STAT 200:
Introduction to Data Science with R

Descriptive Statistics and Visualizing Distributions

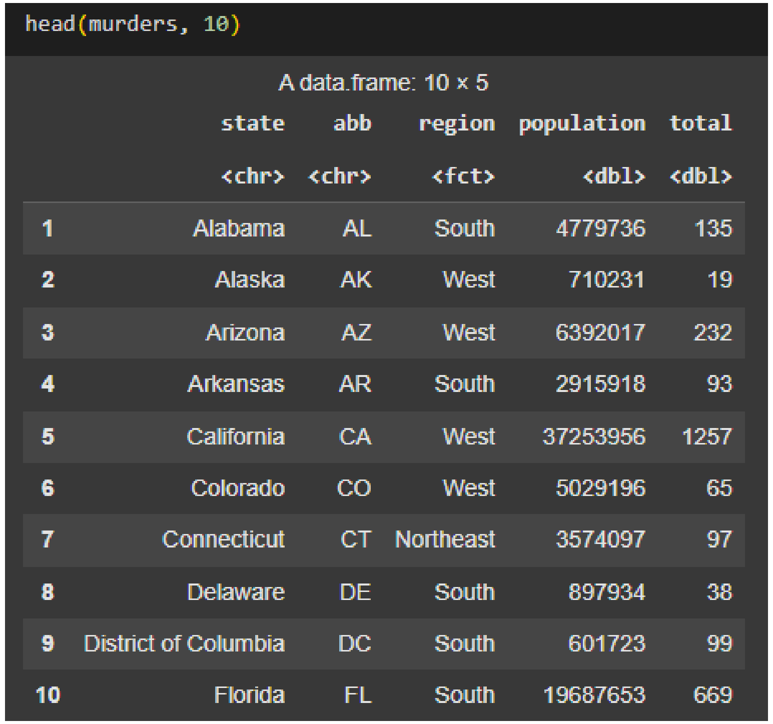
SDSU | San Diego State University

Department | Mathematics and Statistics
Division | Statistics and Data Science

Descriptive Statistics

Descriptive Statistics

- The first step in any data analysis is to gain an understanding of the data itself



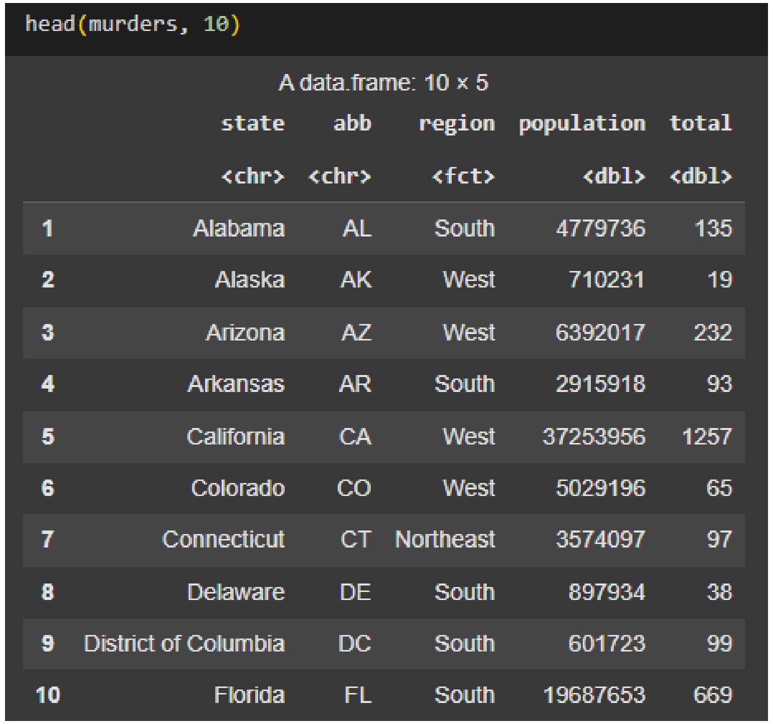
```
head(murders, 10)
```

```
                     A data.frame: 10 × 5
                   state   abb     region  population  total
                   <chr>  <chr>     <fct>       <dbl>  <dbl>
 1               Alabama    AL      South     4779736    135
 2                Alaska    AK       West      710231     19
 3               Arizona    AZ       West     6392017    232
 4              Arkansas    AR      South     2915918     93
 5            California    CA       West    37253956   1257
 6              Colorado    CO       West     5029196     65
 7           Connecticut    CT  Northeast     3574097     97
 8              Delaware    DE      South      897934     38
 9  District of Columbia    DC      South      601723     99
10               Florida    FL      South    19687653    669
```

Descriptive Statistics

- The first step in any data analysis is to gain an understanding of the data itself

- We do this by loading our data into some program (e.g., R, Excel) and exploring it various attributes
  - Number of observations
  - Number of variables
  - Identify errors in data entry
  - Identify missing values
  - etc.



```
head(murders, 10)
```

A data.frame: 10 × 5

	state	abb	region	population	total
	<chr>	<chr>	<fct>	<dbl>	<dbl>
1	Alabama	AL	South	4779736	135
2	Alaska	AK	West	710231	19
3	Arizona	AZ	West	6392017	232
4	Arkansas	AR	South	2915918	93
5	California	CA	West	37253956	1257
6	Colorado	CO	West	5029196	65
7	Connecticut	CT	Northeast	3574097	97
8	Delaware	DE	South	897934	38
9	District of Columbia	DC	South	601723	99
10	Florida	FL	South	19687653	669

Descriptive Statistics – Types of Data

- Once we have an initial understanding of our data, we can then perform basic data analysis through both
  1. Descriptive statistics
  2. Data visualizations

Descriptive Statistics – Types of Data

- Once we have an initial understanding of our data, we can then perform basic data analysis through both
    1. Descriptive statistics
    2. Data visualizations

- What are descriptive statistics?

Descriptive Statistics – Types of Data

- Once we have an initial understanding of our data, we can then perform basic data analysis through both
  1. Descriptive statistics
  2. Data visualizations

- What are descriptive statistics?
  - Measurements that summarize a given dataset

Descriptive Statistics – Types of Data

- Once we have an initial understanding of our data, we can then perform basic data analysis through both
  1. Descriptive statistics
  2. Data visualizations

- What are descriptive statistics?
  - Measurements that summarize a given dataset

- We use different descriptive statistics and data visualizations for different types of data

Descriptive Statistics – Types of Data

- Once we have an initial understanding of our data, we can then perform basic data analysis through both
  1. Descriptive statistics
  2. Data visualizations

- What are descriptive statistics?
  - Measurements that summarize a given dataset

- We use different descriptive statistics and data visualizations for different types of data
  - Numeric data

Descriptive Statistics – Types of Data

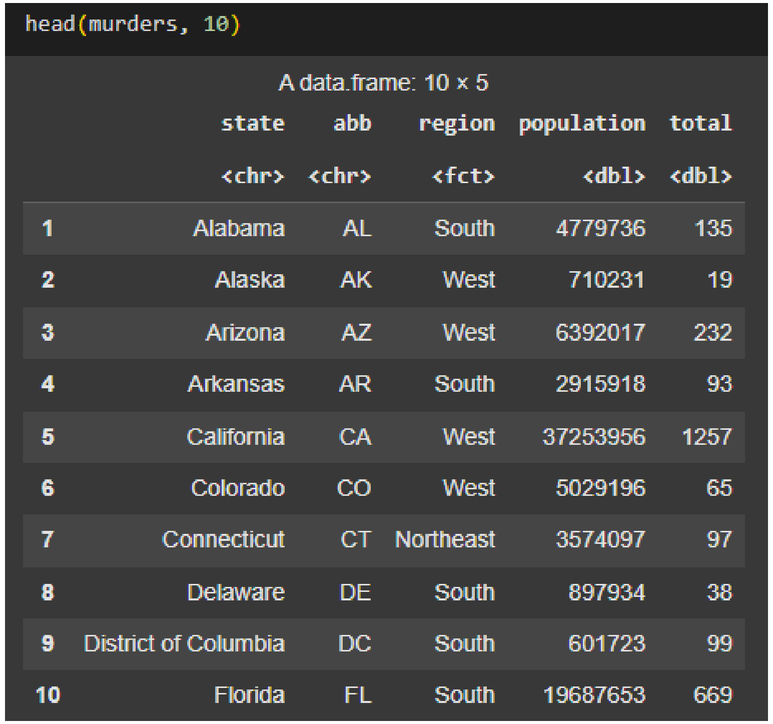
- Once we have an initial understanding of our data, we can then perform basic data analysis through both
  1. Descriptive statistics
  2. Data visualizations

- What are descriptive statistics?
  - Measurements that summarize a given dataset

- We use different descriptive statistics and data visualizations for different types of data
  - Numeric data
  - Categorical data

Descriptive Statistics – Types of Data – Numeric

What is numeric data?



```
head(murders, 10)
```

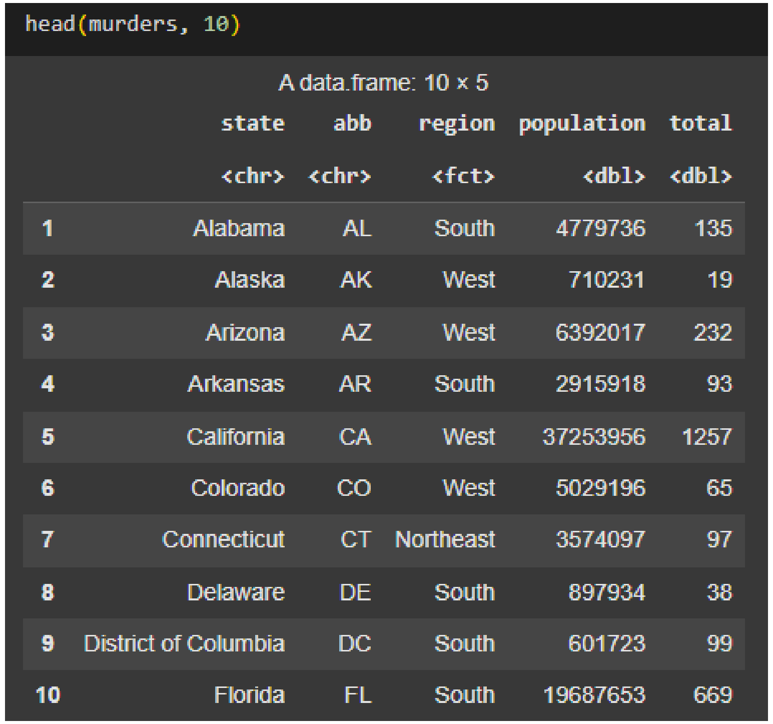
A data.frame: 10 × 5

	state	abb	region	population	total
	<chr>	<chr>	<fct>	<dbl>	<dbl>
1	Alabama	AL	South	4779736	135
2	Alaska	AK	West	710231	19
3	Arizona	AZ	West	6392017	232
4	Arkansas	AR	South	2915918	93
5	California	CA	West	37253956	1257
6	Colorado	CO	West	5029196	65
7	Connecticut	CT	Northeast	3574097	97
8	Delaware	DE	South	897934	38
9	District of Columbia	DC	South	601723	99
10	Florida	FL	South	19687653	669

Descriptive Statistics – Types of Data – Numeric

What is numeric data?

- Data consisting of numbers that are either discrete or continuous



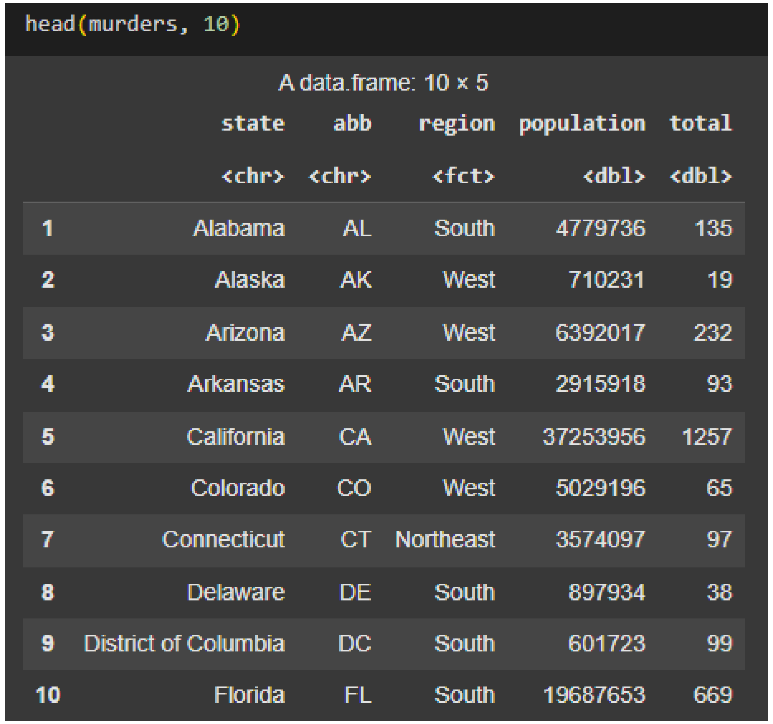
```
head(murders, 10)
```

```
                    A data.frame: 10 × 5
                  state   abb     region  population  total
                  <chr> <chr>      <fct>       <dbl>  <dbl>
 1              Alabama    AL      South     4779736    135
 2               Alaska    AK       West      710231     19
 3              Arizona    AZ       West     6392017    232
 4             Arkansas    AR      South     2915918     93
 5           California    CA       West    37253956   1257
 6             Colorado    CO       West     5029196     65
 7          Connecticut    CT  Northeast     3574097     97
 8             Delaware    DE      South      897934     38
 9 District of Columbia    DC      South      601723     99
10              Florida    FL      South    19687653    669
```

Descriptive Statistics – Types of Data – Numeric

What is numeric data?

- Data consisting of numbers that are either discrete or continuous
- Discrete data consist of numeric values that are distinct and countable, typically integer-valued data
  - e.g., population
  - e.g., number of students in a classroom



```r
head(murders, 10)
```

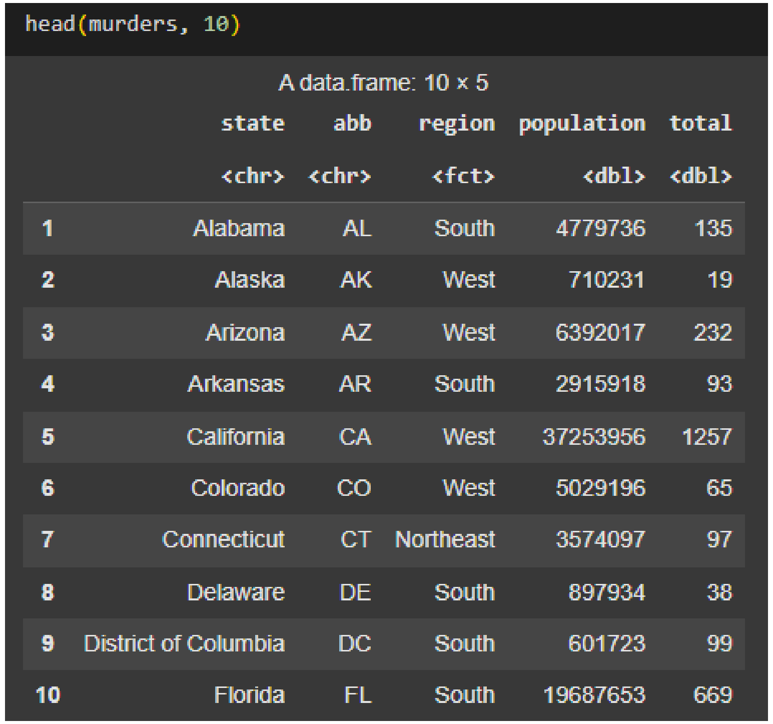
A data.frame: 10 × 5

	state	abb	region	population	total
	<chr>	<chr>	<fct>	<dbl>	<dbl>
1	Alabama	AL	South	4779736	135
2	Alaska	AK	West	710231	19
3	Arizona	AZ	West	6392017	232
4	Arkansas	AR	South	2915918	93
5	California	CA	West	37253956	1257
6	Colorado	CO	West	5029196	65
7	Connecticut	CT	Northeast	3574097	97
8	Delaware	DE	South	897934	38
9	District of Columbia	DC	South	601723	99
10	Florida	FL	South	19687653	669

Descriptive Statistics – Types of Data – Numeric

What is numeric data?

- Data consisting of numbers that are either discrete or continuous
- Discrete data consist of numeric values that are distinct and countable, typically integer-valued data
  - e.g., population
  - e.g., number of students in a classroom
- Continuous numeric data is quantitative data, typically represented by a fraction or decimal
  - e.g., temperature of 98.6 degrees Fahrenheit
  - e.g., time of 2.84 seconds

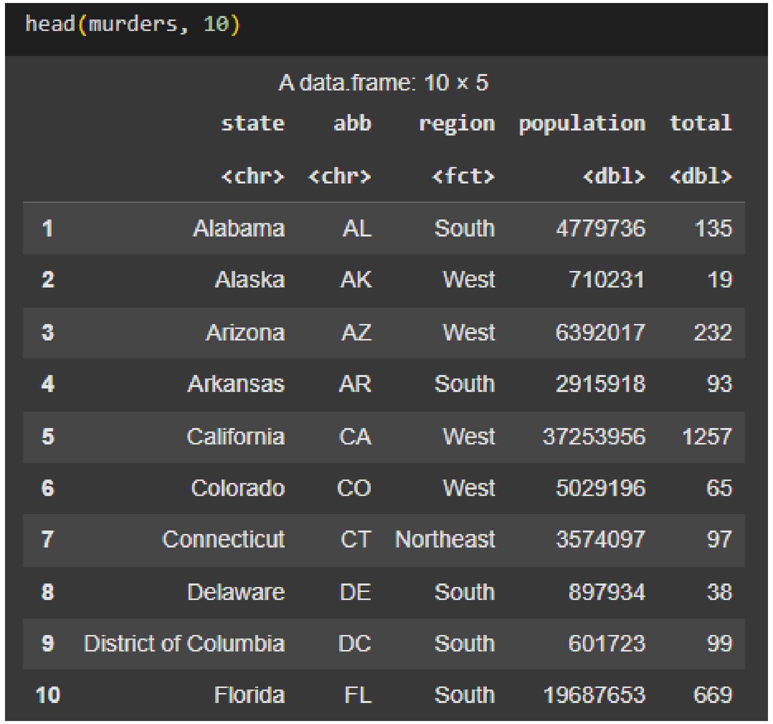


```
head(murders, 10)
```

```
                 A data.frame: 10 × 5
                   state  abb    region  population total
                   <chr> <chr>    <fct>       <dbl> <dbl>
 1               Alabama   AL     South     4779736   135
 2                Alaska   AK      West      710231    19
 3               Arizona   AZ      West     6392017   232
 4              Arkansas   AR     South     2915918    93
 5            California   CA      West    37253956  1257
 6              Colorado   CO      West     5029196    65
 7           Connecticut   CT Northeast     3574097    97
 8              Delaware   DE     South      897934    38
 9  District of Columbia   DC     South      601723    99
10               Florida   FL     South    19687653   669
```

Descriptive Statistics – Types of Data – Categorical

What is categorical data?



