approx 67896$$

Devore, *Probability and Statistics 9th Edition* (2016)

Example

The **May 1, 2009, issue of *The Montclarian*** reported the following home sale amounts for a sample of homes in Alameda, CA that were sold the previous month (1000s of dollars):

590, 815, 575, 608, 350, 1285, 408, 540, 555, 679

$\bar{x} = 640.5$

$s^2 = 67896$

Devore, *Probability and Statistics 9th Edition* (2016)

Example

The **May 1, 2009, issue of** ***The Montclarian*** reported the following home sale amounts for a sample of homes in Alameda, CA that were sold the previous month (1000s of dollars):

590, 815, 575, 608, 350, 1285, 408, 540, 555, 679

$\bar{x} = 640.5$

$s^2 = 67896$

$$s = \sqrt{67896} = 260.6$$

Example

The **May 1, 2009, issue of *The Montclarian*** reported the following home sale amounts for a sample of homes in Alameda, CA that were sold the previous month (1000s of dollars):

590, 815, 575, 608, 350, 1285, 408, 540, 555, 679

$\bar{x} = 640.5$

$s^2 = 67896$

$$\boxed{s = \sqrt{67896} = 260.6}$$

So, a house in Alameda, CA costs on average 640.5 ± 260.6 thousand dollars