Math 11
Calculus-Based Introductory Probability and Statistics

Eddie Aamari
S.E.W. Assistant Professor

eaamari@ucsd.edu
math.ucsd.edu/~eaamari/
AP&M 5880A

Math 11
Calculus-Based Introductory Probability and Statistics

Eddie Aamari
S.E.W. Assistant Professor

eaamari@ucsd.edu
math.ucsd.edu/~eaamari/
AP&M 5880A

Today:

- Presentation of the course
- Introduction to data

Course Home

Instructor's webpage:

math.ucsd.edu/~eaamari/teaching_2018winter_math11.html

- Lecture slides
- Homework sets
- Course syllabus
- Provisional course calendar
- Link to Piazza
- Office hours times and locations (see syllabus)

Homework

- Homework is due weekly on Tuesday's lecture.
- Late assignments will not be accepted.
- Your worst homework grade will be dropped.
- Randomly selected problems on the assignment will be graded.

Tacit homework: Read the textbook!

- Homework handed back on Discussion sections.
- No homework re-grading will be allowed after the section ends. This means that if you come back after you went out the room, your grade is fixed and your homework will not be regraded. Complaints/reclamation during the section will be considered with concern.

Computer Labs

In addition, you will be given eight weekly computer labs.

Check out and follow conscientiously the associated website:

http://www.math.ucsd.edu/math11/W18.html

- These labs will be due on **Friday at 9pm** each week without an exam. Labs should be submitted on TritonEd by their respective due dates.
- Lab assignments can be submitted up to 1 hour late at the cost of a 1-point penalty. Assignments submitted later than this will not be accepted, excepting the first assignment which can be submitted up to a week late for a 1-point penalty.
- Questions about labs should be directed to the head lab TA:

  Selene Xu (yux033@ucsd.edu)
- Lab Office Hours held weekly, on

  Friday, 11am-1pm AP&M B349

How the Course is Graded

The one following formula giving you the better result will be used:

Formula 1		Formula 2	
15%	Homework	15%	Homework
20%	Computer labs	20%	Computer labs
15%	Midterm Exam 1	15%	Best Midterm Exam
15%	Midterm Exam 2	50%	Final Exam
35%	Final Exam		

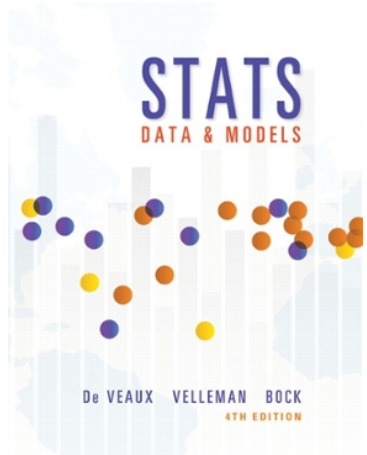
- Your worst homework grade will be dropped for computing your final *Homework* score.
- No makeup exams.
- The grading scheme will be curved and scaled to the best student in class.

Class Calendar

	Monday	Tuesday	Wednesday	Thursday	Friday
Week 1	January 8	9 Chapters 3-4-6	10	11 Chapters 6-7	12 Lab 1 due
Week 2	15 Martin Luther King, Jr. Holiday	16 Chapters 8-9 HW 1 due	17	18 Chapters 13-14	19 Lab 2 due
Week 3	22	23 Chapter 14 HW 2 due	24	25 Chapter 15	26 Lab 3 due
Week 4	29	**30** **Midterm 1**	31	February 1 Chapter 16	2
Week 5	5	6 Chapter 15 HW 3 due	7	8 Chapters 15-5	9 Lab 4 due
Week 6	12	13 Chapters 11-16-17 HW 4 due	14	15 Chapters 17-18	16 Lab 5 due
Week 7	19 Presidents' Day	20 Chapter 19 HW 5 due	21	22 Chapters 20-21	23 Lab 6 due
Week 8	26	**27** **Midterm 2**	28	March 1 Chapters 20-21	2
Week 9	5	6 Chapters 22-23 HW 6 due	7	8 Chapters 22-23	9 Lab 7 due
Week 10	12	13 Chapter 25 HW 7 due	14	15 Chapters 25-24	16 Lab 8 due
Final Week	19	20	21	**22** **Final Exam** **3pm-6pm**	23

Components you Need

Textbook: Stats, Data & Models, 4th Edition.
by De Veaux, Velleman & Bock



Software: Minitab 18
$\rightarrow$ Free
Download instructions on TritonEd.
Used for labs

Calculators:

- Used on exams and homework
- Need not be graphics, nor have statistical functions
- Cannot be your phone or computer (for exams)

Discussion Board

The Piazza forum for our class where questions can be posted and answered.

https://piazza.com/ucsd/winter2018/math11lectureb/home

Before Carrying On...

Any questions so far?

Types of Data

Types of Data

- **Categorical/Qualitative**
  Data that fall into categories or labels; often text ideas; tend NOT to have units
  Examples: Gender, marital status, area code

Types of Data

- **Categorical/Qualitative**
  Data that fall into categories or labels; often text ideas; tend NOT to have units
  Examples: Gender, marital status, area code
- **Numeric/Quantitative**
  Numerical data that have units; it usually makes sense to do math operations on these data
  Examples: Age, salary, # text messages sent last month

Types of Data

- **Categorical/Qualitative**
  Data that fall into categories or labels; often text ideas; tend NOT to have units
  Examples: Gender, marital status, area code
- **Numeric/Quantitative**
  Numerical data that have units; it usually makes sense to do math operations on these data
  Examples: Age, salary, # text messages sent last month

Be Careful About Data Types:
Some variables encoded with numbers are not numeric.