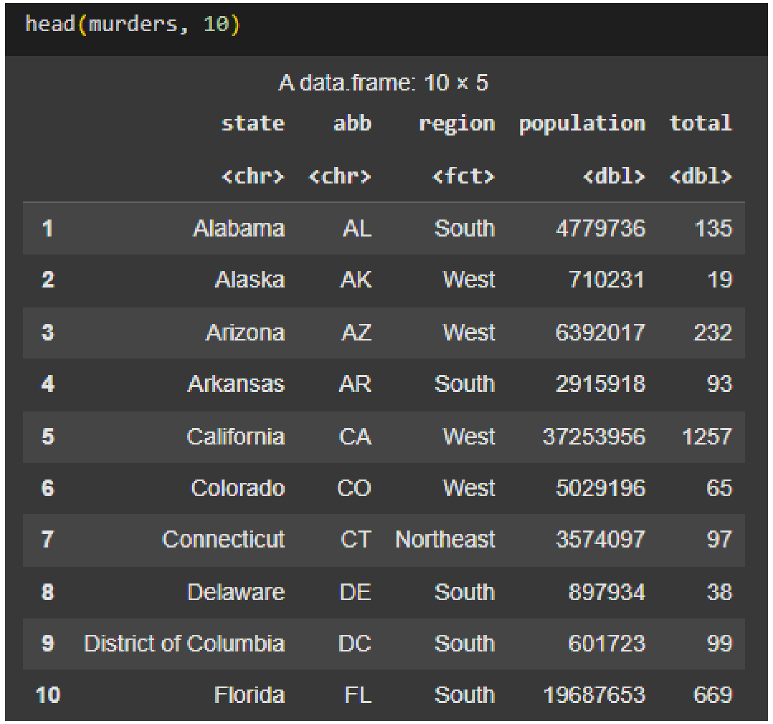
```
head(murders, 10)
```

```
                      A data.frame: 10 × 5
                  state   abb    region population total
                  <chr> <chr>     <fct>      <dbl> <dbl>
 1              Alabama    AL     South    4779736   135
 2               Alaska    AK      West     710231    19
 3              Arizona    AZ      West    6392017   232
 4             Arkansas    AR     South    2915918    93
 5           California    CA      West   37253956  1257
 6             Colorado    CO      West    5029196    65
 7          Connecticut    CT Northeast    3574097    97
 8             Delaware    DE     South     897934    38
 9 District of Columbia    DC     South     601723    99
10              Florida    FL     South   19687653   669
```

Descriptive Statistics – Types of Data – Categorical

What is categorical data?

- Data divided into distinct groups or categories



```
head(murders, 10)
```

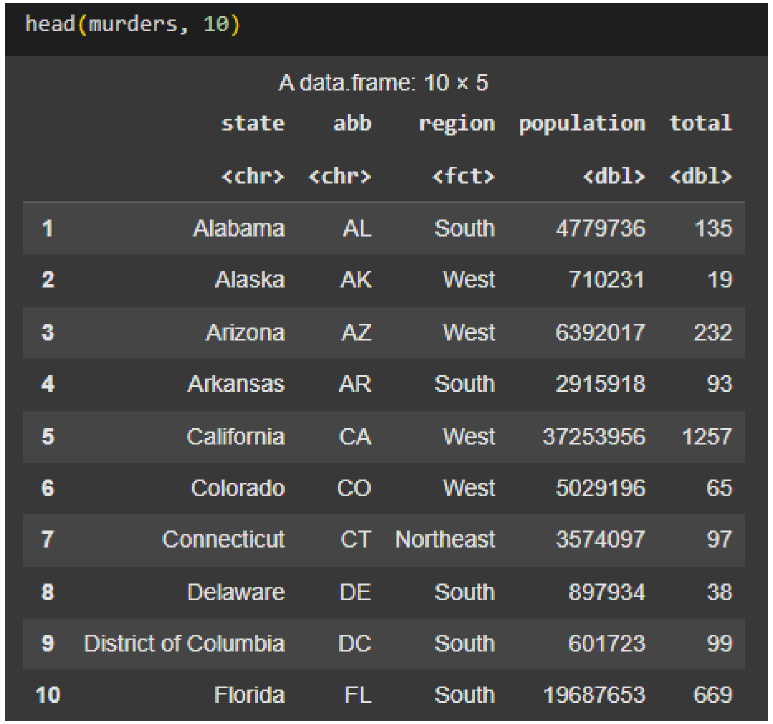
A data.frame: 10 × 5

	state	abb	region	population	total
	<chr>	<chr>	<fct>	<dbl>	<dbl>
1	Alabama	AL	South	4779736	135
2	Alaska	AK	West	710231	19
3	Arizona	AZ	West	6392017	232
4	Arkansas	AR	South	2915918	93
5	California	CA	West	37253956	1257
6	Colorado	CO	West	5029196	65
7	Connecticut	CT	Northeast	3574097	97
8	Delaware	DE	South	897934	38
9	District of Columbia	DC	South	601723	99
10	Florida	FL	South	19687653	669

Descriptive Statistics – Types of Data – Categorical

What is categorical data?

- Data divided into distinct groups or categories
- Two types of categorical data
  - Nominal
  - Ordinal



```r
head(murders, 10)
```

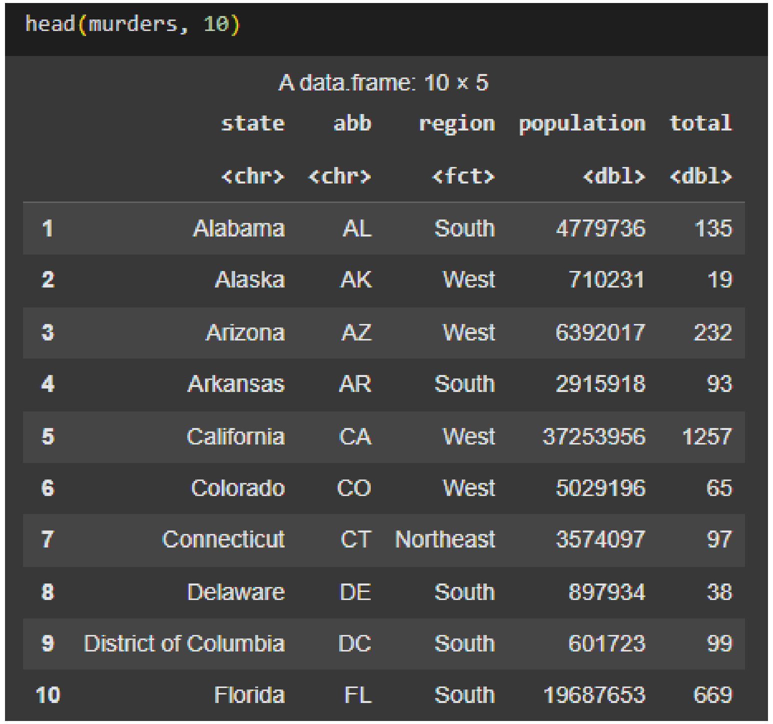
A data.frame: 10 × 5

	state	abb	region	population	total
	<chr>	<chr>	<fct>	<dbl>	<dbl>
1	Alabama	AL	South	4779736	135
2	Alaska	AK	West	710231	19
3	Arizona	AZ	West	6392017	232
4	Arkansas	AR	South	2915918	93
5	California	CA	West	37253956	1257
6	Colorado	CO	West	5029196	65
7	Connecticut	CT	Northeast	3574097	97
8	Delaware	DE	South	897934	38
9	District of Columbia	DC	South	601723	99
10	Florida	FL	South	19687653	669

Descriptive Statistics – Types of Data – Categorical

What is categorical data?

- Data divided into distinct groups or categories
- Two types of categorical data
  - Nominal
  - Ordinal
- Nominal categorical data consist of categories that have no inherent ordering or ranking
  - e.g., states, region



```
head(murders, 10)
```

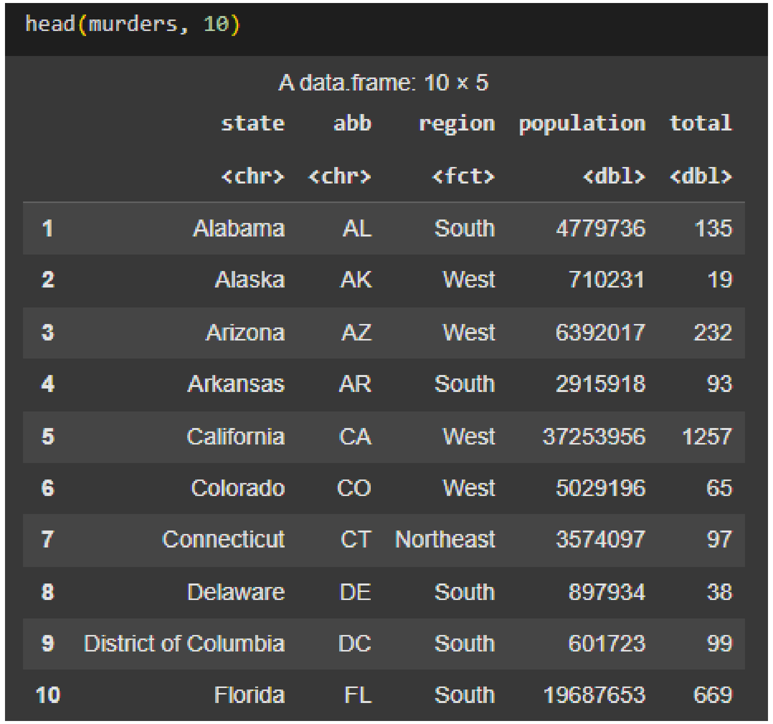
A data.frame: 10 × 5

	state	abb	region	population	total
	<chr>	<chr>	<fct>	<dbl>	<dbl>
1	Alabama	AL	South	4779736	135
2	Alaska	AK	West	710231	19
3	Arizona	AZ	West	6392017	232
4	Arkansas	AR	South	2915918	93
5	California	CA	West	37253956	1257
6	Colorado	CO	West	5029196	65
7	Connecticut	CT	Northeast	3574097	97
8	Delaware	DE	South	897934	38
9	District of Columbia	DC	South	601723	99
10	Florida	FL	South	19687653	669

Descriptive Statistics – Types of Data – Categorical

What is categorical data?

- Data divided into distinct groups or categories
- Two types of categorical data
  - Nominal
  - Ordinal
- Nominal categorical data consist of categories that have no inherent ordering or ranking
  - e.g., states, region
- Ordinal categorical data consist of categories that have inherent ordering or ranking
  - e.g., education level (high school, bachelor’s, master’s, PhD)
  - e.g., mood (1-sad, 2-neutral, 3-happy)



```
head(murders, 10)
```

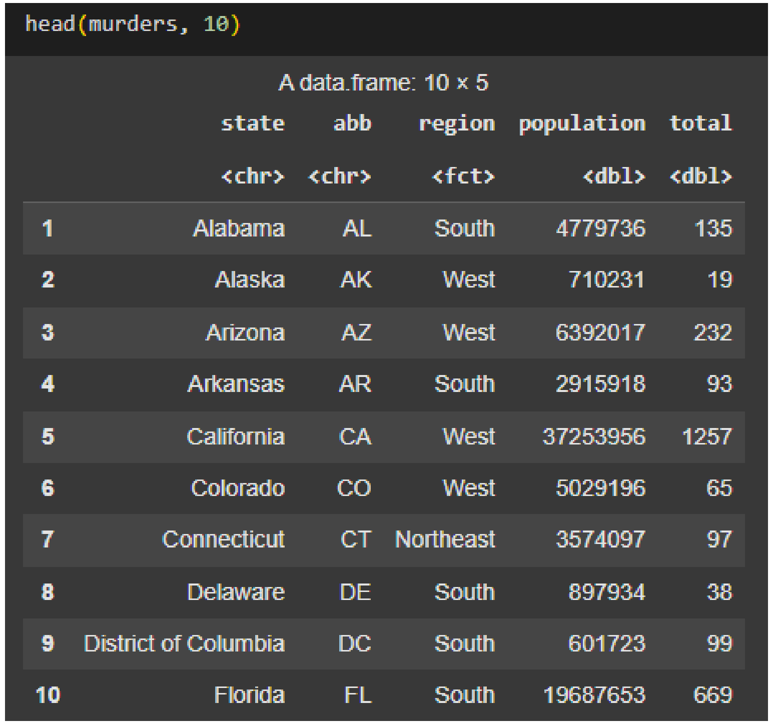
A data.frame: 10 × 5

	state	abb	region	population	total
	<chr>	<chr>	<fct>	<dbl>	<dbl>
1	Alabama	AL	South	4779736	135
2	Alaska	AK	West	710231	19
3	Arizona	AZ	West	6392017	232
4	Arkansas	AR	South	2915918	93
5	California	CA	West	37253956	1257
6	Colorado	CO	West	5029196	65
7	Connecticut	CT	Northeast	3574097	97
8	Delaware	DE	South	897934	38
9	District of Columbia	DC	South	601723	99
10	Florida	FL	South	19687653	669

Descriptive Statistics – Types of Data – Categorical

What is categorical data?

- Data divided into distinct groups or categories
- Two types of categorical data
  - Nominal
  - Ordinal
- Nominal categorical data consist of categories that have no inherent ordering or ranking
  - e.g., states, region
- Ordinal categorical data consist of categories that have inherent ordering or ranking
  - e.g., education level (high school, bachelor’s, master’s, PhD)
  - e.g., mood (1-sad, 2-neutral, 3-happy)



```
head(murders, 10)
```

```
                          A data.frame: 10 × 5
                 state   abb     region  population  total
                 <chr> <chr>      <fct>       <dbl>  <dbl>
 1             Alabama    AL      South     4779736    135
 2              Alaska    AK       West      710231     19
 3             Arizona    AZ       West     6392017    232
 4            Arkansas    AR      South     2915918     93
 5          California    CA       West    37253956   1257
 6            Colorado    CO       West     5029196     65
 7         Connecticut    CT  Northeast     3574097     97
 8            Delaware    DE      South      897934     38
 9  District of Columbia  DC      South      601723     99
10             Florida    FL      South    19687653    669
```

Note that discrete numeric data can be considered ordinal!

Descriptive Statistics – Numeric Data

- For numeric data, there are two common types of descriptive statistics
  - Measures of central tendency
  - Measures of variability (how things change or vary)

Descriptive Statistics – Numeric Data

- For numeric data, there are two common types of descriptive statistics
  - Measures of central tendency
  - Measures of variability (how things change or vary)

- Measures of central tendency
  - Mean (average)
  - Median (50th percentile, 0.50 quantile)
  - Mode (value that occurs most often)

Descriptive Statistics – Numeric Data

- For numeric data, there are two common types of descriptive statistics
  - Measures of central tendency
  - Measures of variability (how things change or vary)

- Measures of central tendency
  - Mean (average)
  - Median (50th percentile, 0.50 quantile)
  - Mode (value that occurs most often)

- Measures of variability
  - Range
  - Variance
  - Standard deviation
  - Quantiles/Percentiles

Descriptive Statistics – Numeric Data

- For numeric data, there are two common types of descriptive statistics
  - Measures of central tendency
  - Measures of variability (how things change or vary)

- Measures of central tendency
  - Mean (average)
  - Median (50th percentile, 0.50 quantile)
  - Mode (value that occurs most often)

- Measures of variability
  - Range
  - Variance
  - Standard deviation
  - Quantiles/Percentiles

<u>These were likely covered in your high school courses, but I will review them here!</u>

Descriptive Statistics:

Measures of Central Tendency

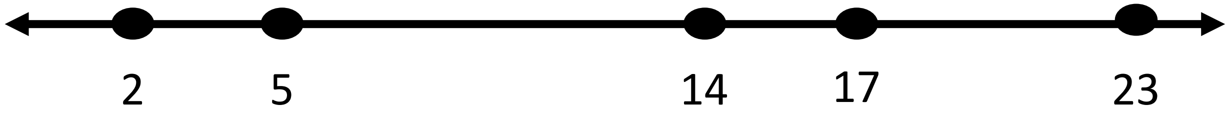
Descriptive Statistics – Numeric Data – Mean

- Consider the following data on the amount of emphysema in the lungs

Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17
6	14

Descriptive Statistics – Numeric Data – Mean

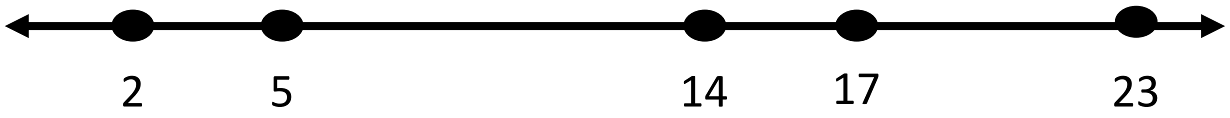
- Consider the following data on the amount of emphysema in the lungs
- We can represent the percentage emphysema on the number line



Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17
6	14

Descriptive Statistics – Numeric Data – Mean

- Consider the following data on the amount of emphysema in the lungs
- We can represent the percentage emphysema on the number line



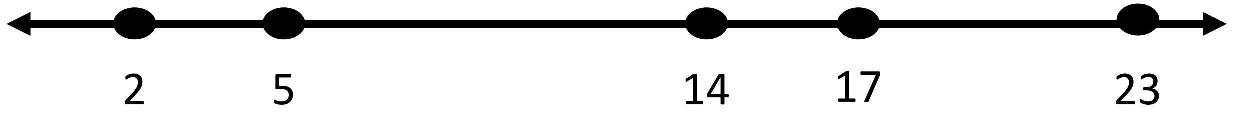
- The mean is a measure of the center of these observations

mean = sum of values / number of values

Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17
6	14

Descriptive Statistics – Numeric Data – Mean

- Consider the following data on the amount of emphysema in the lungs
- We can represent the percentage emphysema on the number line



- The mean is a measure of the center of these observations

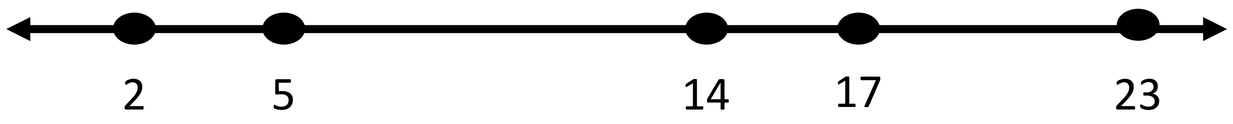
mean = sum of values / number of values

- What is the mean of the variable percentage emphysema?

Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17
6	14

Descriptive Statistics – Numeric Data – Mean

- Consider the following data on the amount of emphysema in the lungs
- We can represent the percentage emphysema on the number line



- The mean is a measure of the center of these observations

mean = sum of values / number of values

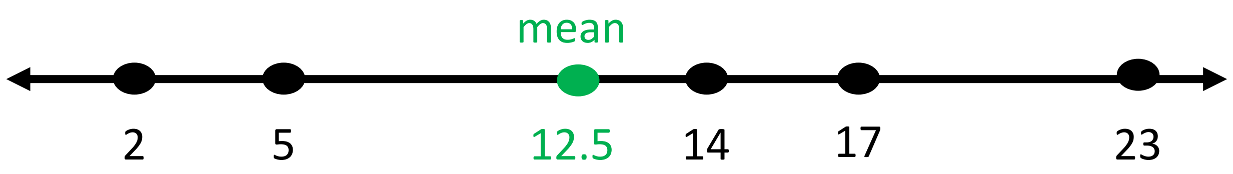
- What is the mean of the variable percentage emphysema?

(5 + 23 + 2 + 14 + 17 + 14) / 6 = 12.5%

Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17
6	14

Descriptive Statistics – Numeric Data – Mean

- Consider the following data on the amount of emphysema in the lungs
- We can represent the percentage emphysema on the number line



- The mean is a measure of the center of these observations

mean = sum of values / number of values

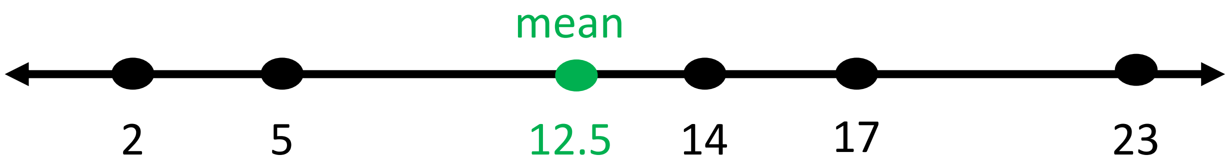
- What is the mean of the variable percentage emphysema?