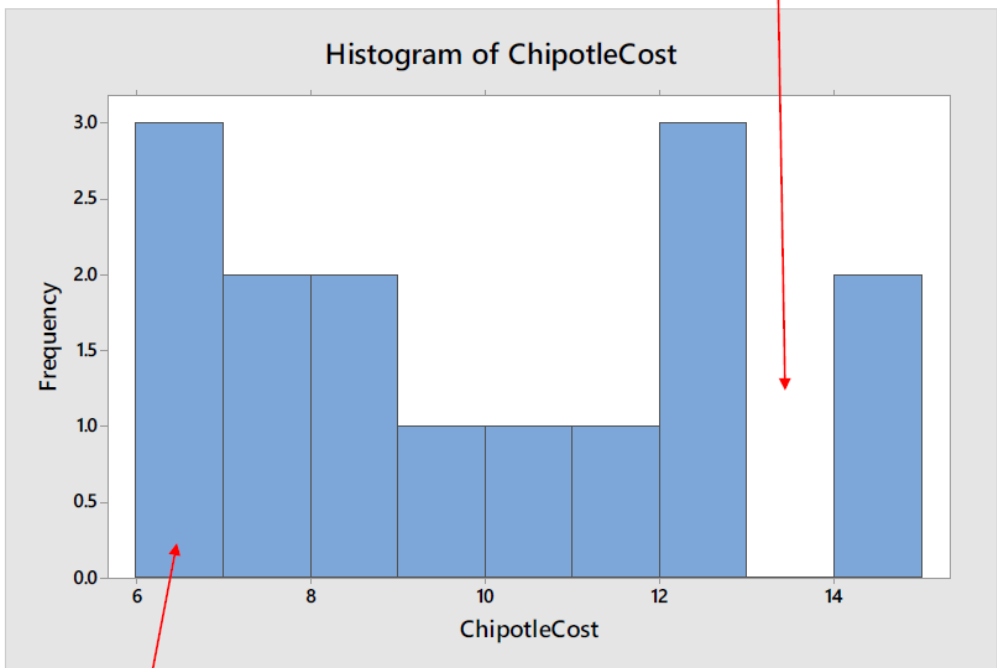
- 0/1 for TRUE/FALSE
- ZIP codes (92093)

Not all numeric variables look like numbers.

- Dates (Friday the 13th, 2017)
- GPS coordinates (40°26' 46" N 79°58' 56" W)

Numerical Data: Histograms

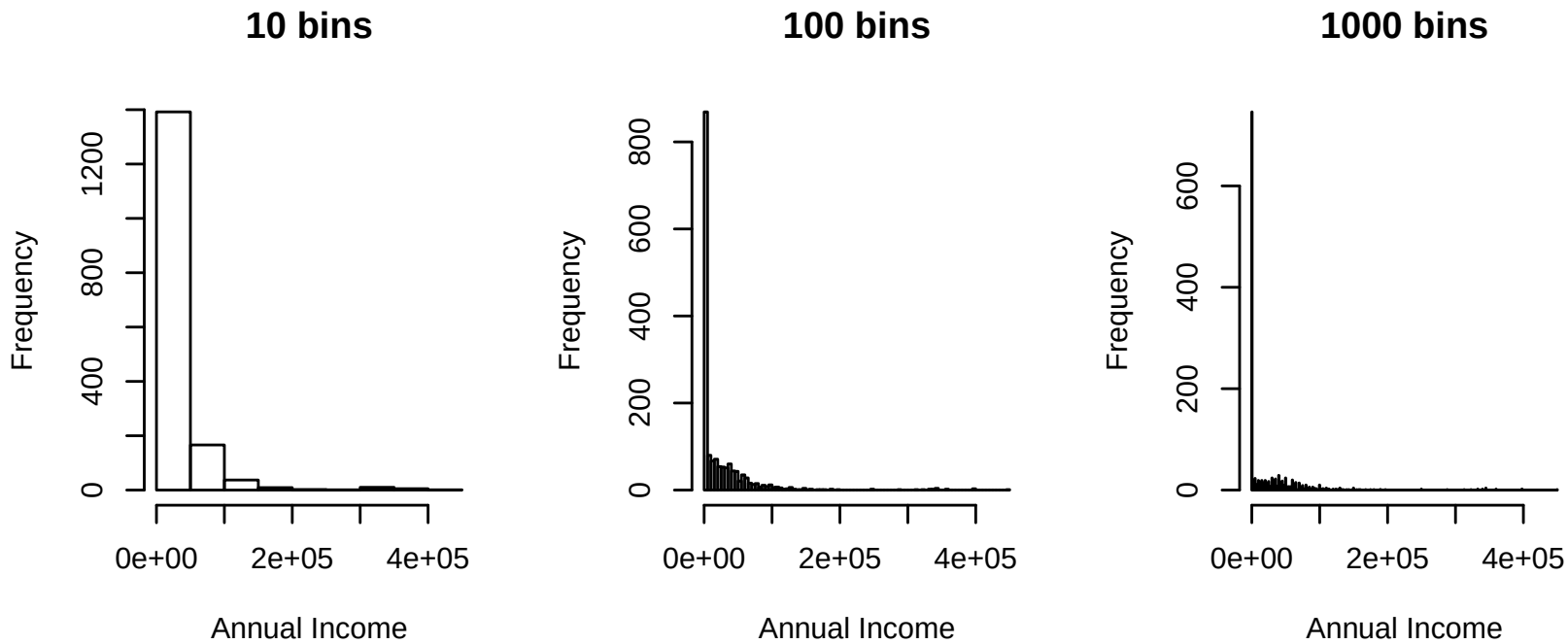
ChipotleCost
6.50
7.29
8.25
6.34
12.31
11.39
14.23
12.65
12.21
7.53
14.21
6.79
8.76
9.32
10.53



3 purchases where $6 \leq \text{cost} < 7$

Numerical Data: Histograms

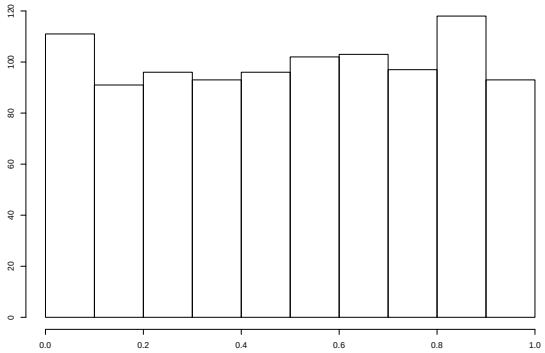
The choice of bin size influences crudely the histogram plot.



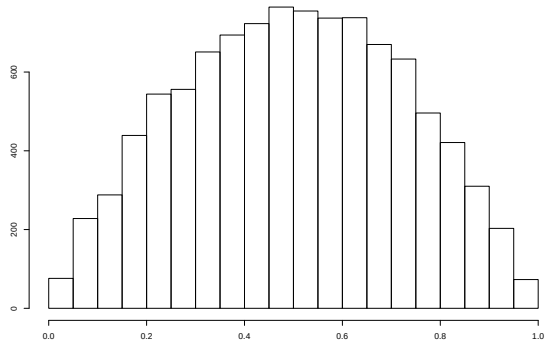
By default, Minitab tries its best to display an informative histogram.

Numerical Data Vocabulary: Modes

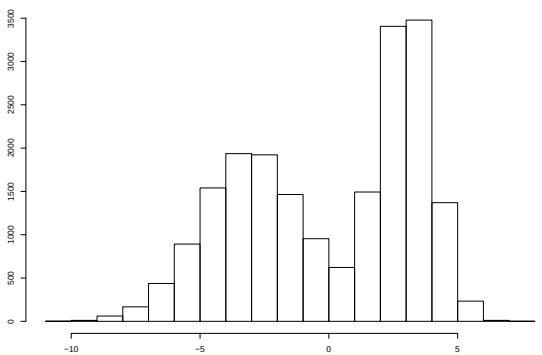
We say an histogram has a **mode** when it is peaked somewhere.



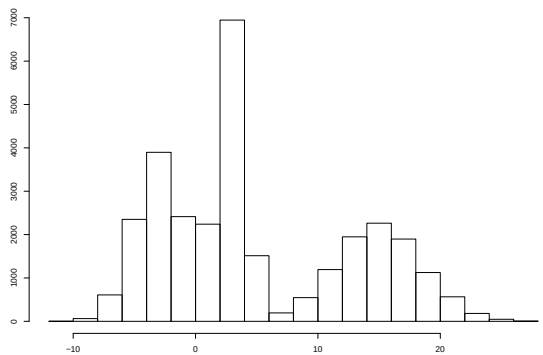
(a) No mode



(b) Unimodal (1 mode)



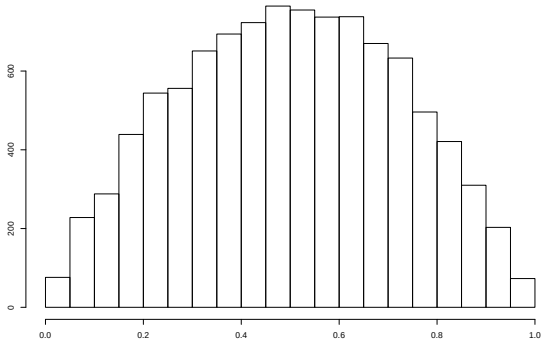
(c) Bimodal (2 modes)



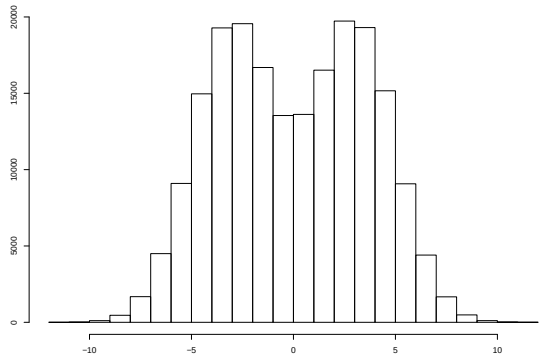
(d) Multimodal (> 2)

Numerical Data Vocabulary: Symmetry

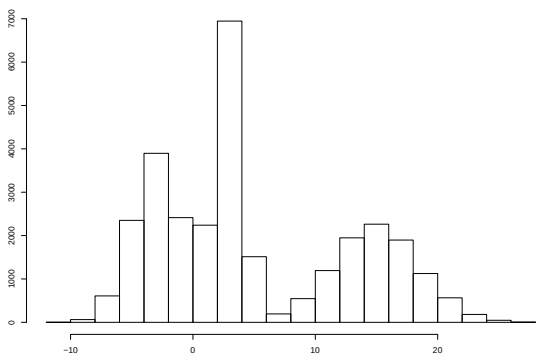
An histogram is **symmetric** if both sides of mode look the same.