(5 + 23 + 2 + 14 + 17 + 14) / 6 = 12.5%

Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17
6	14

Descriptive Statistics – Numeric Data – Median

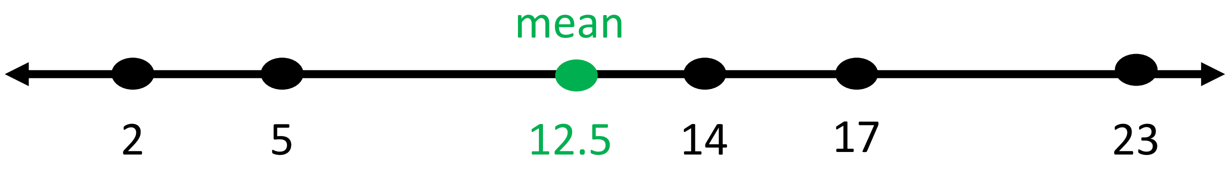
- Consider the following data on the amount of emphysema in the lungs
- We can represent the percentage emphysema on the number line



Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17
6	14

Descriptive Statistics – Numeric Data – Median

- Consider the following data on the amount of emphysema in the lungs
- We can represent the percentage emphysema on the number line

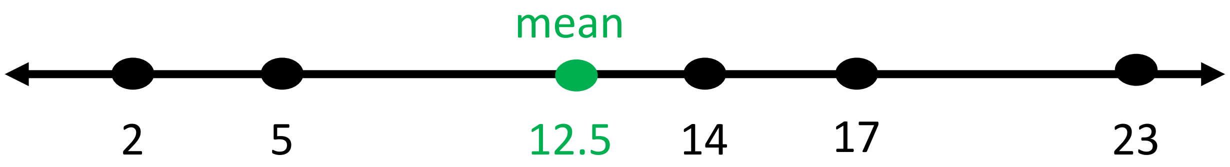


- The median is the value such that half the data is above and half the data is below

Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17
6	14

Descriptive Statistics – Numeric Data – Median

- Consider the following data on the amount of emphysema in the lungs
- We can represent the percentage emphysema on the number line

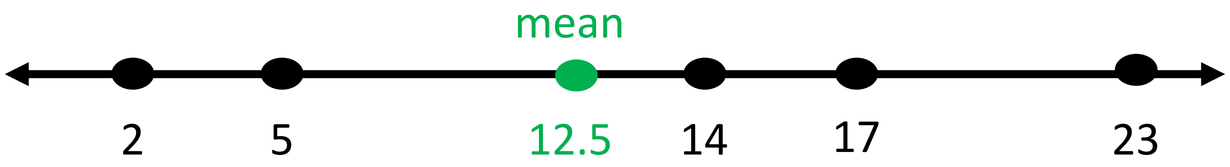


- The median is the value such that half the data is above and half the data is below
- To find the median
  1. Rank the data – 2, 5, 14, 14, 17, 23

Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17
6	14

Descriptive Statistics – Numeric Data – Median

- Consider the following data on the amount of emphysema in the lungs
- We can represent the percentage emphysema on the number line



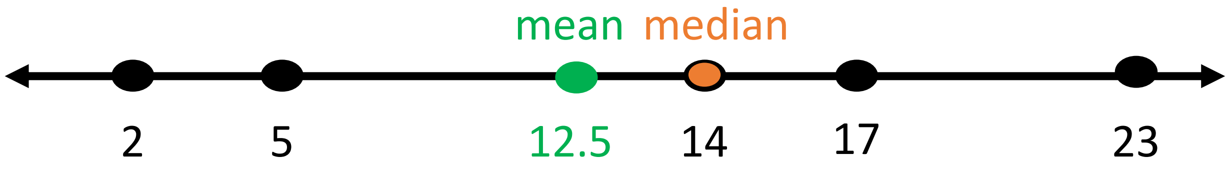
Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17
6	14

- The median is the value such that half the data is above and half the data is below

- To find the median
  1. Rank the data – 2, 5, 14, 14, 17, 23
  2. Find the “middle” number
     - If there is an even number of data points, average the two middle numbers

Descriptive Statistics – Numeric Data – Median

- Consider the following data on the amount of emphysema in the lungs
- We can represent the percentage emphysema on the number line



Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17
6	14

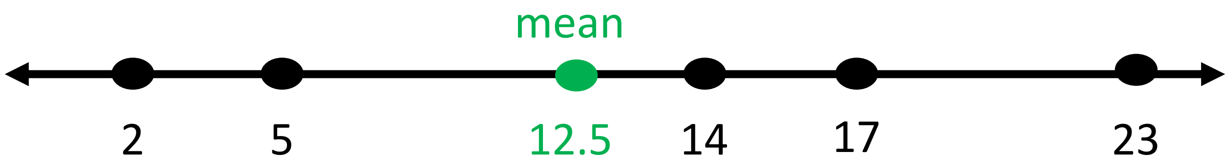
- The median is the value such that half the data is above and half the data is below

- To find the median
  1. Rank the data – 2, 5, 14, 14, 17, 23
  2. Find the "middle" number
     - If there is an even number of data points, average the two middle numbers

- The median is 14
  - Half the data points are greater than 14 and half are less than 14

Descriptive Statistics – Numeric Data – Mode

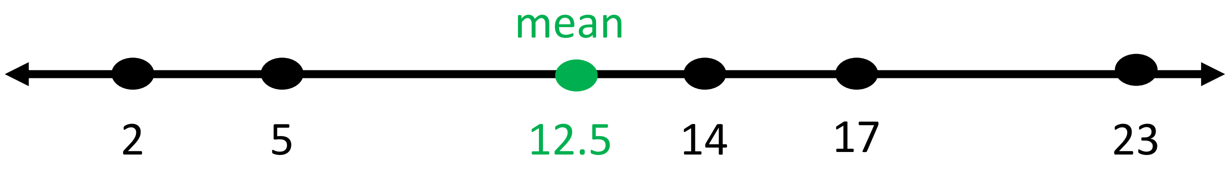
- Consider the following data on the amount of emphysema in the lungs
- We can represent the percentage emphysema on the number line



Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17
6	14

Descriptive Statistics – Numeric Data – Mode

- Consider the following data on the amount of emphysema in the lungs
- We can represent the percentage emphysema on the number line

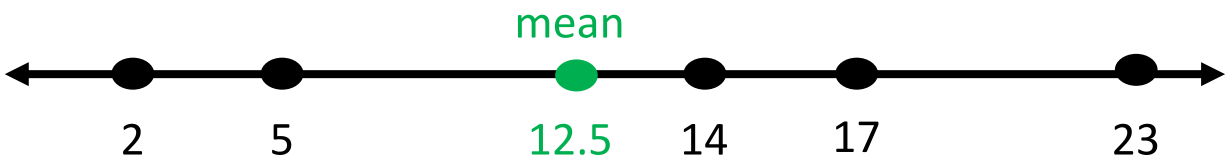


- The mode is the value that appears most often

Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17
6	14

Descriptive Statistics – Numeric Data – Mode

- Consider the following data on the amount of emphysema in the lungs
- We can represent the percentage emphysema on the number line

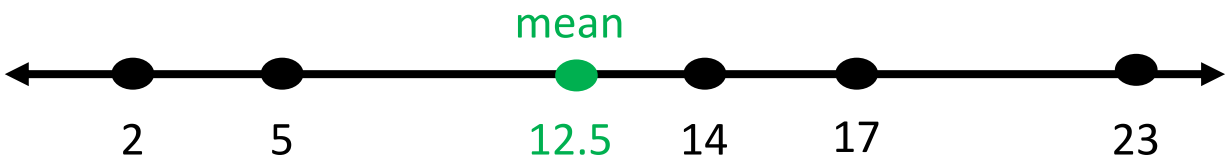


Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17
6	14

- The mode is the value that appears most often
- To find the mode
  1. Count the occurrence of each value
  2. Choose the value with the largest occurrence

Descriptive Statistics – Numeric Data – Mode

- Consider the following data on the amount of emphysema in the lungs
- We can represent the percentage emphysema on the number line



- The mode is the value that appears most often

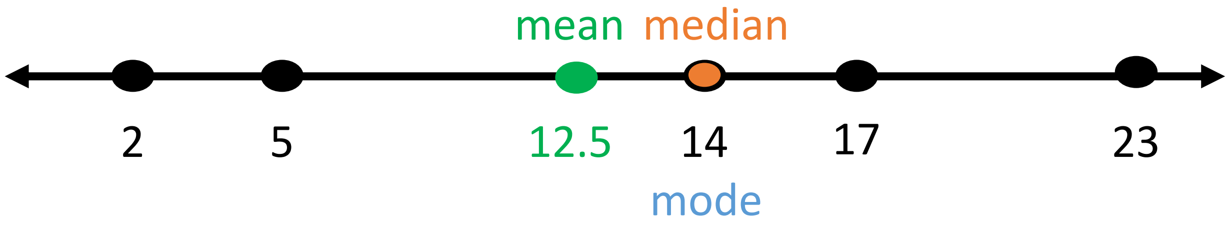
- To find the mode
  1. Count the occurrence of each value
  2. Choose the value with the largest occurrence

- What is the mode of percentage emphysema?

Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17
6	14

Descriptive Statistics – Numeric Data – Mode

- Consider the following data on the amount of emphysema in the lungs
- We can represent the percentage emphysema on the number line



- The mode is the value that appears most often

- To find the mode
  1. Count the occurrence of each value
  2. Choose the value with the largest occurrence

- What is the mode of percentage emphysema?
  - 14 since this value occurs 2 times
  - All other values occur less than 2 times

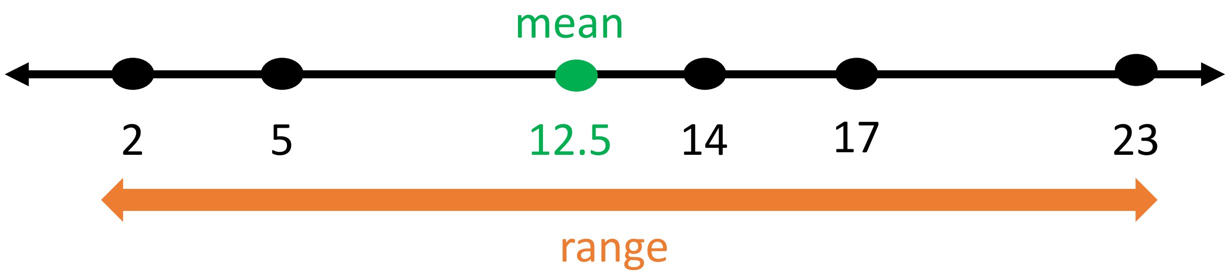
Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17
6	14

Descriptive Statistics:

Measures of Variability

Descriptive Statistics – Numeric Data – Range

- Consider the following data on the amount of emphysema in the lungs
- We can represent the percentage emphysema on the number line

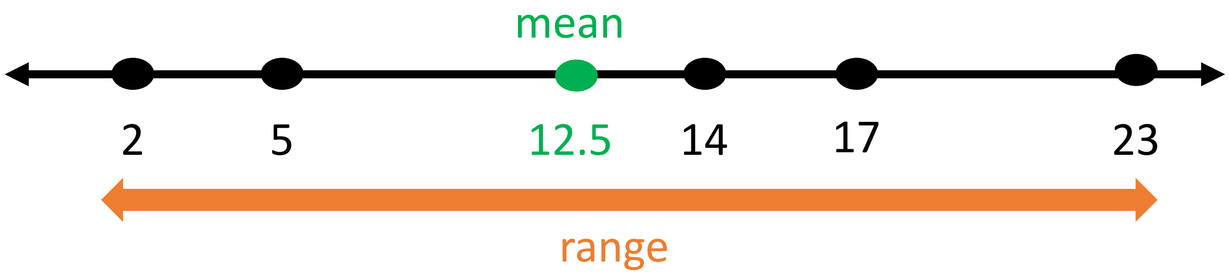


Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17
6	14

- The **range** is the maximum – minimum

Descriptive Statistics – Numeric Data – Range

- Consider the following data on the amount of emphysema in the lungs
- We can represent the percentage emphysema on the number line

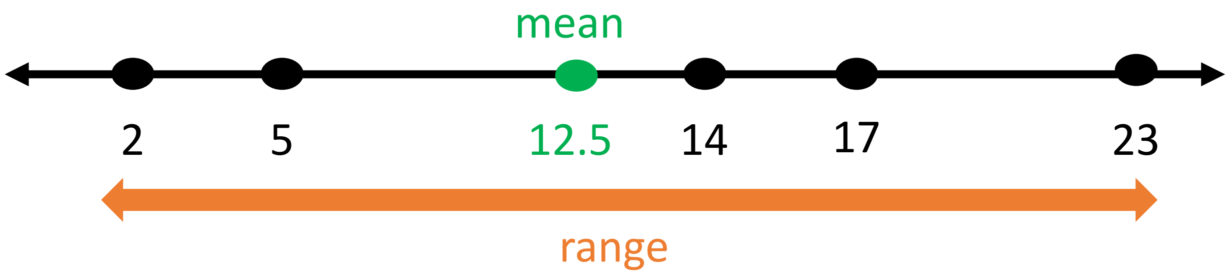


- The **range** is the maximum – minimum
- The range measures the dispersion (spread) of the data

Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17
6	14

Descriptive Statistics – Numeric Data – Range

- Consider the following data on the amount of emphysema in the lungs
- We can represent the percentage emphysema on the number line

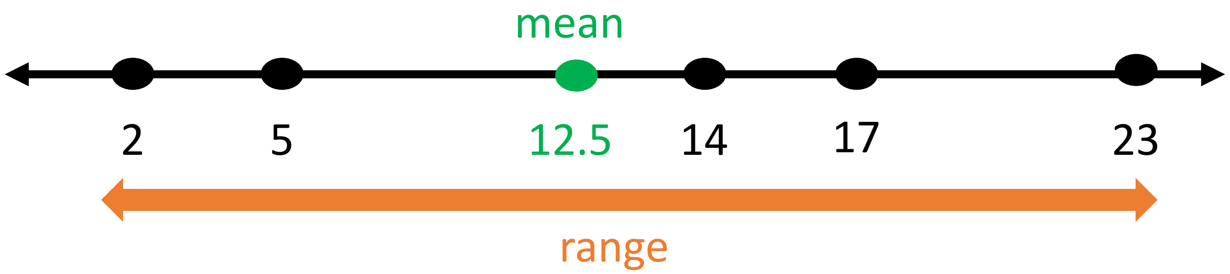


- The **range** is the maximum – minimum
- The range measures the dispersion (spread) of the data
- What is the range of percentage emphysema?

Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17
6	14

Descriptive Statistics – Numeric Data – Range

- Consider the following data on the amount of emphysema in the lungs
- We can represent the percentage emphysema on the number line



Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17
6	14

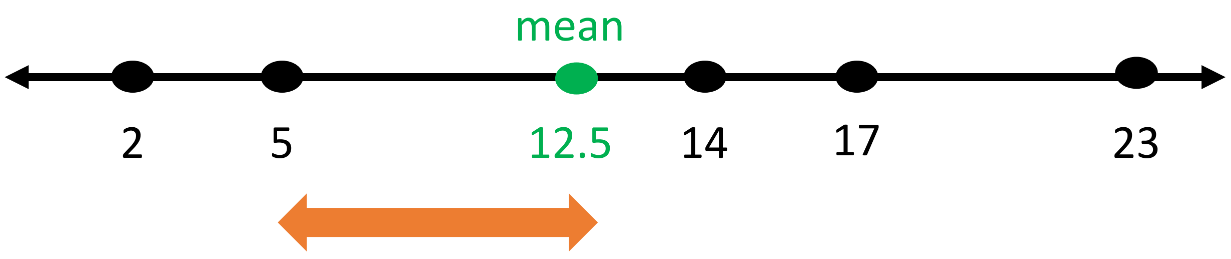
- The **range** is the maximum – minimum
- The range measures the dispersion (spread) of the data
- What is the range of percentage emphysema?

range = max – min = 23 – 2 = 21%

Descriptive Statistics – Numeric Data – Variance

- **Variance** is another measure of dispersion

Patient	%emph	%emph - mean	(%emph – mean)^2
1	5	-7.5	56.25
2	23	10.5	110.25
3	2	-10.5	110.25
4	14	1.5	2.25
5	17	4.5	20.25
6	14	1.5	2.25
		Total:	301.5
		Total/(6-1):	60.3

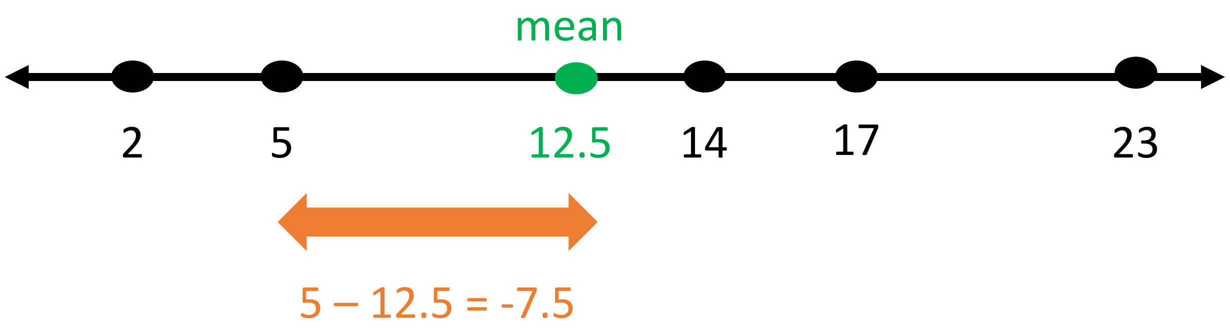


5 – 12.5 = -7.5

Descriptive Statistics – Numeric Data – Variance

- **Variance** is another measure of dispersion
- Unlike range, variance is calculated from all data points (more informative)

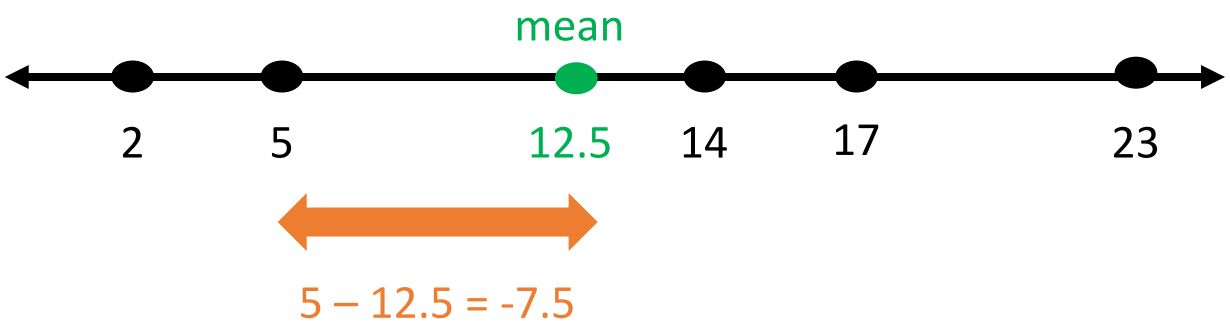
Patient	%emph	%emph - mean	(%emph – mean)^2
1	5	-7.5	56.25
2	23	10.5	110.25
3	2	-10.5	110.25
4	14	1.5	2.25
5	17	4.5	20.25
6	14	1.5	2.25
		Total:	301.5
		Total/(6-1):	60.3



Descriptive Statistics – Numeric Data – Variance

- **Variance** is another measure of dispersion
- Unlike range, variance is calculated from all data points (more informative)
- Variance is the average sum of squared distances from the mean

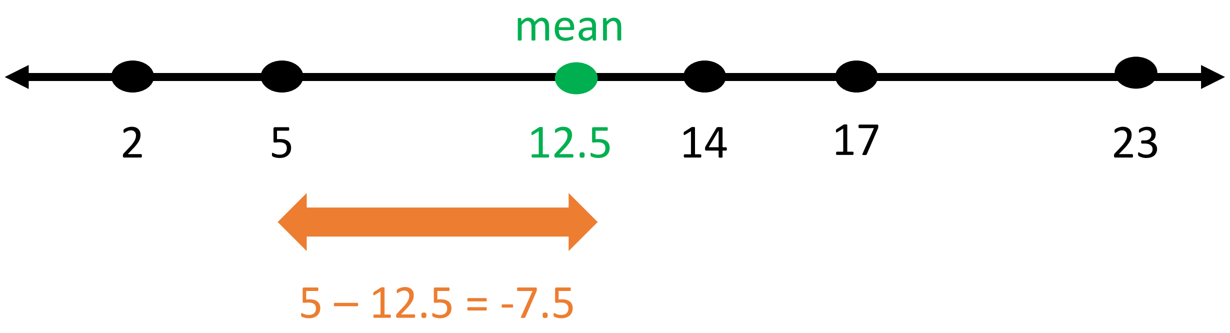
Patient	%emph	%emph - mean	(%emph – mean)^2
1	5	-7.5	56.25
2	23	10.5	110.25
3	2	-10.5	110.25
4	14	1.5	2.25
5	17	4.5	20.25
6	14	1.5	2.25
		Total:	301.5
		Total/(6-1):	60.3