(a) Symmetric



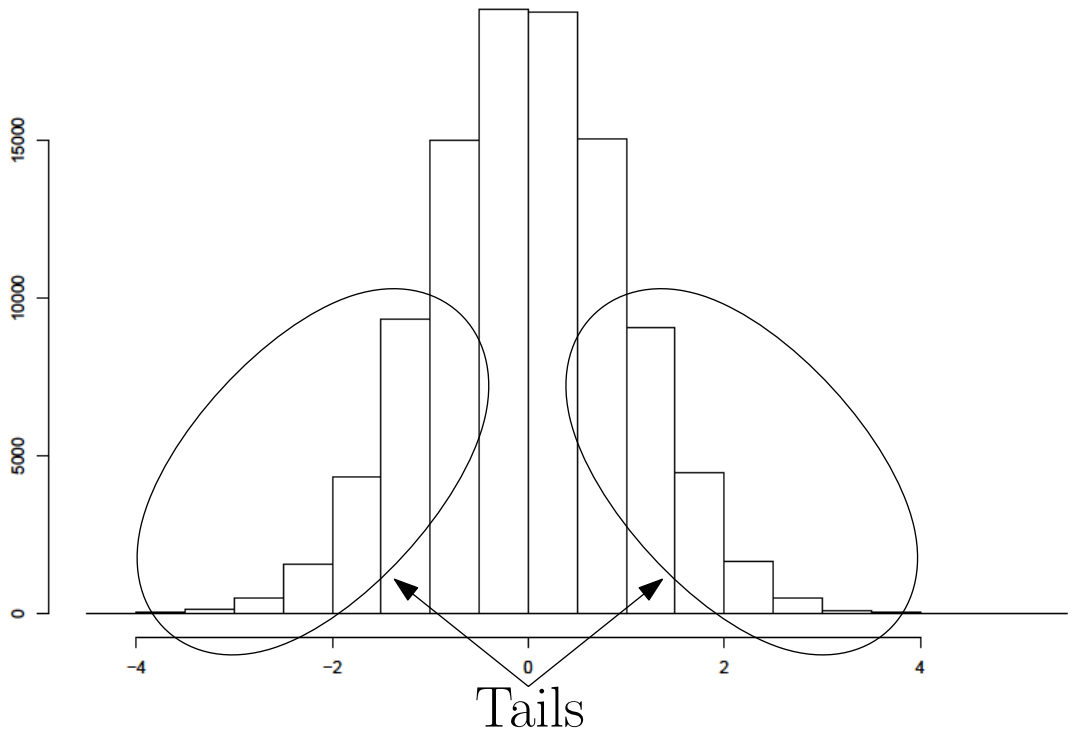
(b) Symmetric



(c) Not symmetric

Numerical Data Vocabulary: Tails

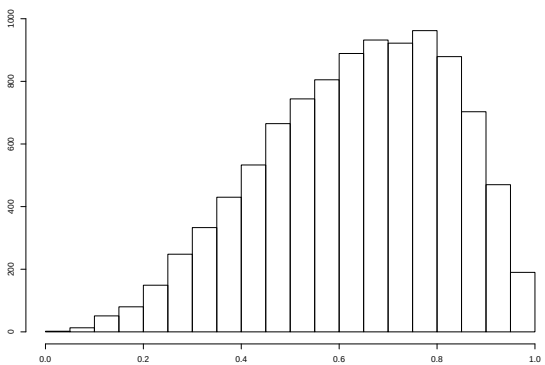
The **tails** of an histogram are the parts away from the center.



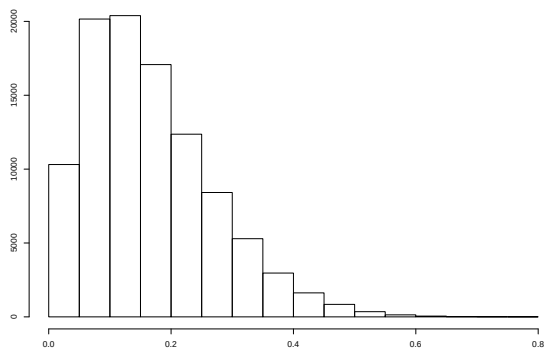
Numerical Data Vocabulary: Skewness

When an histogram is not symmetric, we can describe further its asymmetry by saying it is

- **Skewed left:** if the left tail is longer than the right tail.
- **Skewed right:** if the right tail is longer than the left tail.



(a) Skewed left

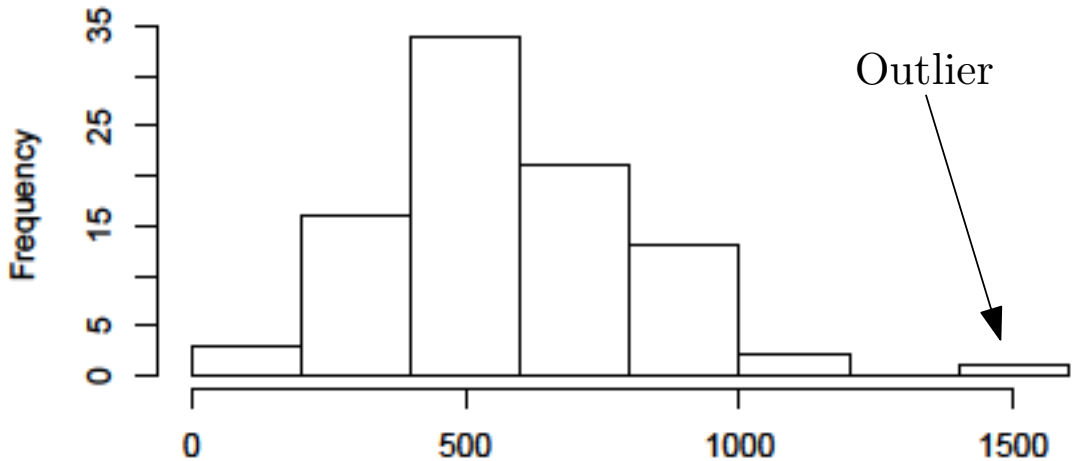


(b) Skewed right

Skewed left/right = the left/right tail stretches out longer.

Numerical Data Vocabulary: Outlier

An **outlier** is an observation that appears extreme relative to the rest of the data. (= Not conventional)