Descriptive Statistics – Numeric Data – Variance

- **Variance** is another measure of dispersion
- Unlike range, variance is calculated from all data points (more informative)
- Variance is the average sum of squared distances from the mean
  1. Subtract the mean from each point

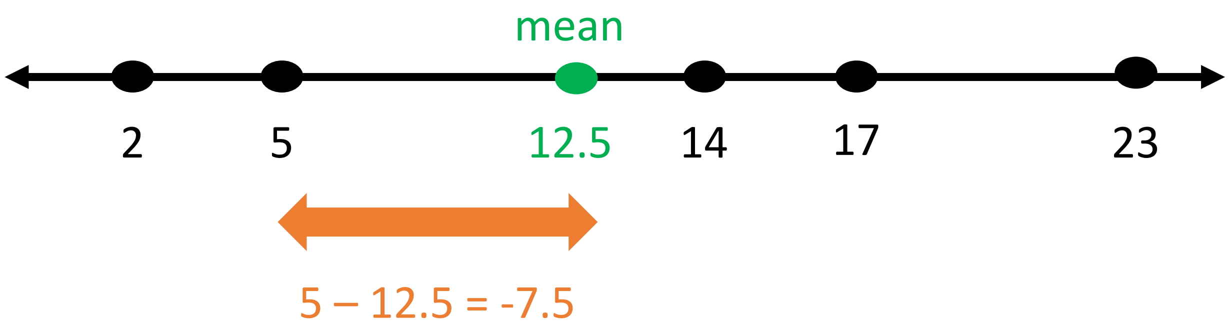
Patient	%emph	%emph - mean	(%emph – mean)^2
1	5	-7.5	56.25
2	23	10.5	110.25
3	2	-10.5	110.25
4	14	1.5	2.25
5	17	4.5	20.25
6	14	1.5	2.25
		Total:	301.5
		Total/(6-1):	60.3



Descriptive Statistics – Numeric Data – Variance

- **<u>Variance</u>** is another measure of dispersion
- Unlike range, variance is calculated from all data points (more informative)
- Variance is the average sum of squared distances from the mean
  1. Subtract the mean from each point
  2. Square each difference

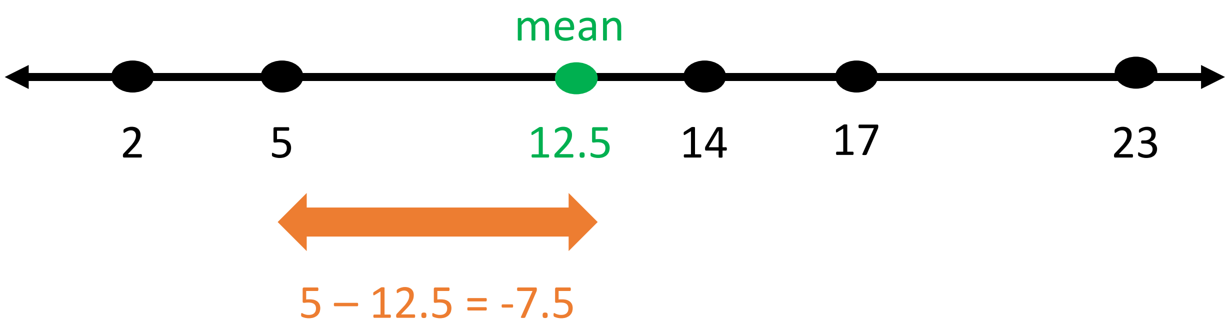
Patient	%emph	%emph - mean	(%emph – mean)^2
1	5	-7.5	56.25
2	23	10.5	110.25
3	2	-10.5	110.25
4	14	1.5	2.25
5	17	4.5	20.25
6	14	1.5	2.25
		Total:	301.5
		Total/(6-1):	60.3



Descriptive Statistics – Numeric Data – Variance

- **<u>Variance</u>** is another measure of dispersion
- Unlike range, variance is calculated from all data points (more informative)
- Variance is the average sum of squared distances from the mean
  1. Subtract the mean from each point
  2. Square each difference
  3. Add the squared differences

Patient	%emph	%emph - mean	(%emph – mean)^2
1	5	-7.5	56.25
2	23	10.5	110.25
3	2	-10.5	110.25
4	14	1.5	2.25
5	17	4.5	20.25
6	14	1.5	2.25
		Total:	301.5
		Total/(6-1):	60.3



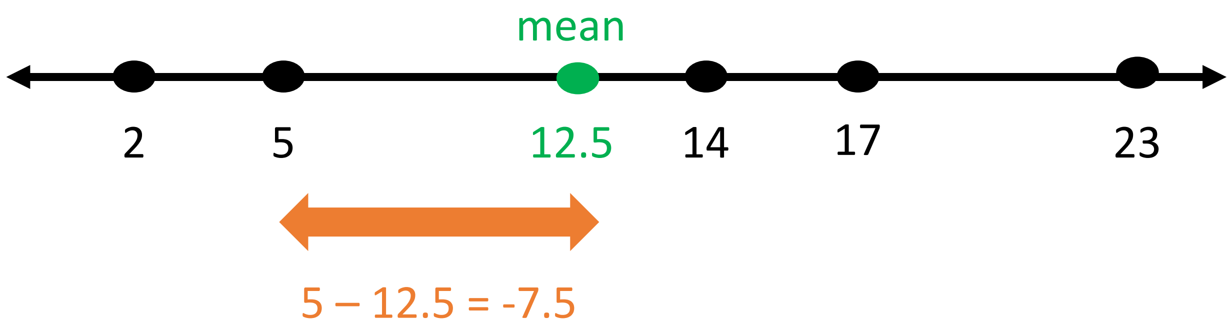
Descriptive Statistics – Numeric Data – Variance

- **Variance** is another measure of dispersion

- Unlike range, variance is calculated from all data points (more informative)

- Variance is the average sum of squared distances from the mean
  1. Subtract the mean from each point
  2. Square each difference
  3. Add the squared differences
  4. Divide the sum by the number of points minus one

Patient	%emph	%emph - mean	(%emph – mean)^2
1	5	-7.5	56.25
2	23	10.5	110.25
3	2	-10.5	110.25
4	14	1.5	2.25
5	17	4.5	20.25
6	14	1.5	2.25
		Total:	301.5
		Total/(6-1):	60.3

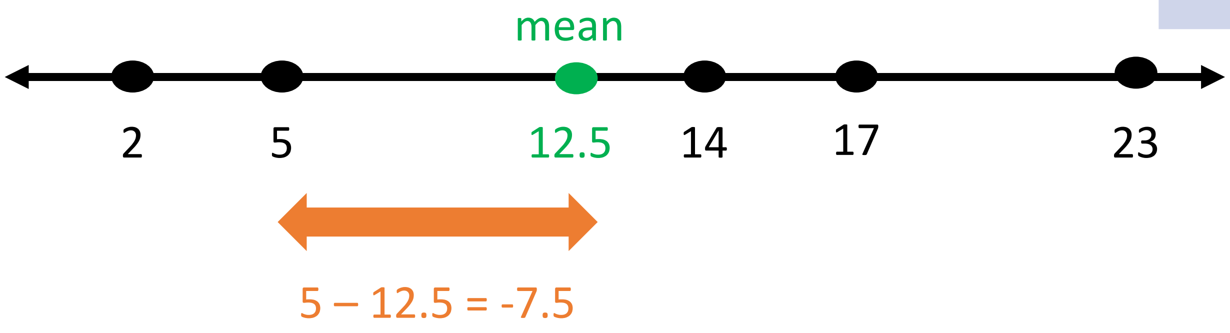


Descriptive Statistics – Numeric Data – Standard Deviation

- **Standard deviation** is the square root of the variance
- Since variance squares the units (%^2 in this case), we take square root to convert back to original units (%)

$$SD = \sqrt{Var} = \sqrt{60.3} = 7.77\%$$

Patient	%emph	%emph - mean	(%emph – mean)^2
1	5	-7.5	56.25
2	23	10.5	110.25
3	2	-10.5	110.25
4	14	1.5	2.25
5	17	4.5	20.25
6	14	1.5	2.25
		Total:	301.5
		Total/(6-1):	60.3
		$SD = \sqrt{60.3}$:	7.77



Descriptive Statistics – Numeric Data

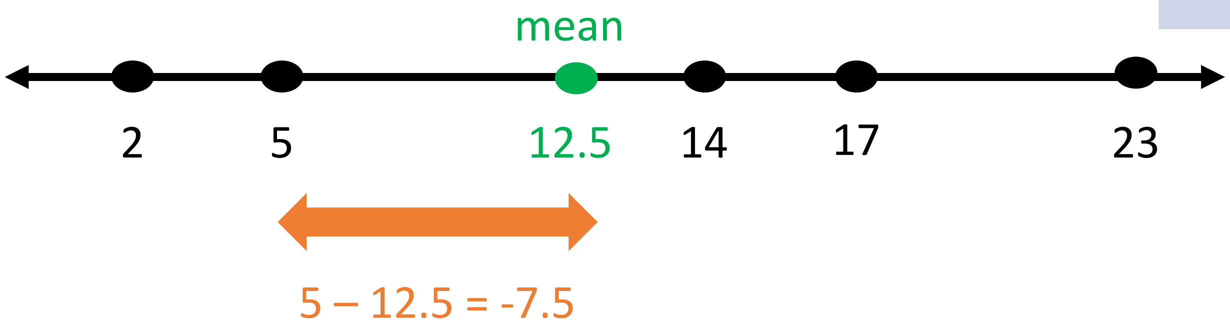
What happens to the variance when points are further from the mean?

A) Variance is larger

B) Variance is smaller

C) Variance stays the same

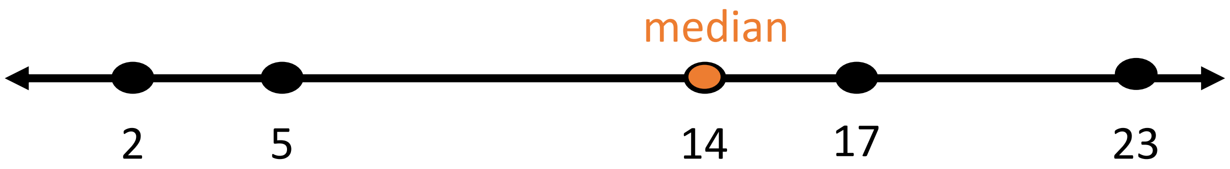
Patient	%emph	%emph - mean	(%emph – mean)^2
1	5	-7.5	56.25
2	23	10.5	110.25
3	2	-10.5	110.25
4	14	1.5	2.25
5	17	4.5	20.25
6	14	1.5	2.25
		Total:	301.5
		Total/(6-1):	60.3
		$SD = \sqrt{60.3}$:	7.77



Descriptive Statistics – Numeric Data – Quantiles and Percentiles

- Recap: The median is the value such that half the data is above and half the data is below

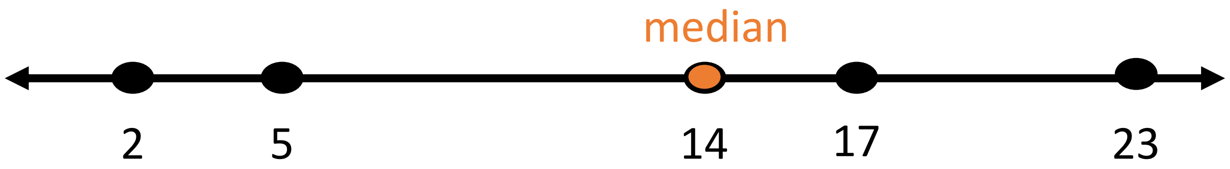
Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17



Descriptive Statistics – Numeric Data – Quantiles and Percentiles

- Recap: The median is the value such that half the data is above and half the data is below
- i.e. 50% of data falls below the median and 50% falls above the median

Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17



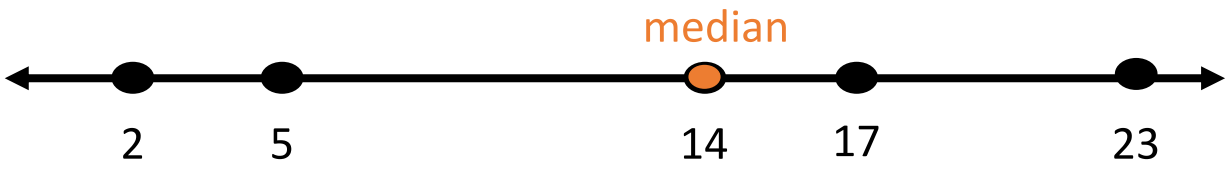
Descriptive Statistics – Numeric Data – Quantiles and Percentiles

- Recap: The median is the value such that half the data is above and half the data is below

- i.e. 50% of data falls below the median and 50% falls above the median

- Therefore, other names for the median are the
  - 50% percentile
  - 0.5 quantile

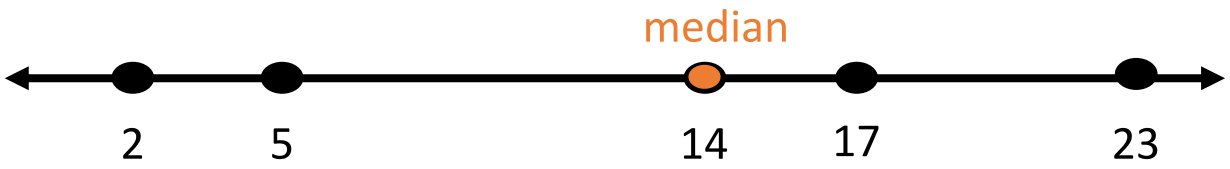
Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17



Descriptive Statistics – Numeric Data – Quantiles and Percentiles

- Recap: The median is the value such that half the data is above and half the data is below
- i.e. 50% of data falls below the median and 50% falls above the median
- Therefore, other names for the median are the
  - 50% percentile
  - 0.5 quantile
- The pth percentile is the value where
  - p% of data falls below the pth percentile
  - (1-p)% of data fall above the pth percentile

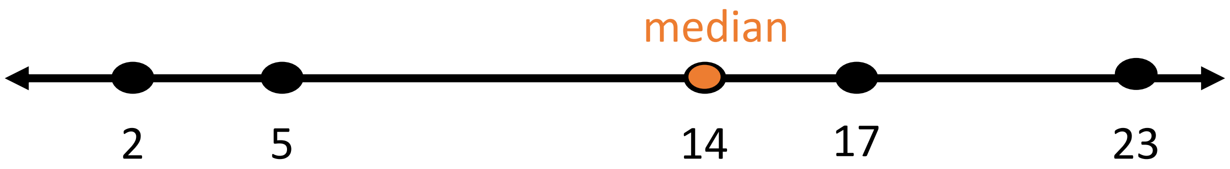
Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17



Descriptive Statistics – Numeric Data – Quantiles and Percentiles

- Two common percentiles are the
  - 25th percentile or “lower quartile”
  - 75th percentile or “upper quartile”

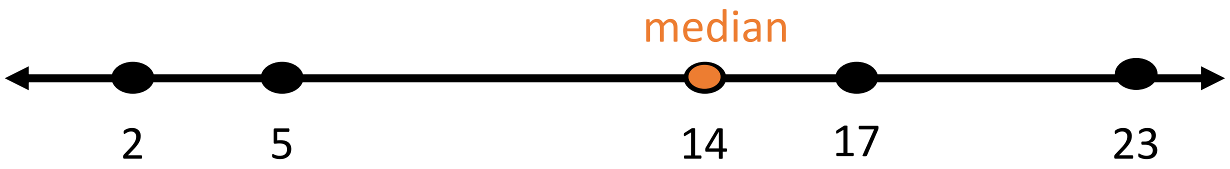
Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17



Descriptive Statistics – Numeric Data – Quantiles and Percentiles

- Two common percentiles are the
  - 25th percentile or “lower quartile”
  - 75th percentile or “upper quartile”
- For the emphysema table on the right,
  - What is the lower quartile?

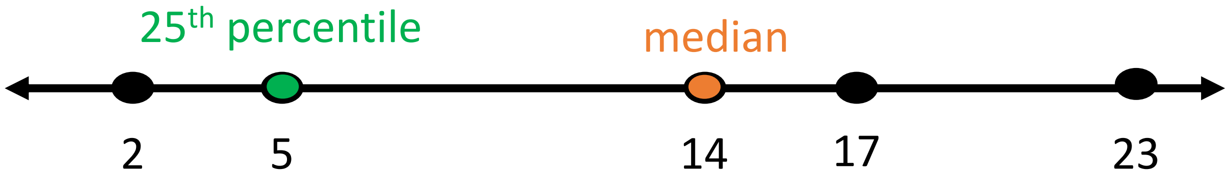
Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17



Descriptive Statistics – Numeric Data – Quantiles and Percentiles

- Two common percentiles are the
  - 25th percentile or "lower quartile"
  - 75th percentile or "upper quartile"
- For the emphysema table on the right,
  - What is the lower quartile? 5

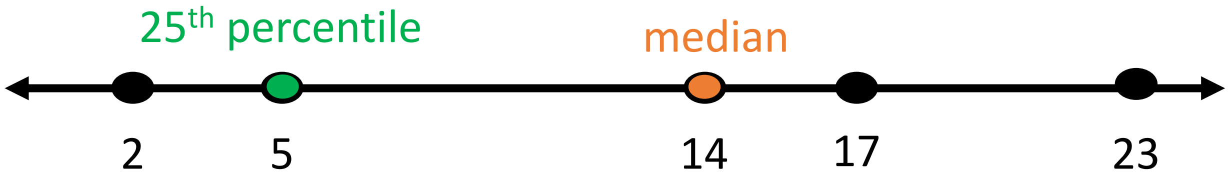
Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17



Descriptive Statistics – Numeric Data – Quantiles and Percentiles

- Two common percentiles are the
  - 25th percentile or “lower quartile”
  - 75th percentile or “upper quartile”
- For the emphysema table on the right,
  - What is the lower quartile? 5
    - That is, 25% of the data falls below 5
    - 75% of the data falls above 5

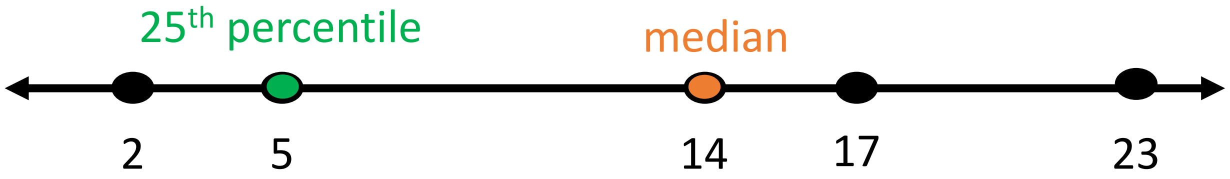
Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17



Descriptive Statistics – Numeric Data – Quantiles and Percentiles

- Two common percentiles are the
  - 25th percentile or “lower quartile”
  - 75th percentile or “upper quartile”
- For the emphysema table on the right,
  - What is the lower quartile? 5
    - That is, 25% of the data falls below 5
    - 75% of the data falls above 5
  - What is the upper quartile?