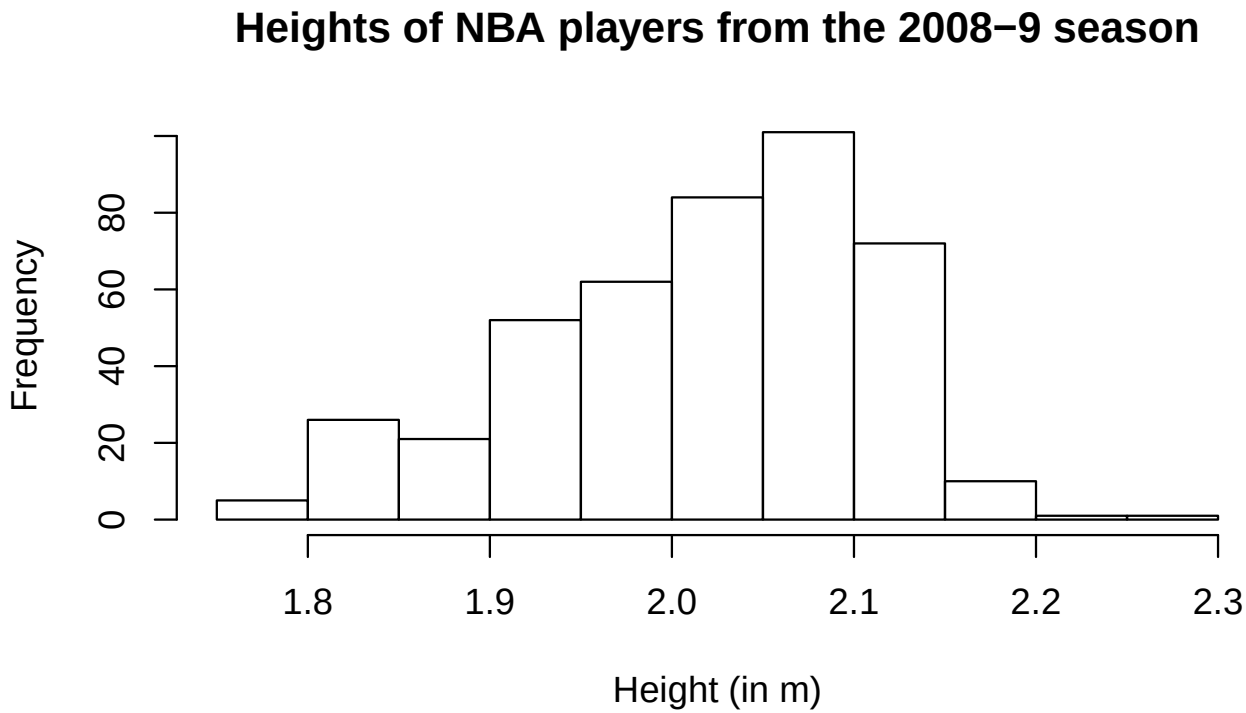
Examples:

- Extreme values in precision measurements for astrophysics
- Trolls' answers in online questionnaires

Sometimes outliers are informative, sometimes just annoying.

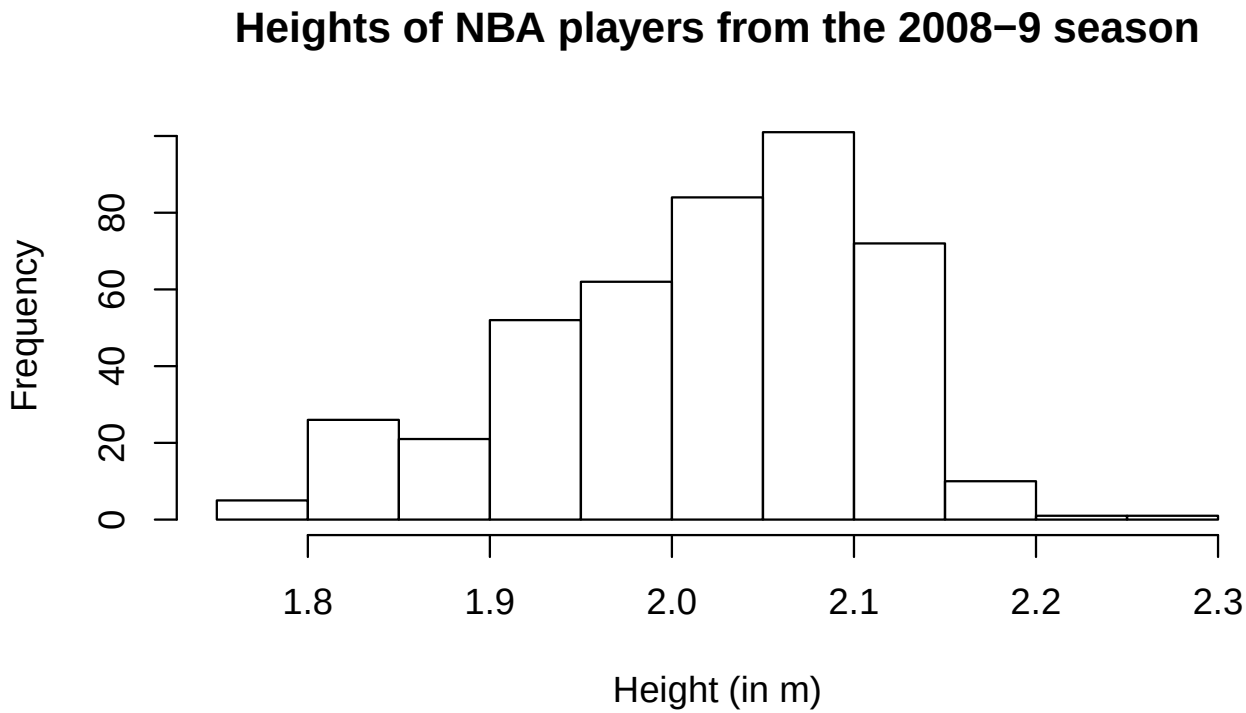
Practice

Describe this histogram.



Practice

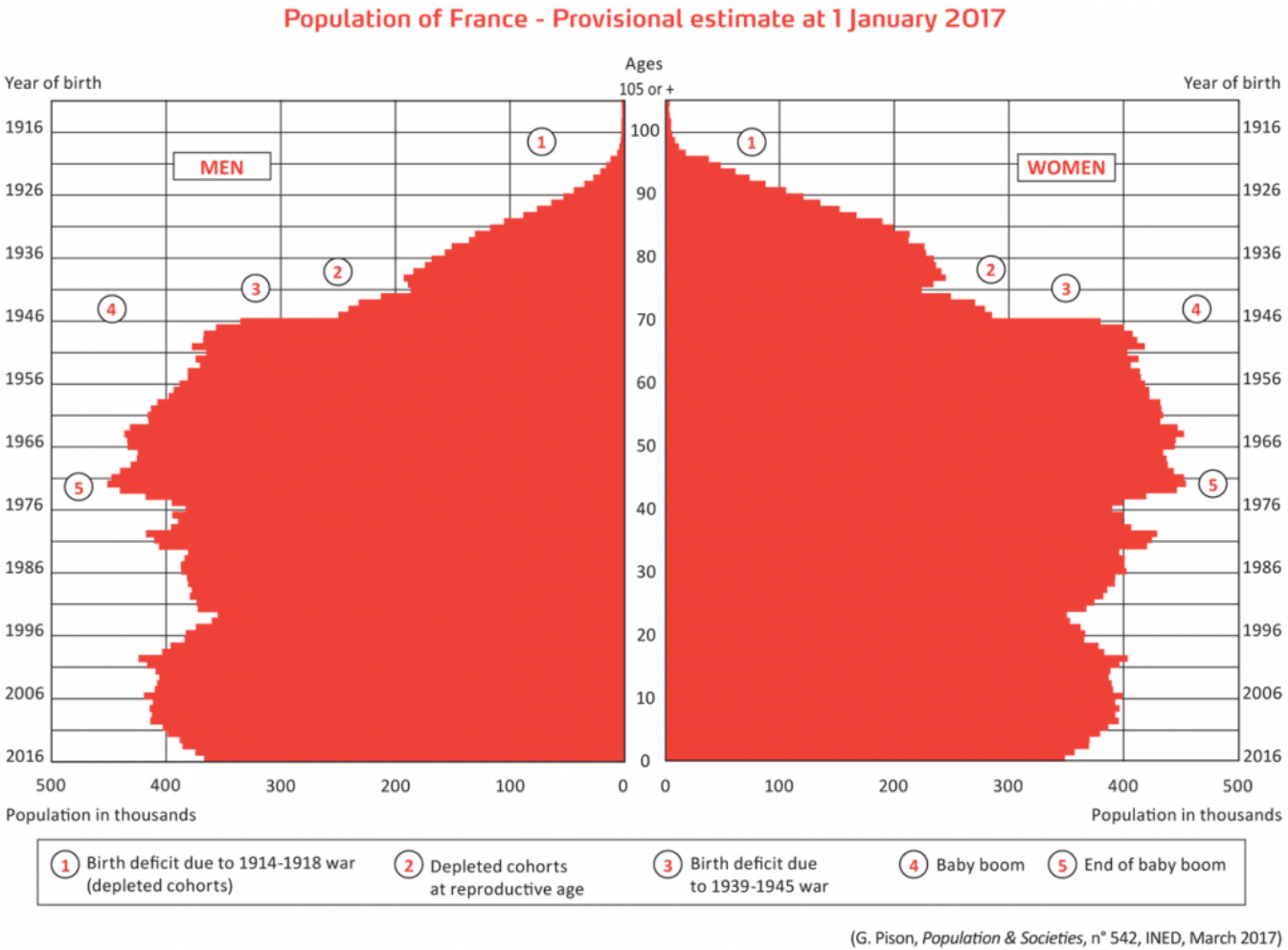
Describe this histogram.



Unimodal, skewed left, no outlier.

Practice

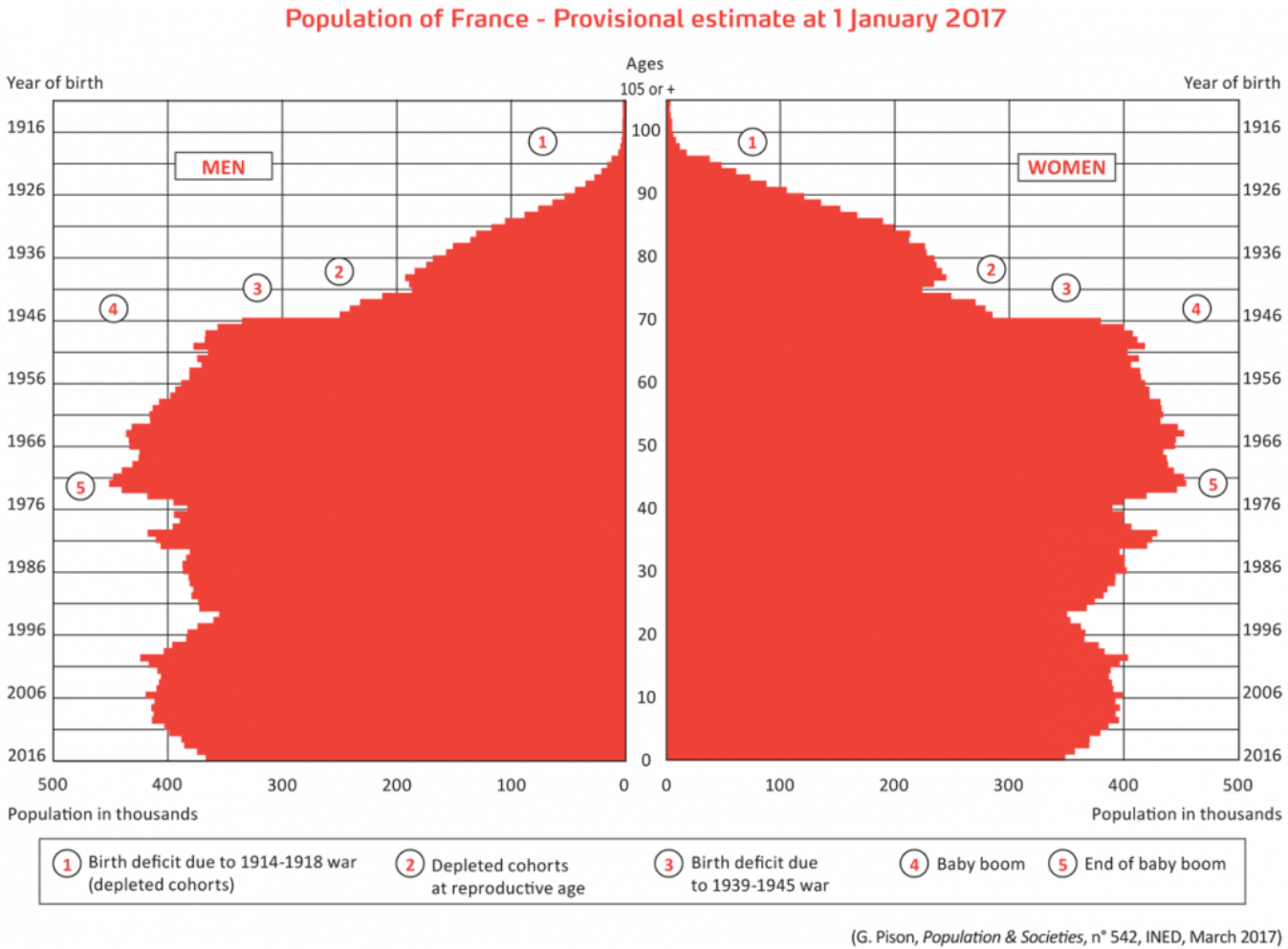
Describe these histograms.