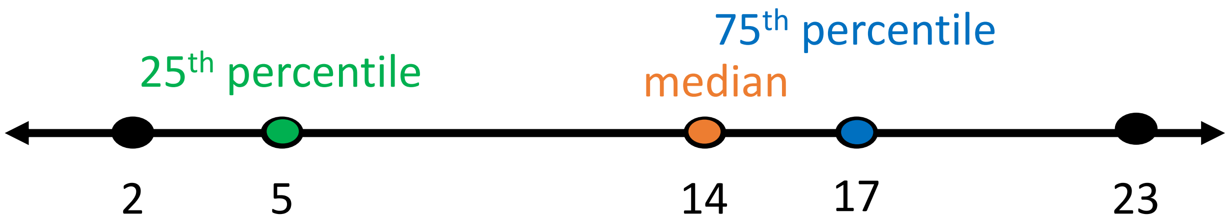
Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17



Descriptive Statistics – Numeric Data – Quantiles and Percentiles

- Two common percentiles are the
  - 25th percentile or “lower quartile”
  - 75th percentile or “upper quartile”
- For the emphysema table on the right,
  - What is the lower quartile? 5
    - That is, 25% of the data falls below 5
    - 75% of the data falls above 5
  - What is the upper quartile? 17

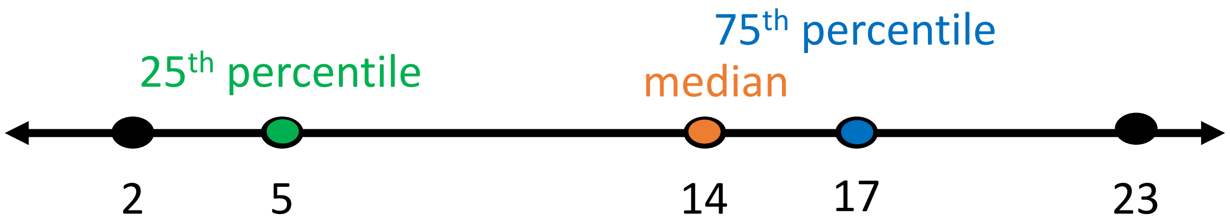
Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17



Descriptive Statistics – Numeric Data – Quantiles and Percentiles

- Two common percentiles are the
  - 25th percentile or “lower quartile”
  - 75th percentile or “upper quartile”
- For the emphysema table on the right,
  - What is the lower quartile? 5
    - That is, 25% of the data falls below 5
    - 75% of the data falls above 5
  - What is the upper quartile? 17
    - That is, 75% of the data falls below 17
    - 25% of the data falls above 17

Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17

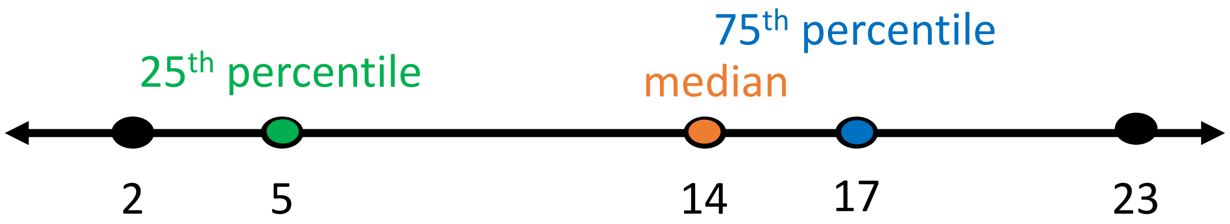


Descriptive Statistics – Numeric Data – Quantiles and Percentiles

- Two common percentiles are the
  - 25th percentile or “lower quartile”
  - 75th percentile or “upper quartile”
- For the emphysema table on the right,
  - What is the lower quartile? 5
    - That is, 25% of the data falls below 5
    - 75% of the data falls above 5
  - What is the upper quartile? 17
    - That is, 75% of the data falls below 17
    - 25% of the data falls above 17

Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17

- Note there are variations for calculating percentiles

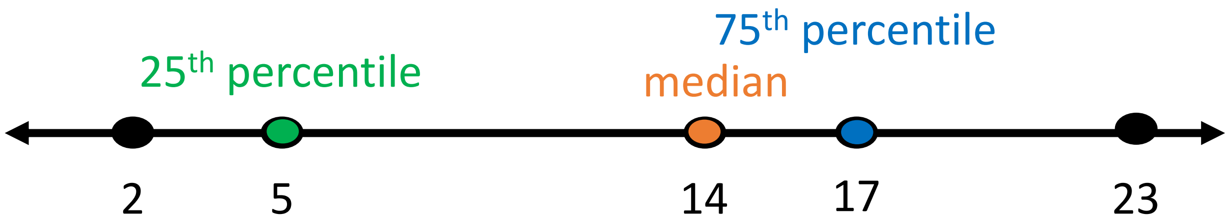


Descriptive Statistics – Numeric Data – Quantiles and Percentiles

- Two common percentiles are the
  - 25th percentile or “lower quartile”
  - 75th percentile or “upper quartile”
- For the emphysema table on the right,
  - What is the lower quartile? 5
    - That is, 25% of the data falls below 5
    - 75% of the data falls above 5
  - What is the upper quartile? 17
    - That is, 75% of the data falls below 17
    - 25% of the data falls above 17

Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17

- Note there are variations for calculating percentiles
- For this class, we will use R to calculate percentiles for us

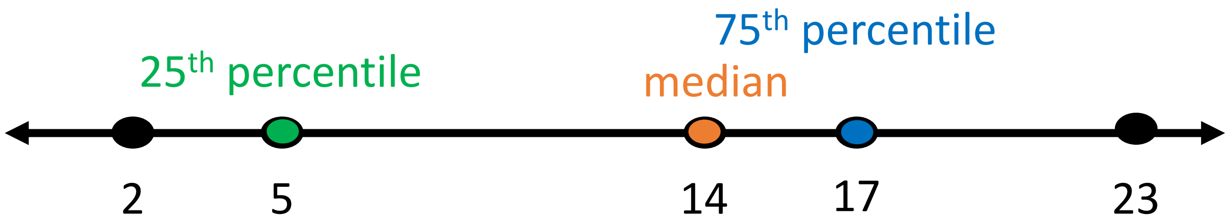


Descriptive Statistics – Numeric Data – Quantiles and Percentiles

- Two common percentiles are the
  - 25th percentile or “lower quartile”
  - 75th percentile or “upper quartile”
- For the emphysema table on the right,
  - What is the lower quartile? 5
    - That is, 25% of the data falls below 5
    - 75% of the data falls above 5
  - What is the upper quartile? 17
    - That is, 75% of the data falls below 17
    - 25% of the data falls above 17

Patient	Percentage Emphysema
1	5
2	23
3	2
4	14
5	17

- Note there are variations for calculating percentiles
- For this class, we will use R to calculate percentiles for us
- I primarily want you to know how to interpret these values



Descriptive Statistics:

Categorical Variables

Descriptive Statistics – Categorical Data

- Summarization of categorical data is much easier!
- We simply count the frequency of each category

Descriptive Statistics – Categorical Data

- Summarization of categorical data is much easier!
- We simply count the frequency of each category
- The table of counts is referred as a
  - Frequency table

    OR
  - Contingency table

Descriptive Statistics – Categorical Data

- Summarization of categorical data is much easier!
- We simply count the frequency of each category
- The table of counts is referred as a
  - Frequency table

    OR
  - Contingency table
- For example, we can categorize % emphysema into three ordinal categories
  - >0, ≤10
  - >10, ≤ 20
  - >20

Patient	Percentage Emphysema	Emphysema Category
1	5	>0, ≤10
2	23	>20
3	2	>0, ≤10
4	14	>10, ≤ 20
5	17	>10, ≤ 20
6	14	>10, ≤ 20

Descriptive Statistics – Categorical Data

- Summarization of categorical data is much easier!
- We simply count the frequency of each category
- The table of counts is referred as a
  - Frequency table

    OR
  - Contingency table
- For example, we can categorize % emphysema into three ordinal categories
  - $>0, \leq 10$
  - $>10, \leq 20$
  - >20
- We then count the number of observations per category

Patient	Percentage Emphysema	Emphysema Category
1	5	$>0, \leq 10$
2	23	>20
3	2	$>0, \leq 10$
4	14	$>10, \leq 20$
5	17	$>10, \leq 20$
6	14	$>10, \leq 20$

Frequency Table

Emphysema Category	Frequency
$>0, \leq 10$	2
$>10, \leq 20$	3
>20	1

Distributions and Histograms

Histograms to Visualize Data Distributions

Measures of central tendency

- Mean (average)
- Median (50th percentile, 0.50 quantile)
- Mode (value that occurs most often)

Measures of variability

- Range
- Variance
- Standard deviation
- Quantiles/Percentiles

- These descriptive statistics describe the attributes/shape of what we call the **data distribution**

Histograms to Visualize Data Distributions

Measures of central tendency

- Mean (average)
- Median (50th percentile, 0.50 quantile)
- Mode (value that occurs most often)

Measures of variability

- Range
- Variance
- Standard deviation
- Quantiles/Percentiles

- These descriptive statistics describe the attributes/shape of what we call the **data distribution**
- The data distribution describes how often values in your data occur

Histograms to Visualize Data Distributions

Measures of central tendency

- Mean (average)
- Median (50th percentile, 0.50 quantile)
- Mode (value that occurs most often)

Measures of variability

- Range
- Variance
- Standard deviation
- Quantiles/Percentiles

- These descriptive statistics describe the attributes/shape of what we call the **data distribution**
- The data distribution describes how often values in your data occur
- The frequency in which data occurs determines the shape of the distribution

Histograms to Visualize Data Distributions

Measures of central tendency

- Mean (average)
- Median (50th percentile, 0.50 quantile)
- Mode (value that occurs most often)

Measures of variability

- Range
- Variance
- Standard deviation
- Quantiles/Percentiles

Why do we care?

Histograms to Visualize Data Distributions

Measures of central tendency

- Mean (average)
- Median (50th percentile, 0.50 quantile)
- Mode (value that occurs most often)

Measures of variability

- Range
- Variance
- Standard deviation
- Quantiles/Percentiles

Why do we care?

- By summarizing the data distribution using descriptive statistics, we can

Histograms to Visualize Data Distributions

Measures of central tendency

- Mean (average)
- Median (50th percentile, 0.50 quantile)
- Mode (value that occurs most often)

Measures of variability

- Range
- Variance
- Standard deviation
- Quantiles/Percentiles

Why do we care?

- By summarizing the data distribution using descriptive statistics, we can
  - Better understand data within a group
    - e.g., salary, survival for cancer patients

Histograms to Visualize Data Distributions

Measures of central tendency

- Mean (average)
- Median (50th percentile, 0.50 quantile)
- Mode (value that occurs most often)

Measures of variability

- Range
- Variance
- Standard deviation
- Quantiles/Percentiles

Why do we care?

- By summarizing the data distribution using descriptive statistics, we can
  - Better understand data within a group
    - e.g., salary, survival for cancer patients
  - and compare data between groups
    - e.g., salary by gender, survival for different cancer treatments, etc.

Histograms to Visualize Data Distributions

Measures of central tendency

- Mean (average)
- Median (50th percentile, 0.50 quantile)
- Mode (value that occurs most often)

Measures of variability

- Range
- Variance
- Standard deviation
- Quantiles/Percentiles

Why do we care?

- By summarizing the data distribution using descriptive statistics, we can
  - Better understand data within a group
    - e.g., salary, survival for cancer patients
  - and compare data between groups
    - e.g., salary by gender, survival for different cancer treatments, etc.

Comparing groups using information about their data distributions is ***statistics!***

Histograms to Visualize Data Distributions

- Data distributions are most commonly visualized using histograms

Histograms to Visualize Data Distributions

- Data distributions are most commonly visualized using histograms

- Histograms describe how often values in your data occur

Histograms to Visualize Data Distributions

- Data distributions are most commonly visualized using histograms
- Histograms describe how often values in your data occur
- Consider the following data set:
    - You collect data on 9 students on the number of years a student has played an instrument

Person	Years Playing Music
1	8
2	6
3	5
4	7
5	6
6	7
7	8
8	7
9	9

Histograms to Visualize Data Distributions

- Data distributions are most commonly visualized using histograms
- Histograms describe how often values in your data occur
- Consider the following data set:
    - You collect data on 9 students on the number of years a student has played an instrument
    - How many times does 5 occur?

Person	Years Playing Music
1	8
2	6
3	5
4	7
5	6
6	7
7	8
8	7
9	9

Histograms to Visualize Data Distributions

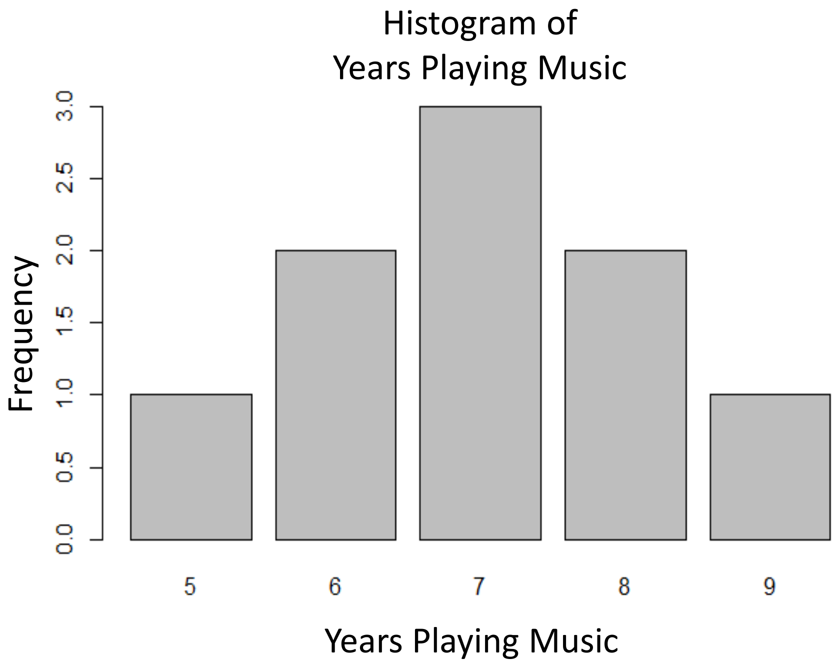
- Data distributions are most commonly visualized using histograms
- Histograms describe how often values in your data occur
- Consider the following data set:
  - You collect data on 9 students on the number of years a student has played an instrument
  - How many times does 5 occur?
  - How many times does 7 occur?

Person	Years Playing Music
1	8
2	6
3	5
4	7
5	6
6	7
7	8
8	7
9	9

Histograms to Visualize Data Distributions

- Data distributions are most commonly visualized using histograms
- Histograms describe how often values in your data occur
- Consider the following data set:
  - You collect data on 9 students on the number of years a student has played an instrument
  - How many times does 5 occur?
  - How many times does 7 occur?
- The plot of these frequencies vs the data values is a **histogram**

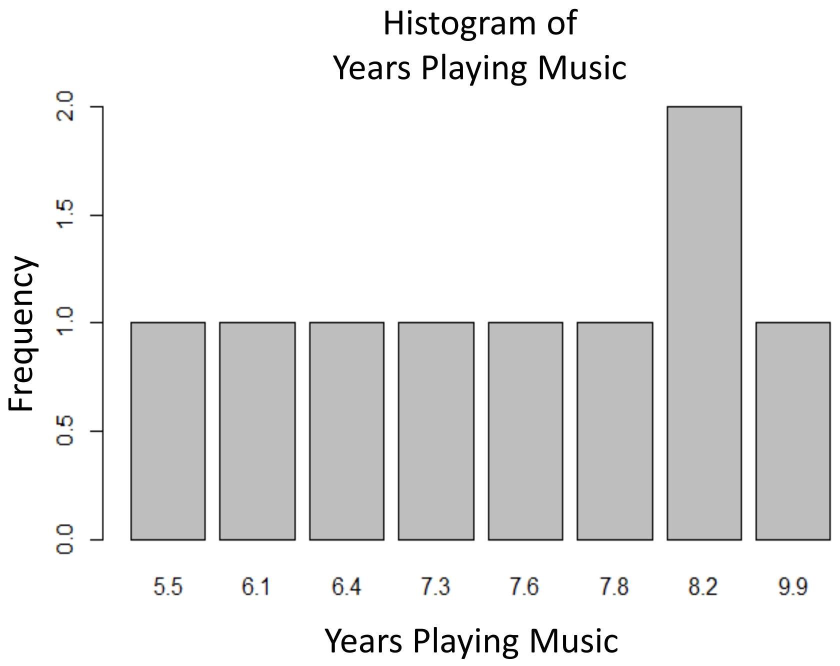
Person	Years Playing Music
1	8
2	6
3	5
4	7
5	6
6	7
7	8
8	7
9	9



Histograms to Visualize Data Distributions

- But what if the data are continuous (decimals) and not discrete (integer-valued)?

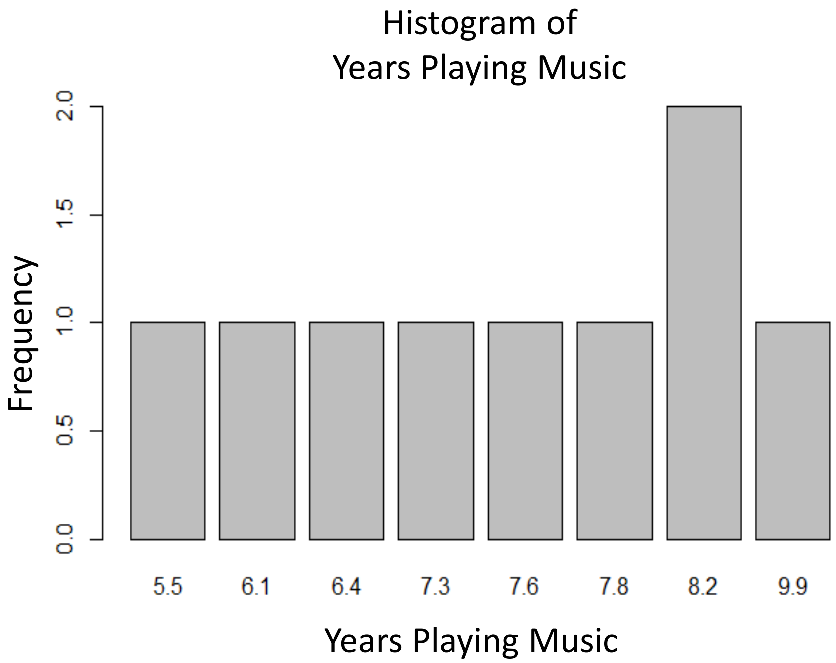
Person	Years Playing Music
1	8.2
2	6.1
3	5.5
4	7.8
5	6.4
6	7.6
7	8.2
8	7.3
9	9.9



Histograms to Visualize Data Distributions

- But what if the data are continuous (decimals) and not discrete (integer-valued)?
- Not very informative!

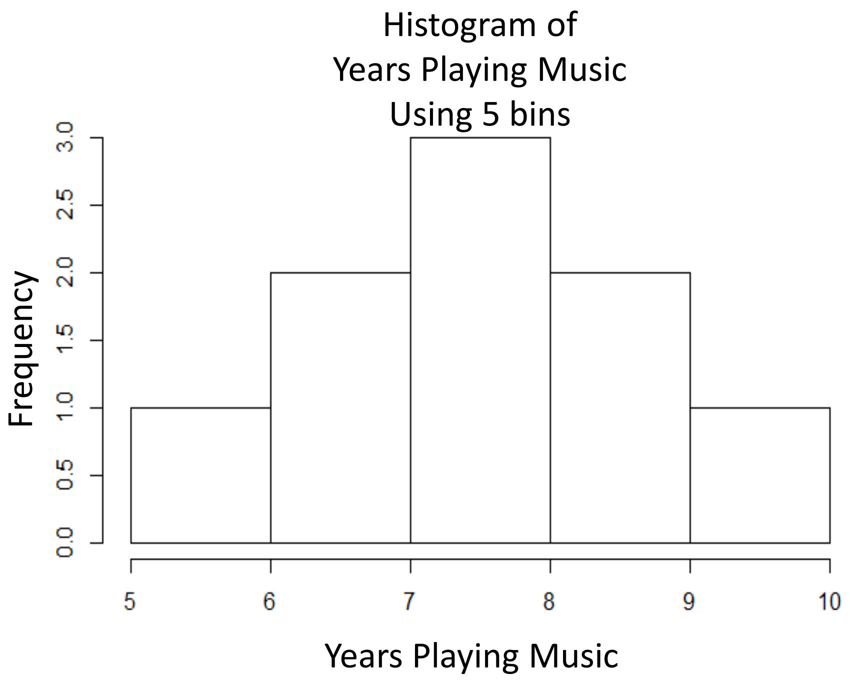
Person	Years Playing Music
1	8.2
2	6.1
3	5.5
4	7.8
5	6.4
6	7.6
7	8.2
8	7.3
9	9.9



Histograms to Visualize Data Distributions

- But what if the data are continuous (decimals) and not discrete (integer-valued)?
- Not very informative!
- In this case, we place the data in “bins” to visualize the distribution
  - e.g., 4-5, 5-6, 6-7

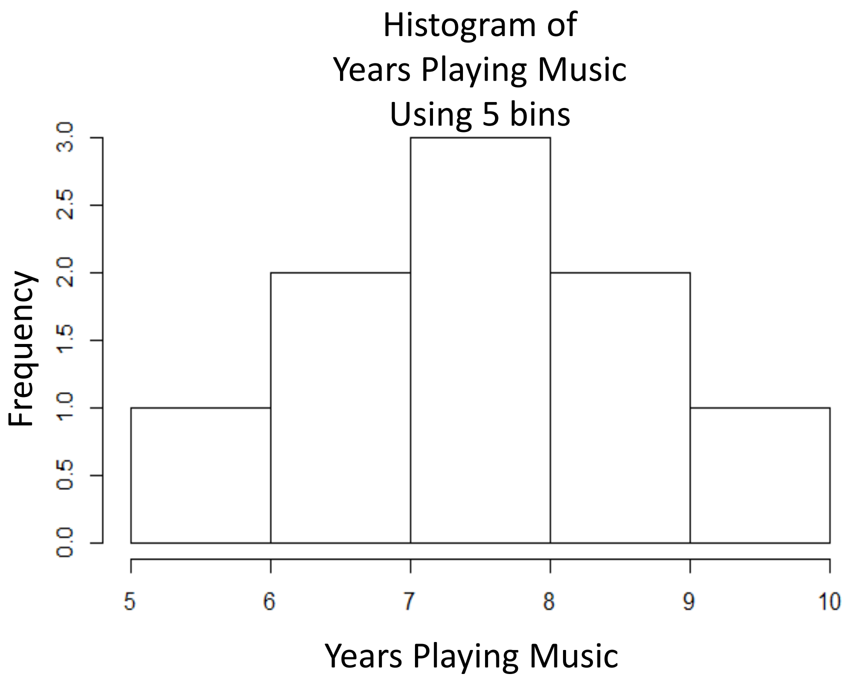
Person	Years Playing Music	Bins
1	8.2	8-9
2	6.1	6-7
3	5.5	5-6
4	7.8	7-8
5	6.4	6-7
6	7.6	7-8
7	8.2	8-9
8	7.3	7-8
9	9.9	9-10



Histograms to Visualize Data Distributions

- But what if the data are continuous (decimals) and not discrete (integer-valued)?
- Not very informative!
- In this case, we place the data in “bins” to visualize the distribution
  - e.g., 4-5, 5-6, 6-7
- The number of bins is chosen by the data scientist

Person	Years Playing Music	Bins
1	8.2	8-9
2	6.1	6-7
3	5.5	5-6
4	7.8	7-8
5	6.4	6-7
6	7.6	7-8
7	8.2	8-9
8	7.3	7-8
9	9.9	9-10

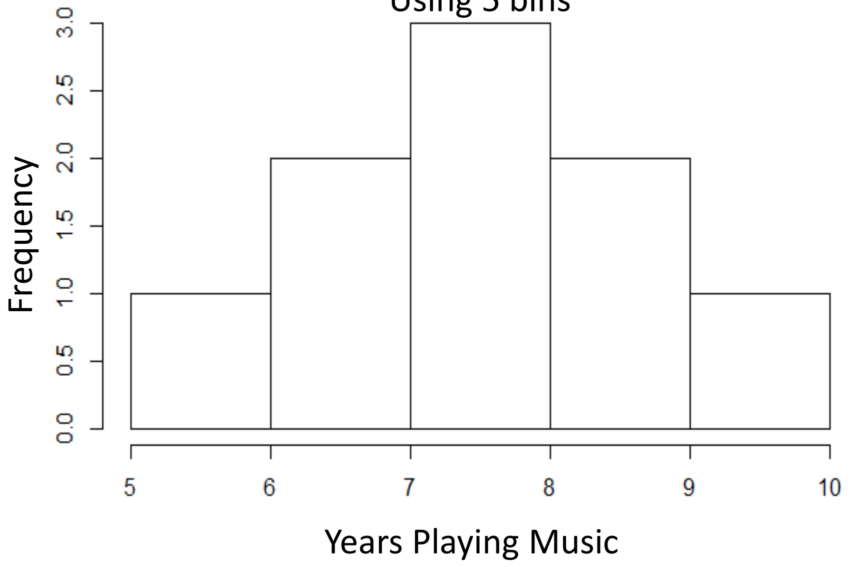


Histograms to Visualize Data Distributions

- But what if the data are continuous (decimals) and not discrete (integer-valued)?
- Not very informative!
- In this case, we place the data in "bins" to visualize the distribution
  - e.g., 4-5, 5-6, 6-7
- The number of bins is chosen by the data scientist
- The best number of bins is subjective

Person	Years Playing Music	Bins
1	8.2	8-9
2	6.1	6-7
3	5.5	5-6
4	7.8	7-8
5	6.4	6-7
6	7.6	7-8
7	8.2	8-9
8	7.3	7-8
9	9.9	9-10

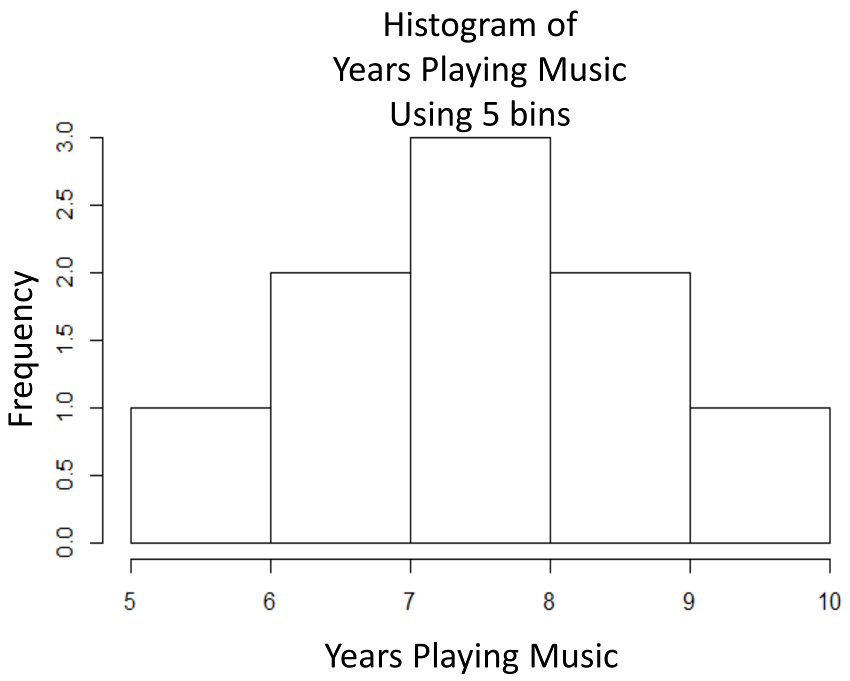
Histogram of
Years Playing Music
Using 5 bins



Histograms to Visualize Data Distributions

- But what if the data are continuous (decimals) and not discrete (integer-valued)?
- Not very informative!
- In this case, we place the data in "bins" to visualize the distribution
    - e.g., 4-5, 5-6, 6-7
- The number of bins is chosen by the data scientist
- The best number of bins is subjective
- You may need to try a few times before acquiring a useful histogram

Person	Years Playing Music	Bins
1	8.2	8-9
2	6.1	6-7
3	5.5	5-6
4	7.8	7-8
5	6.4	6-7
6	7.6	7-8
7	8.2	8-9
8	7.3	7-8
9	9.9	9-10



Histograms to Visualize Data Distributions

- Thus far, we have defined a data distribution and visualized it using a histogram

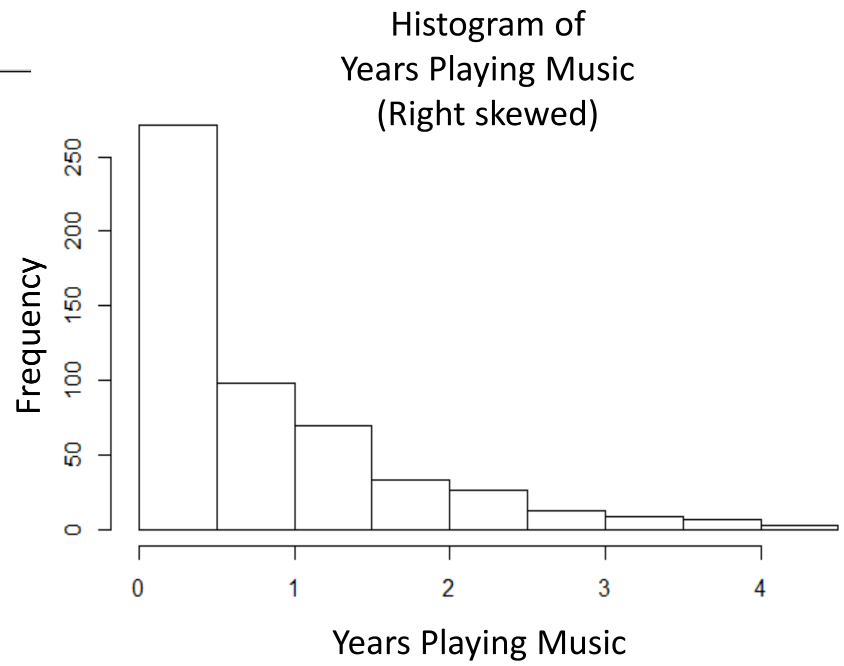
Histograms to Visualize Data Distributions

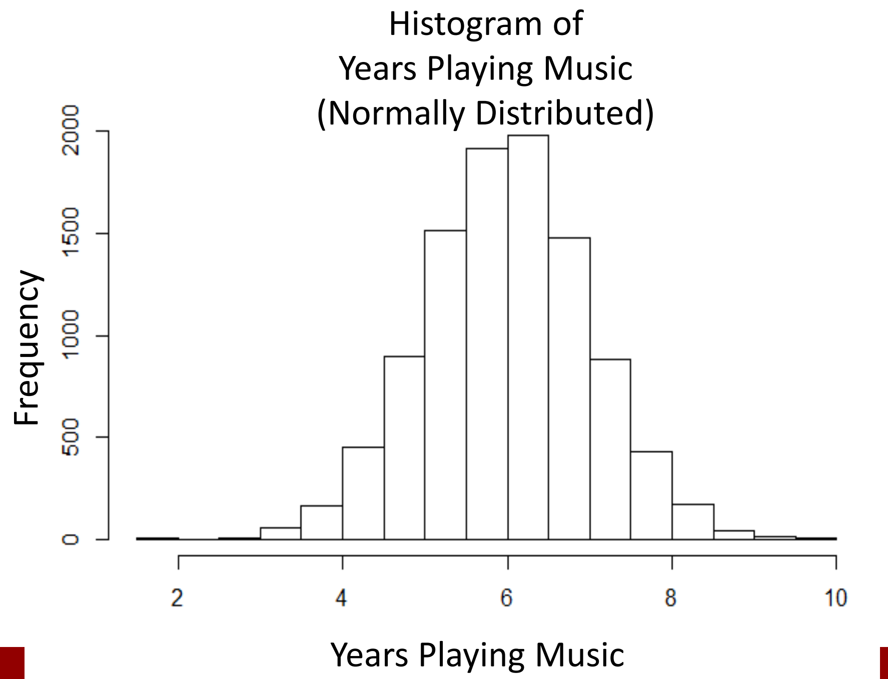
- Thus far, we have defined a data distribution and visualized it using a histogram
- Data distributions can take on many, many different shapes

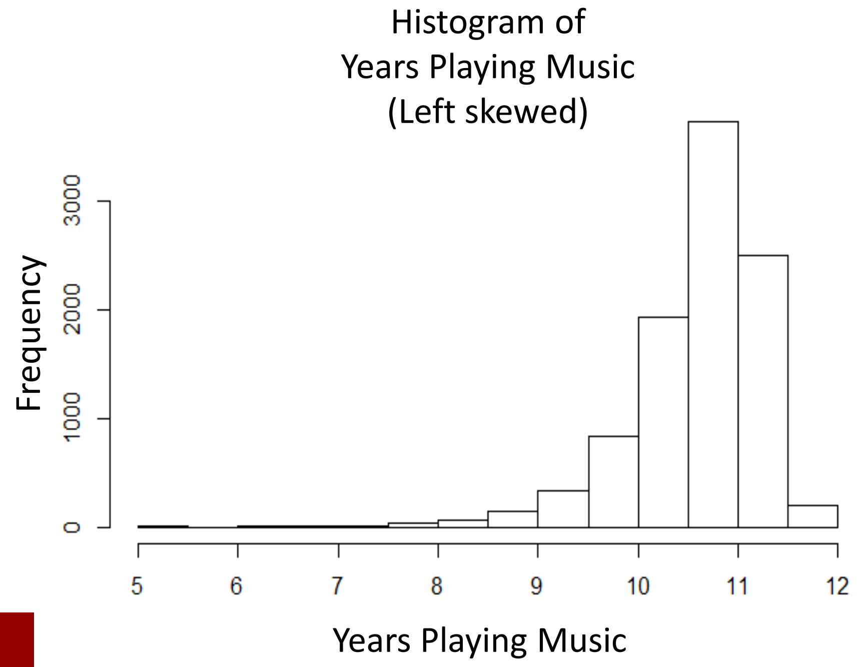
Histograms to Visualize Data Distributions

- Thus far, we have defined a data distribution and visualized it using a histogram

- Data distributions can take on many, many different shapes







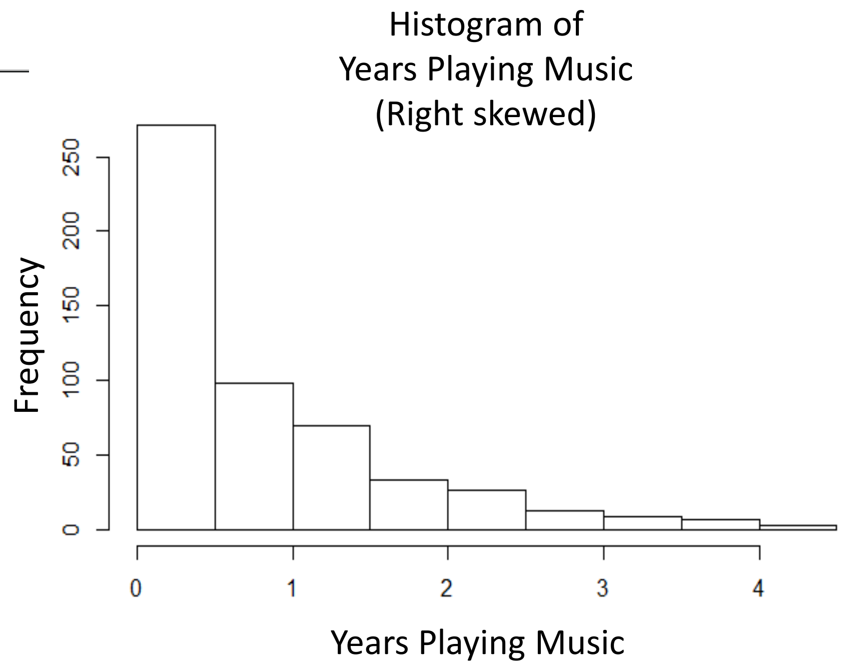
Histograms to Visualize Data Distributions

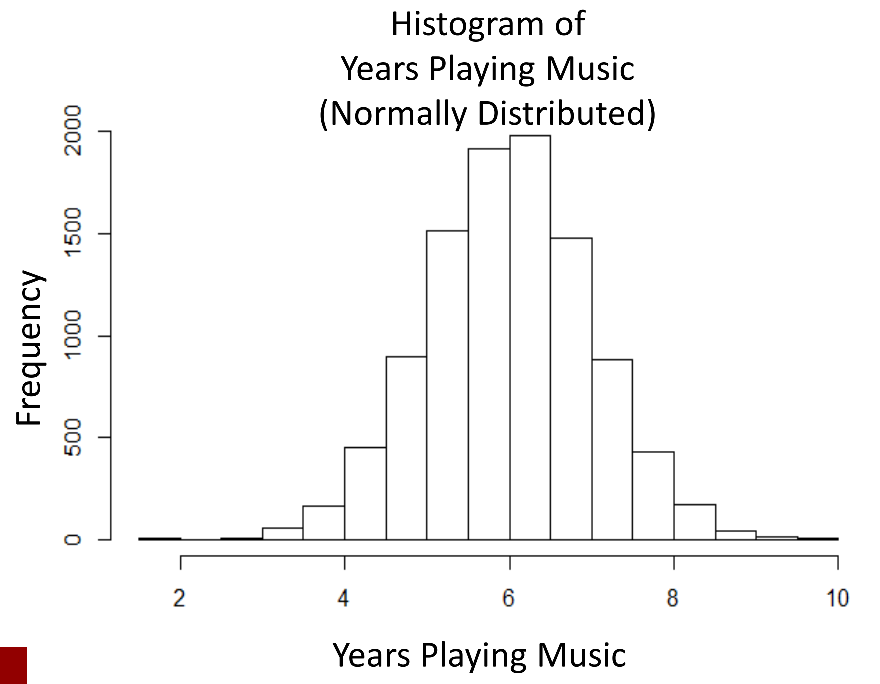
- Thus far, we have defined a data distribution and visualized it using a histogram

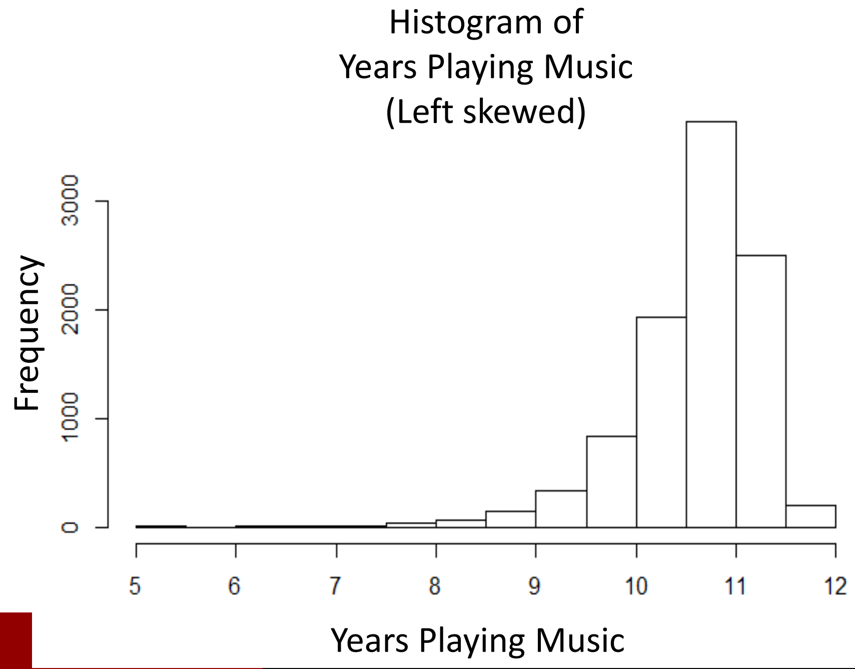
- Data distributions can take on many, many different shapes

- However, the most common is a bell-shaped curve called the normal distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$