(G. Pison, *Population & Societies*, n° 542, INED, March 2017)

Practice

Describe these histograms.



Multimodal, skewed right (up), no outlier.

The Center of a Data Set/Histogram/Distribution

Idea 1: The center of a distribution should be the data value "in the middle of the list of data" –there should be the same number of data values on each side of "the center". **MEDIAN**

The Center of a Data Set/Histogram/Distribution

Idea 1: The center of a distribution should be the data value "in the middle of the list of data" –there should be the same number of data values on each side of "the center". **MEDIAN**

4, 6, 0, -2, 45 → -2, 0, 4, 6, 45

Put the data in order, choose the middle number.

4, 6, 0, 3 -2, 45 → -2, 0, 3, 4, 6, 45

The median is 3.5

If there is no "middle number", average the two in the middle of the ordered list

The Center of a Data Set/Histogram/Distribution

Idea 1: The center of a distribution should be the data value "in the middle of the list of data" –there should be the same number of data values on each side of "the center". **MEDIAN**

4, 6, 0, -2, 45 → -2, 0, 4, 6, 45

Put the data in order, choose the middle number.

4, 6, 0, 3 -2, 45 → -2, 0, 3, 4, 6, 45

The median is 3.5

If there is no "middle number", average the two in the middle of the ordered list

Idea 2: The center of a distribution must take into account the data values themselves, not just the order they are in. **MEAN**

The Center of a Data Set/Histogram/Distribution

Idea 1: The center of a distribution should be the data value "in the middle of the list of data" –there should be the same number of data values on each side of "the center". **MEDIAN**

4, 6, 0, -2, 45 → -2, 0, 4, 6, 45

Put the data in order, choose the middle number.

4, 6, 0, 3 -2, 45 → -2, 0, 3, 4, 6, 45

The median is 3.5

If there is no "middle number", average the two in the middle of the ordered list

Idea 2: The center of a distribution must take into account the data values themselves, not just the order they are in. **MEAN**

4,5,6 and 4,5, 990 have the same median

$$\frac{4+5+6}{3} = 5 \qquad \frac{4+5+990}{3} = 333$$

Add the data values, divide by how many there are.

Which Center Idea Do I Use?

$\{2, 3, 4, 5, 6, 7\}$

Median: 4.5
Mean: 4.5

$\{2, 3, 4, 5, 6, 70\}$

Median: 4.5
Mean: 15

$\{2, 3, 4, 5, 6, 700\}$

Median: 4.5
Mean: 120

Which Center Idea Do I Use?

$\{2, 3, 4, 5, 6, 7\}$

Median: 4.5
Mean: 4.5

$\{2, 3, 4, 5, 6, 70\}$

Median: 4.5
Mean: 15

$\{2, 3, 4, 5, 6, 700\}$

Median: 4.5
Mean: 120

Moral: The median is resistant to outliers and skew. For this reason, it is called "robust".

Which Center Idea Do I Use?

$\{2, 3, 4, 5, 6, 7\}$

Median: 4.5
Mean: 4.5

$\{2, 3, 4, 5, 6, 70\}$

Median: 4.5
Mean: 15

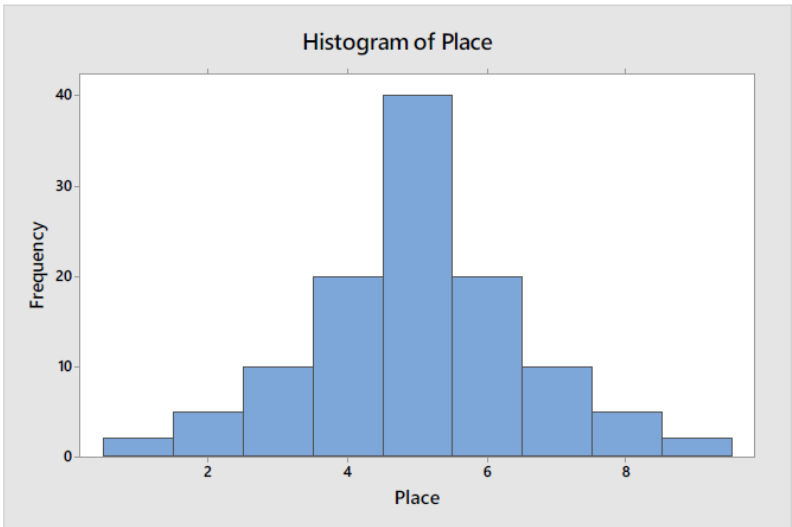
$\{2, 3, 4, 5, 6, 700\}$

Median: 4.5
Mean: 120

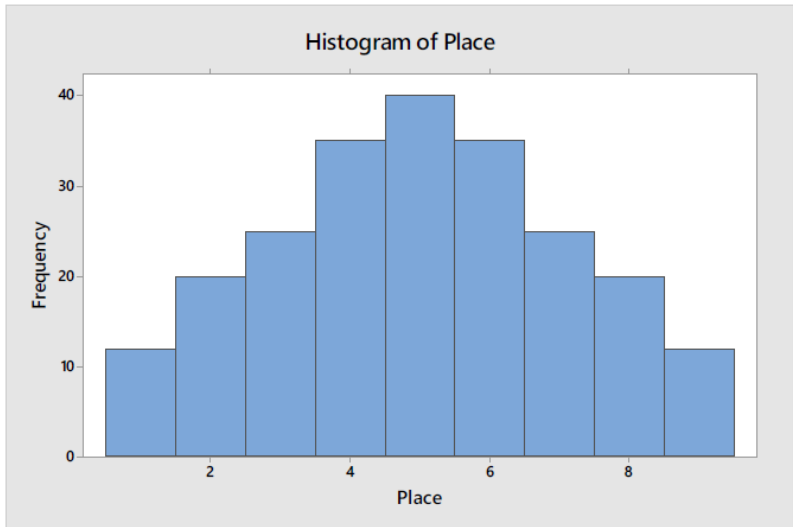
Moral: The median is resistant to outliers and skew. For this reason, it is called "robust".