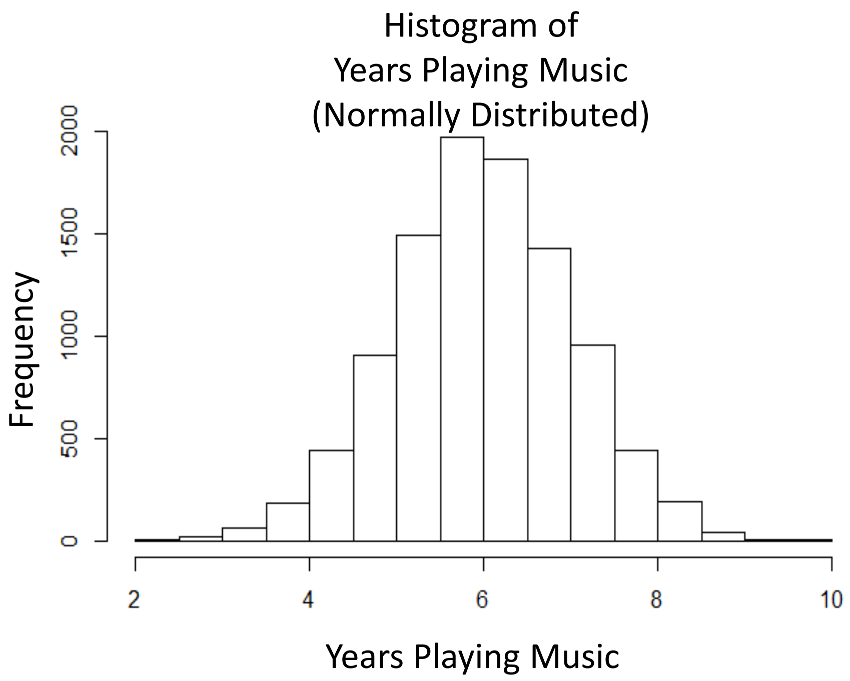


Distributions and Descriptive Statistics

Histograms to Descriptive Statistics

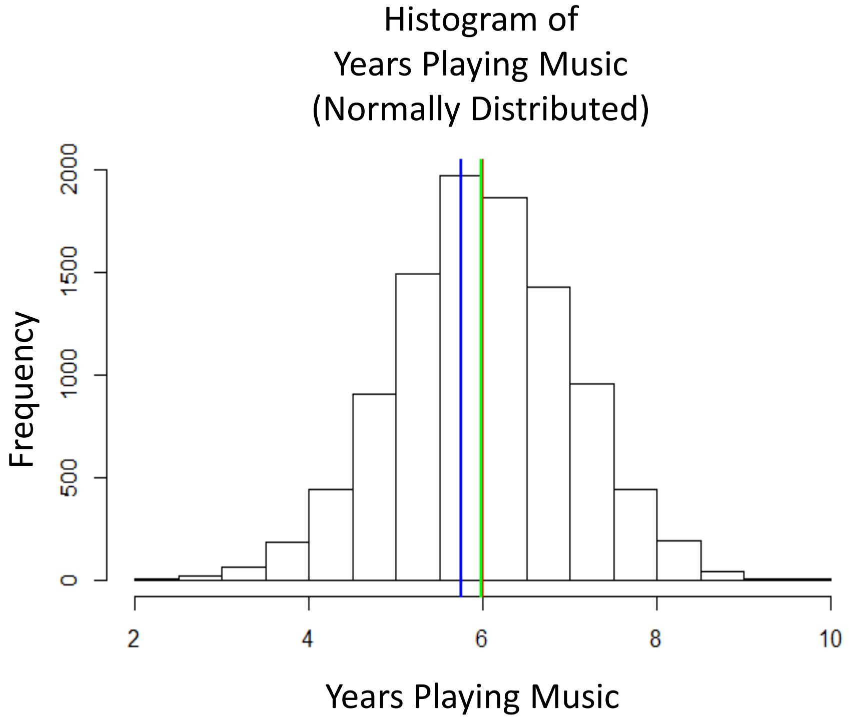
- Descriptive statistics describe different aspects of the distribution

- Central tendency – Where is the center of the distribution?



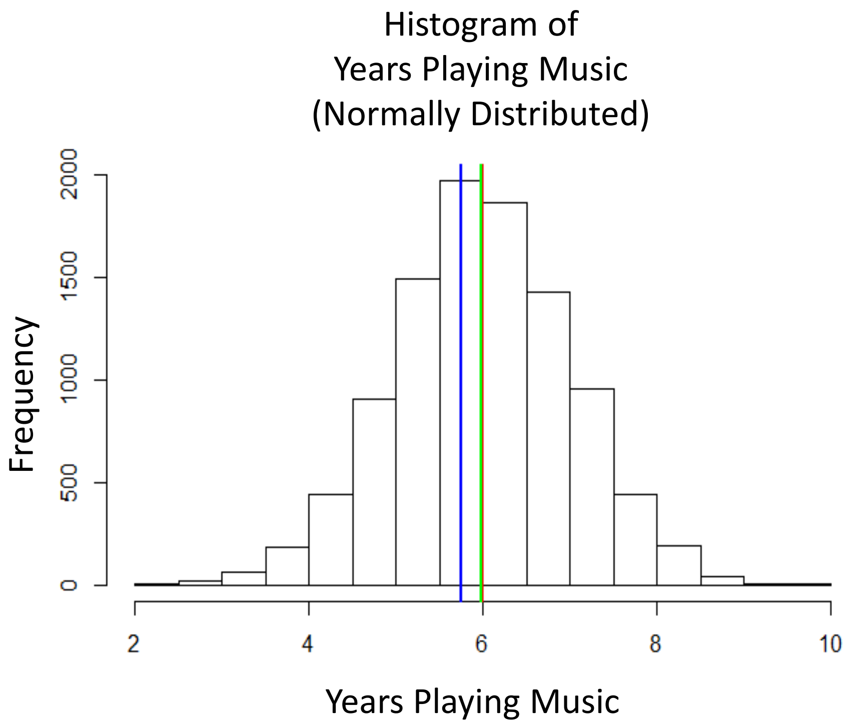
Histograms to Descriptive Statistics

- Descriptive statistics describe different aspects of the distribution
- Central tendency – Where is the center of the distribution?



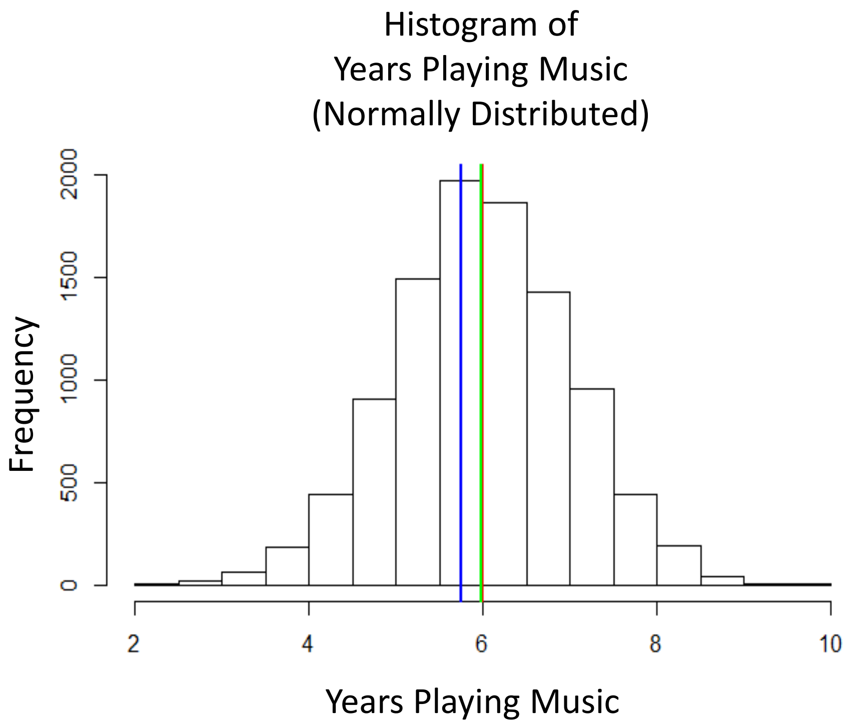
Histograms to Descriptive Statistics

- Descriptive statistics describe different aspects of the distribution
- Central tendency – Where is the center of the distribution?
  - Mean = 5.99, Median = 5.98, Mode = ~5.75



Histograms to Descriptive Statistics

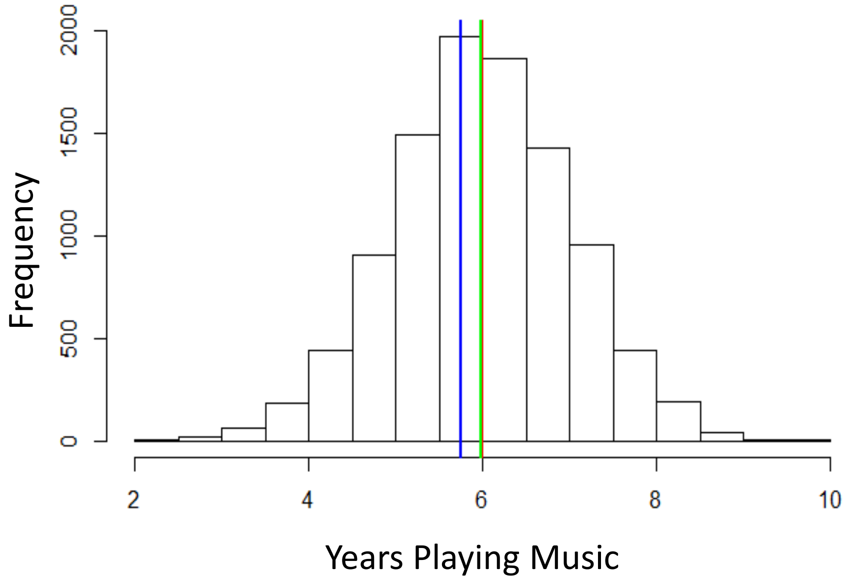
- Descriptive statistics describe different aspects of the distribution
- Central tendency – Where is the center of the distribution?
  - Mean = 5.99, Median = 5.98, Mode = ~5.75
- What can we say about years playing music among students in our sample?



Histograms to Descriptive Statistics

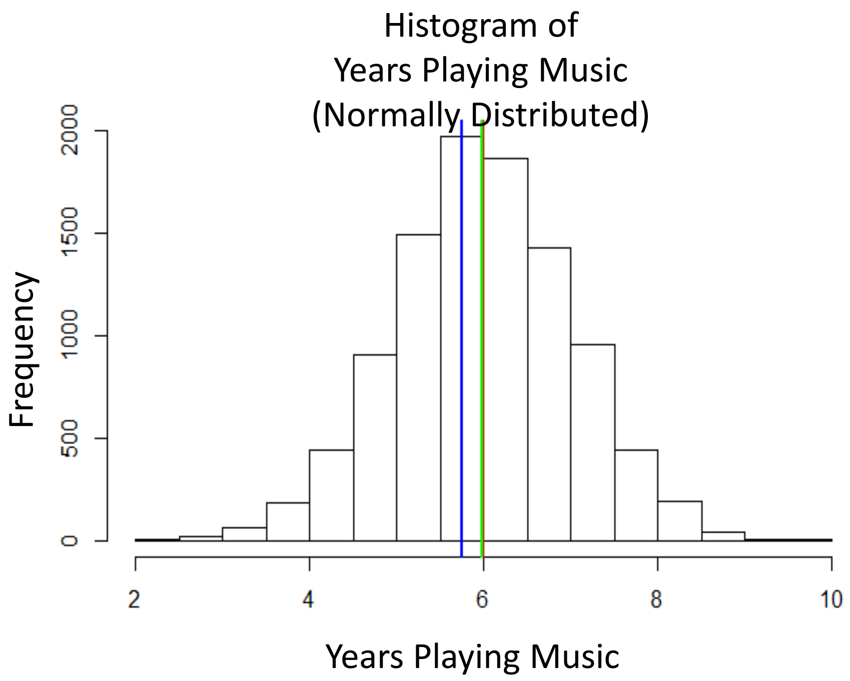
- Descriptive statistics describe different aspects of the distribution
- Central tendency – Where is the center of the distribution?
  - Mean = 5.99, Median = 5.98, Mode = ~5.75
- What can we say about years playing music among students in our sample?
- **Note:** For the theoretical normal distribution, the mean = median = mode

Histogram of
Years Playing Music
(Normally Distributed)



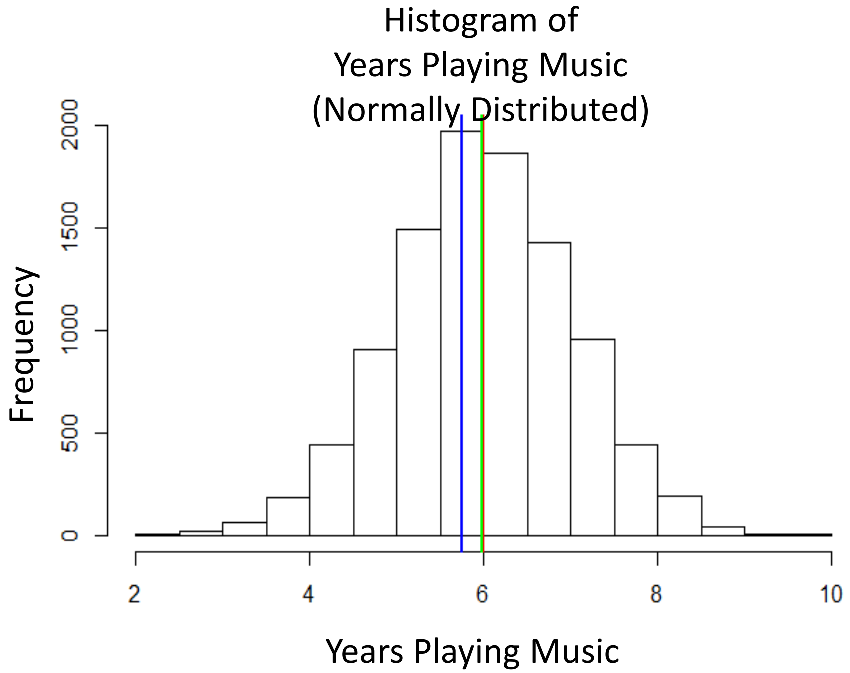
Histograms to Descriptive Statistics

- Descriptive statistics describe different aspects of the distribution
- Dispersion – What is the range of years of music played?



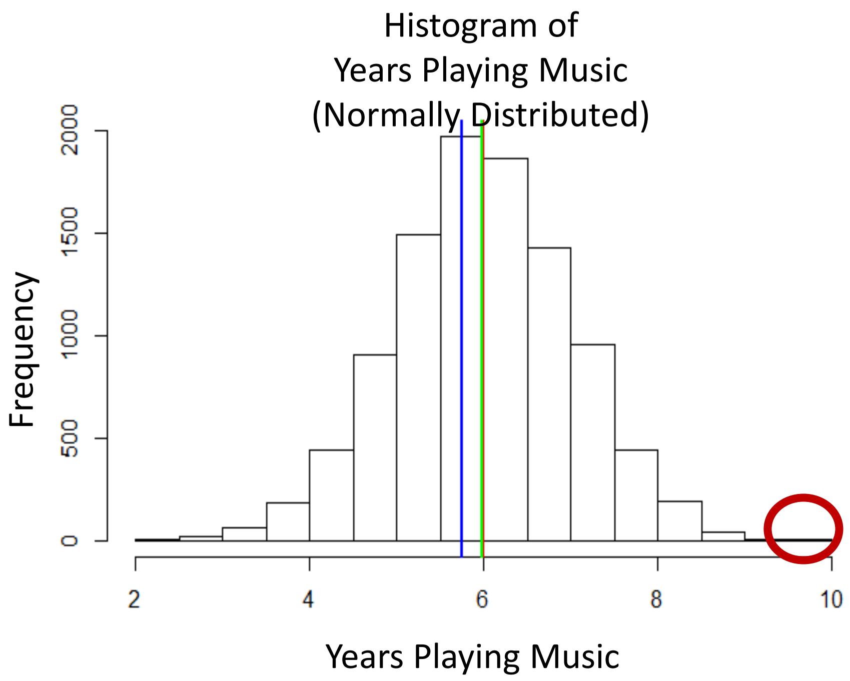
Histograms to Descriptive Statistics

- Descriptive statistics describe different aspects of the distribution
- Dispersion – What is the range of years of music played?
  - 2 to 10 years



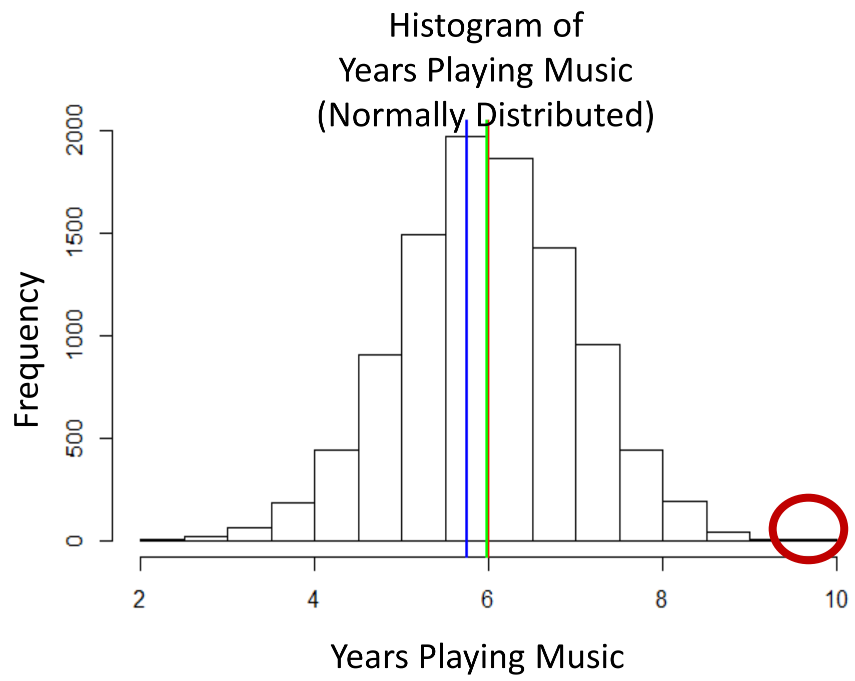
Histograms to Descriptive Statistics

- Descriptive statistics describe different aspects of the distribution
- Dispersion – What is the range of years of music played?
  - 2 to 10 years
- What can we say about this student?



Histograms to Descriptive Statistics

- Descriptive statistics describe different aspects of the distribution
- Dispersion – What is the range of years of music played?
  - 2 to 10 years
- What can we say about this student?
- Where is the student with the least amount of experience playing music?

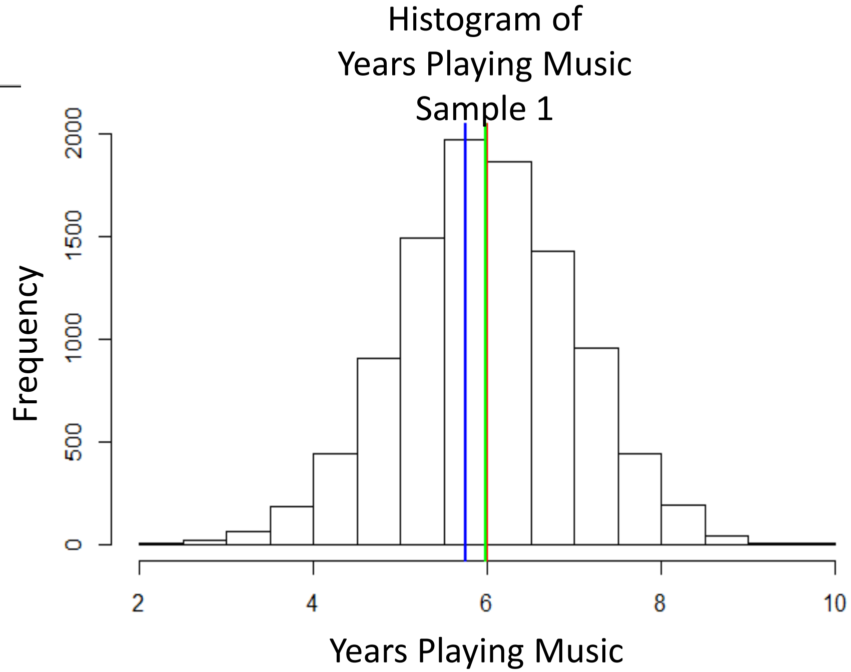


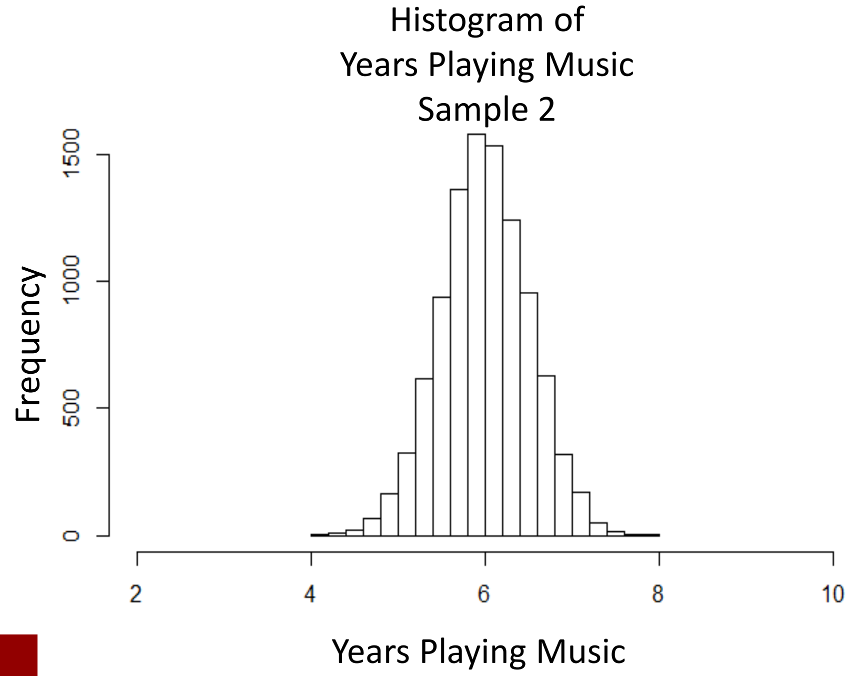
Histograms to Descriptive Statistics

- Imagine we collect data on two different groups of students (shown on the right)
- Which sample (1 or 2) has a smaller standard deviation in years of playing music?

A) Sample 1

B) Sample 2



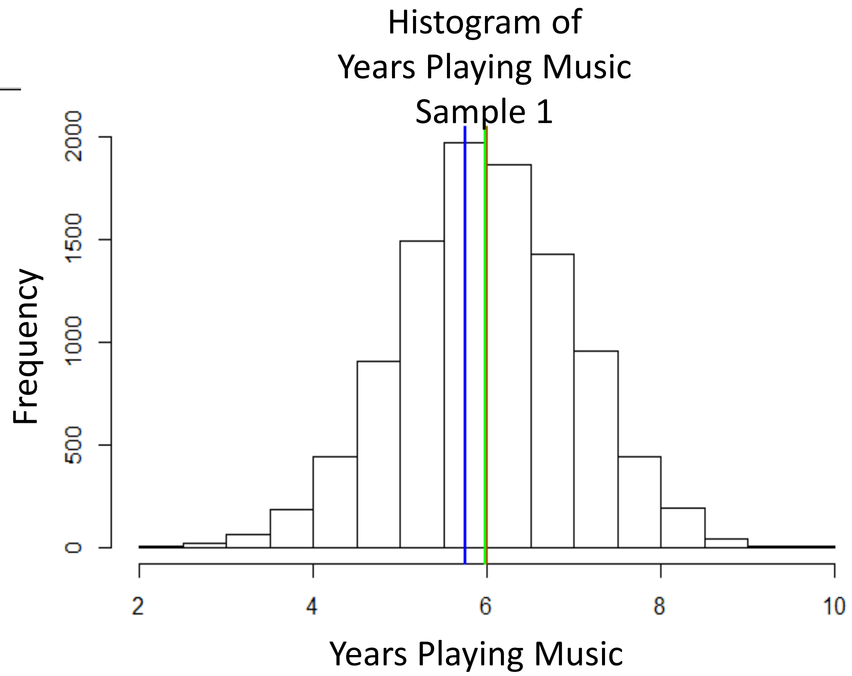


Histograms to Descriptive Statistics

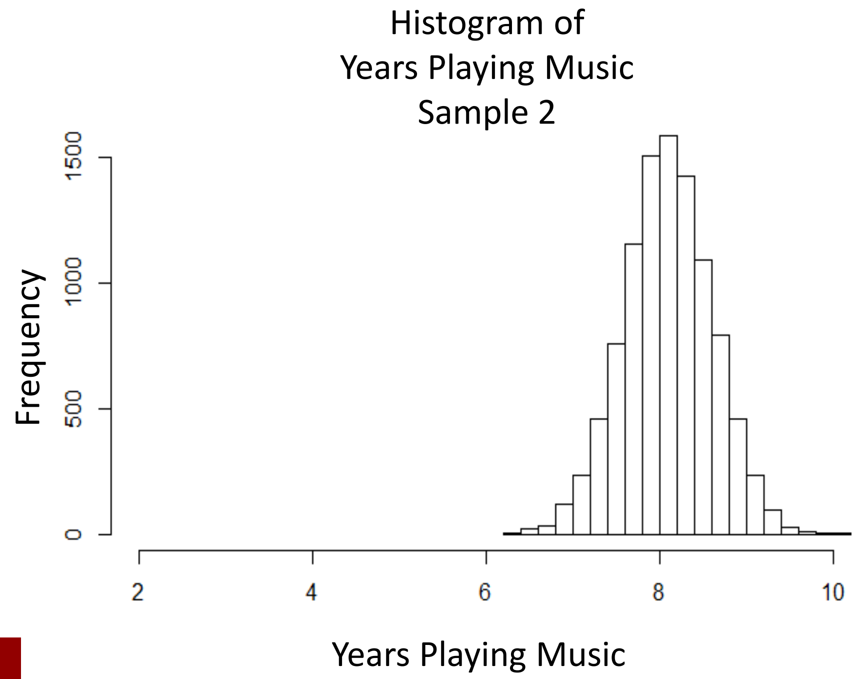
- Imagine we collect data on two different groups of students (shown on the right)
- Which sample (1 or 2) has a smaller standard deviation in years of playing music?

A) Sample 1

B) Sample 2



Histogram of Years Playing Music Sample 1

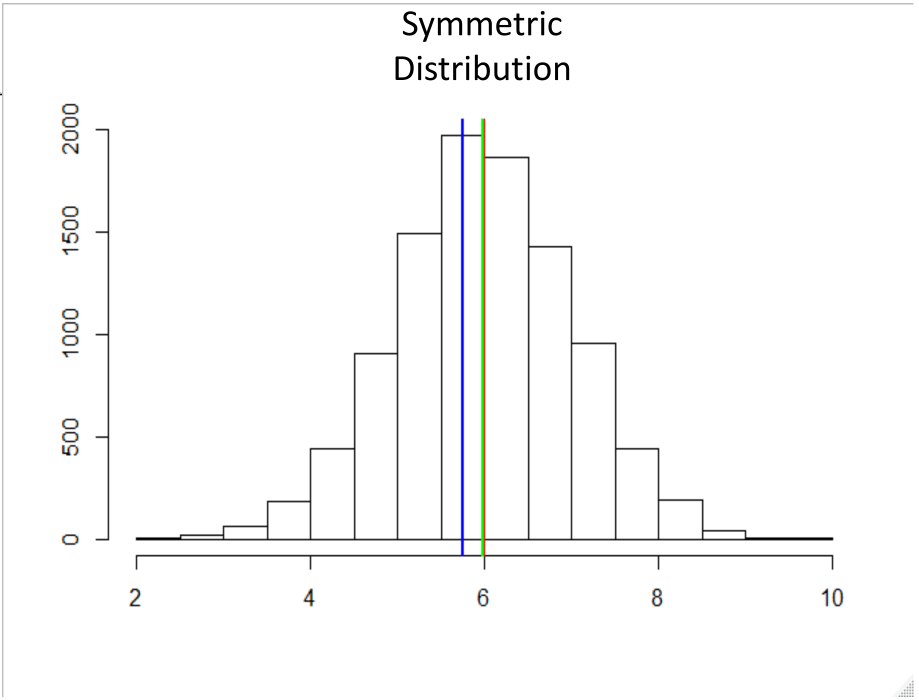


Histogram of Years Playing Music Sample 2

Histograms to Visualize Data Distributions

- When the distribution is close to symmetric the

  mean ≈ median



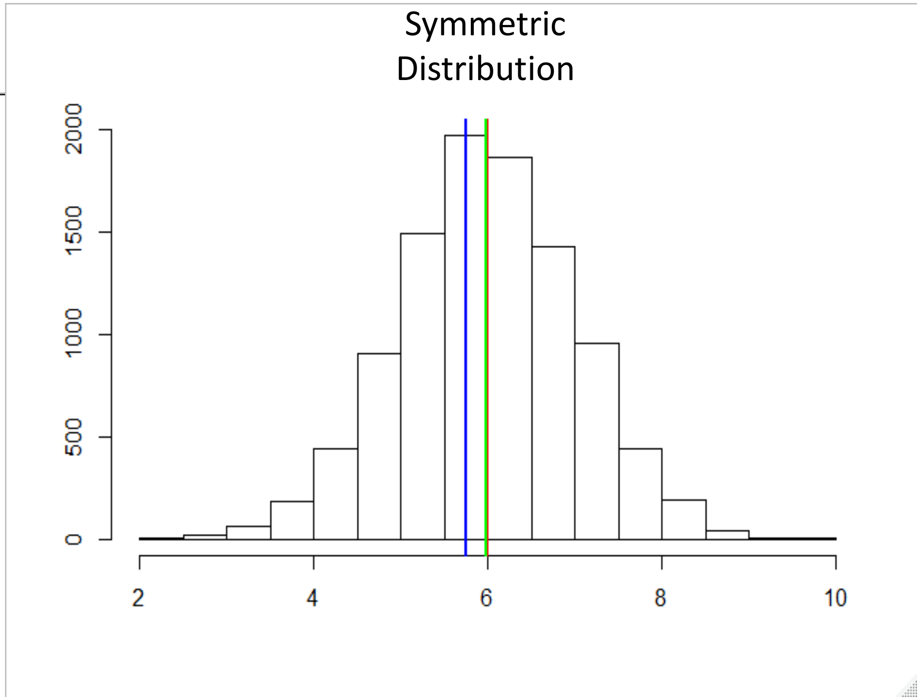
Histograms to Visualize Data Distributions

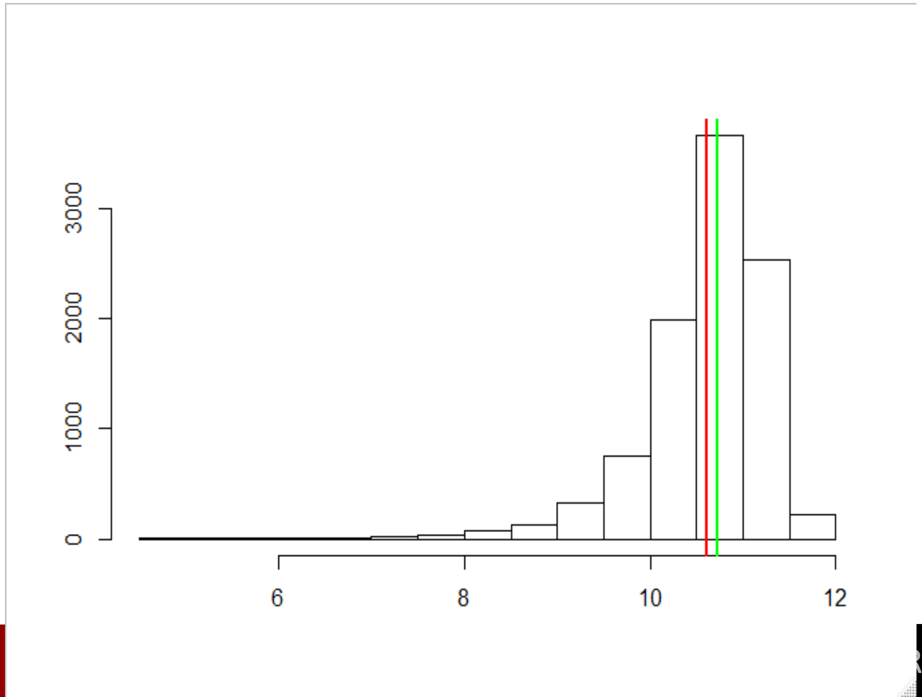
- When the distribution is close to symmetric the

  mean ≈ median

- Left skewed distributions

  mean < median





Histograms to Visualize Data Distributions

- When the distribution is close to symmetric the

  mean ≈ median

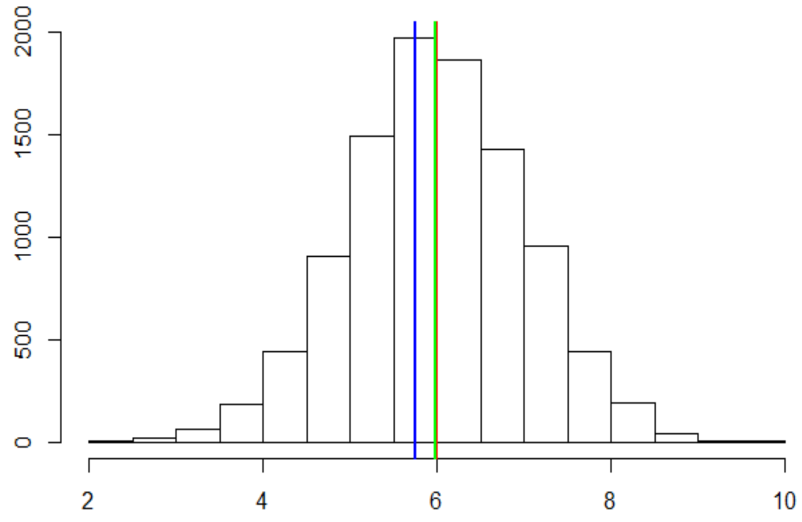
- Left skewed distributions

  mean < median

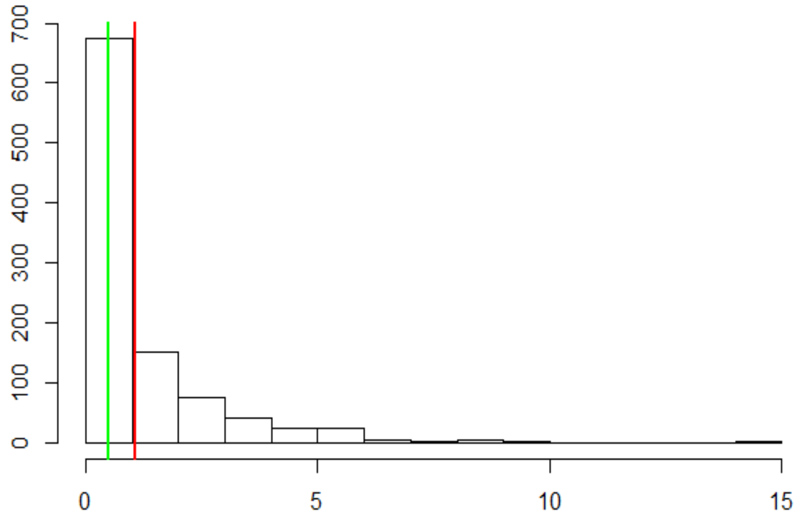
- Right skewed distributions

  mean > median

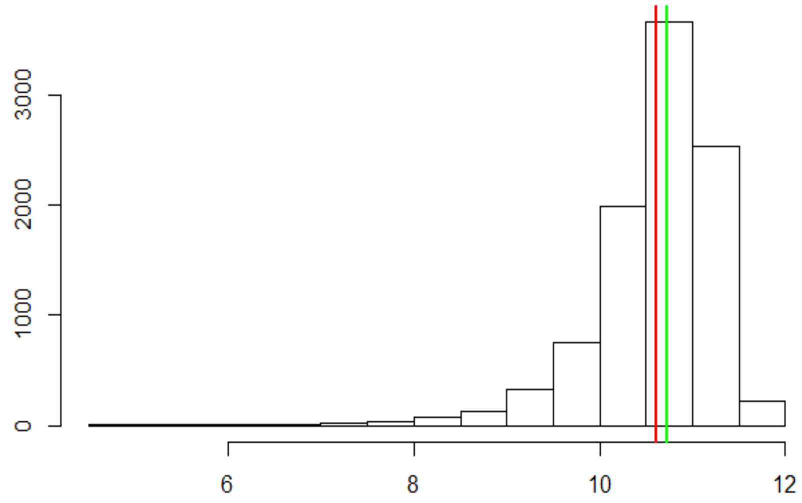
Symmetric Distribution



Right skewed

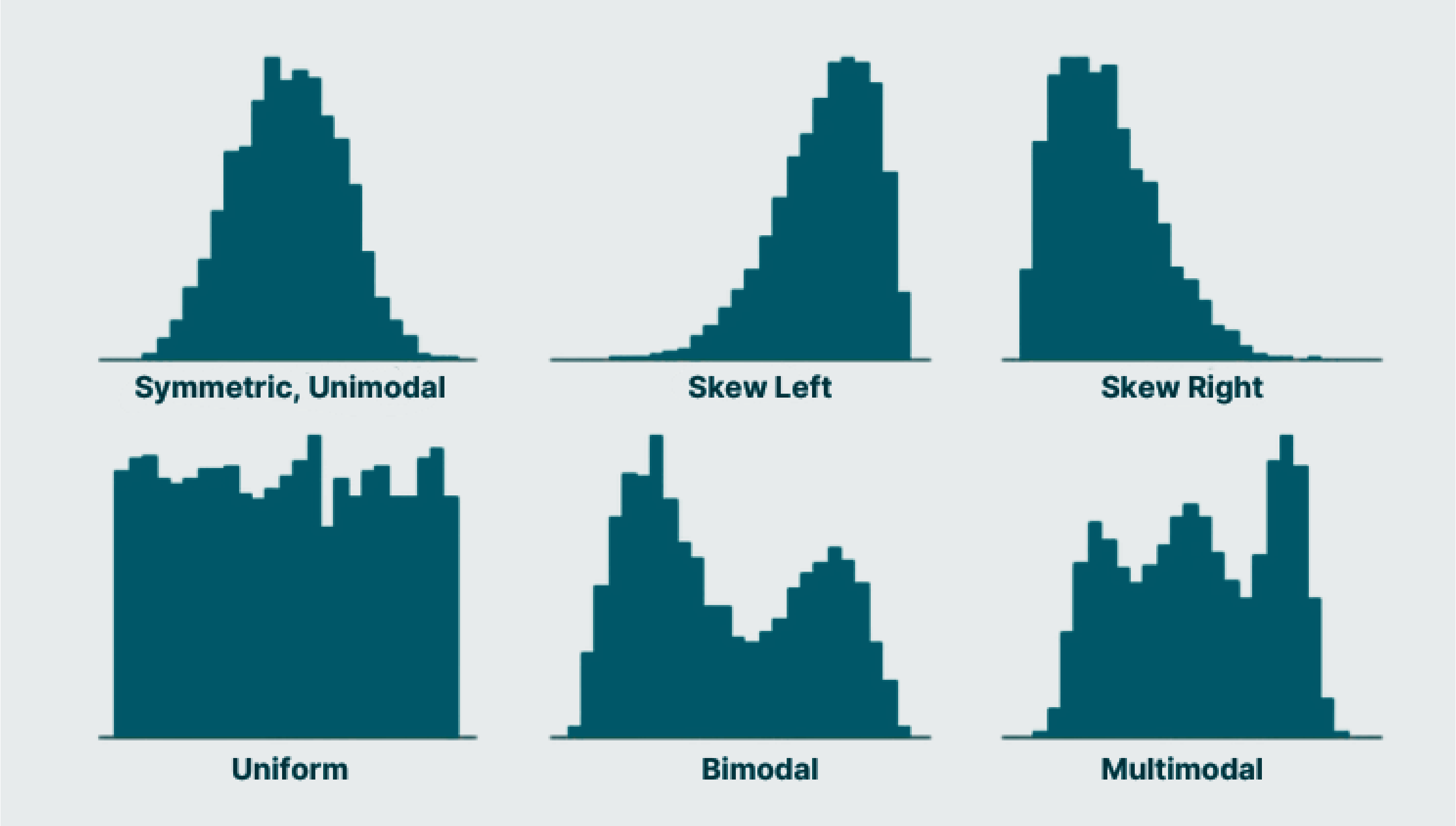


Left skewed



Histograms to Visualize Data Distributions

Data distributions can take on many, many different shapes!



R Code Covered in Practice Assignment