Median: Use for asymmetric distributions or data with outliers.
Mean: Use for symmetric distributions without outliers.

The Spread of a Distribution



Most values in here



Most values in here

Three Ideas for Measuring Spread

The Range:

(maximum value in data set) – (minimum value in data set)

- *Pros*: Easy to calculate, gives a sense of the total span of the data
- *Cons*: Summarizes the data using only two of the data points (the extremes); not resistant to outliers

Three Ideas for Measuring Spread

The Range:

(maximum value in data set) – (minimum value in data set)

- *Pros*: Easy to calculate, gives a sense of the total span of the data
- *Cons*: Summarizes the data using only two of the data points (the extremes); not resistant to outliers

The Interquartile Range (IQR):

(upper quartile) – (lower quartile)

Three Ideas for Measuring Spread

The Range:

(maximum value in data set) – (minimum value in data set)

- *Pros*: Easy to calculate, gives a sense of the total span of the data
- *Cons*: Summarizes the data using only two of the data points (the extremes); not resistant to outliers

The Interquartile Range (IQR):

(upper quartile) – (lower quartile)

- *Pros*: Resistant to skew and outliers, easy to communicate
- *Cons*: Takes work to calculate, no universal definition, not widely known

Three Ideas for Measuring Spread

The Range:

(maximum value in data set) – (minimum value in data set)

- *Pros*: Easy to calculate, gives a sense of the total span of the data
- *Cons*: Summarizes the data using only two of the data points (the extremes); not resistant to outliers

The Interquartile Range (IQR):

(upper quartile) – (lower quartile)

- *Pros*: Resistant to skew and outliers, easy to communicate
- *Cons*: Takes work to calculate, no universal definition, not widely known

Range: 19 - 2 = 17

IQR: 13 – 3 = 10

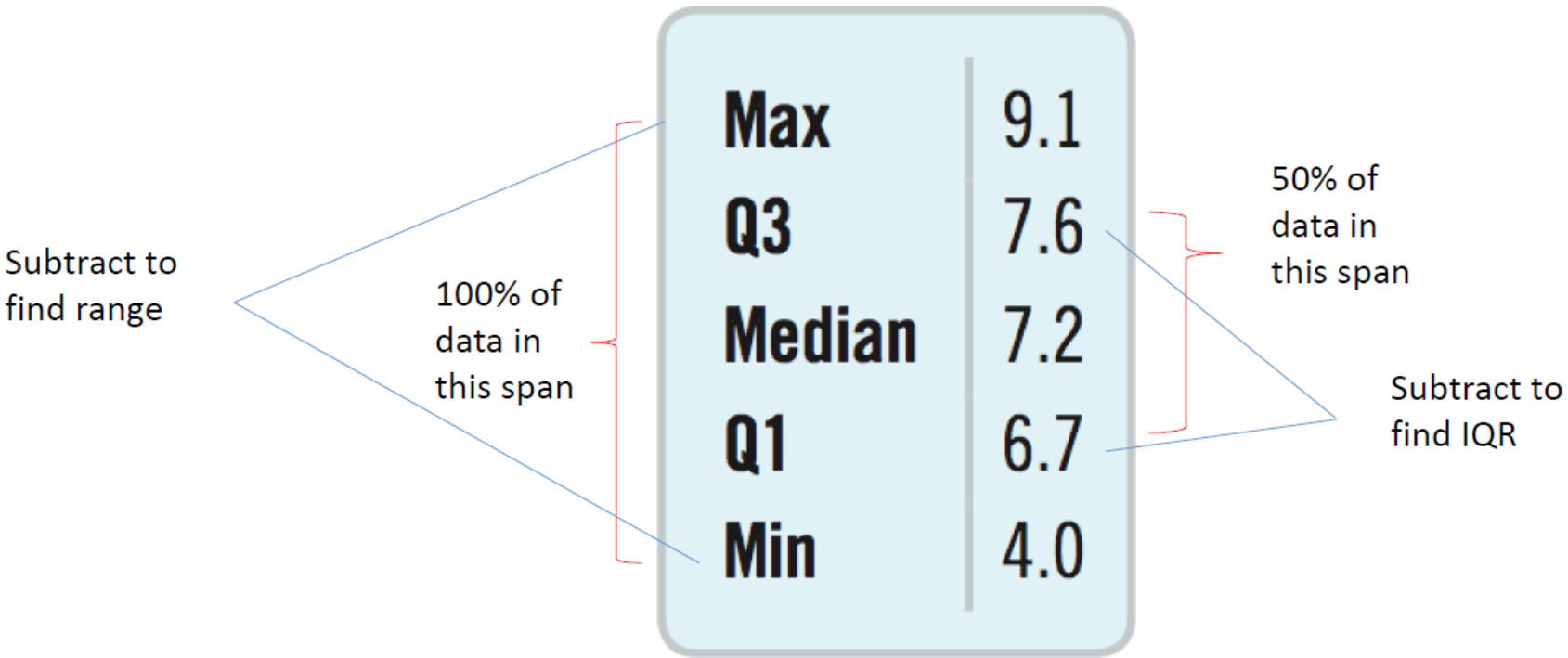
Upper quartile (Q3): Find the median of the data to the right of Q2 (here: 13)

$$\{2, 2, 3, 3, 5, 5, 5, 7, 11, 13, 13, 13, 13, 17, 19\}$$

Lower quartile (Q1): Find the median of the data to the left of Q2 (here: 3)

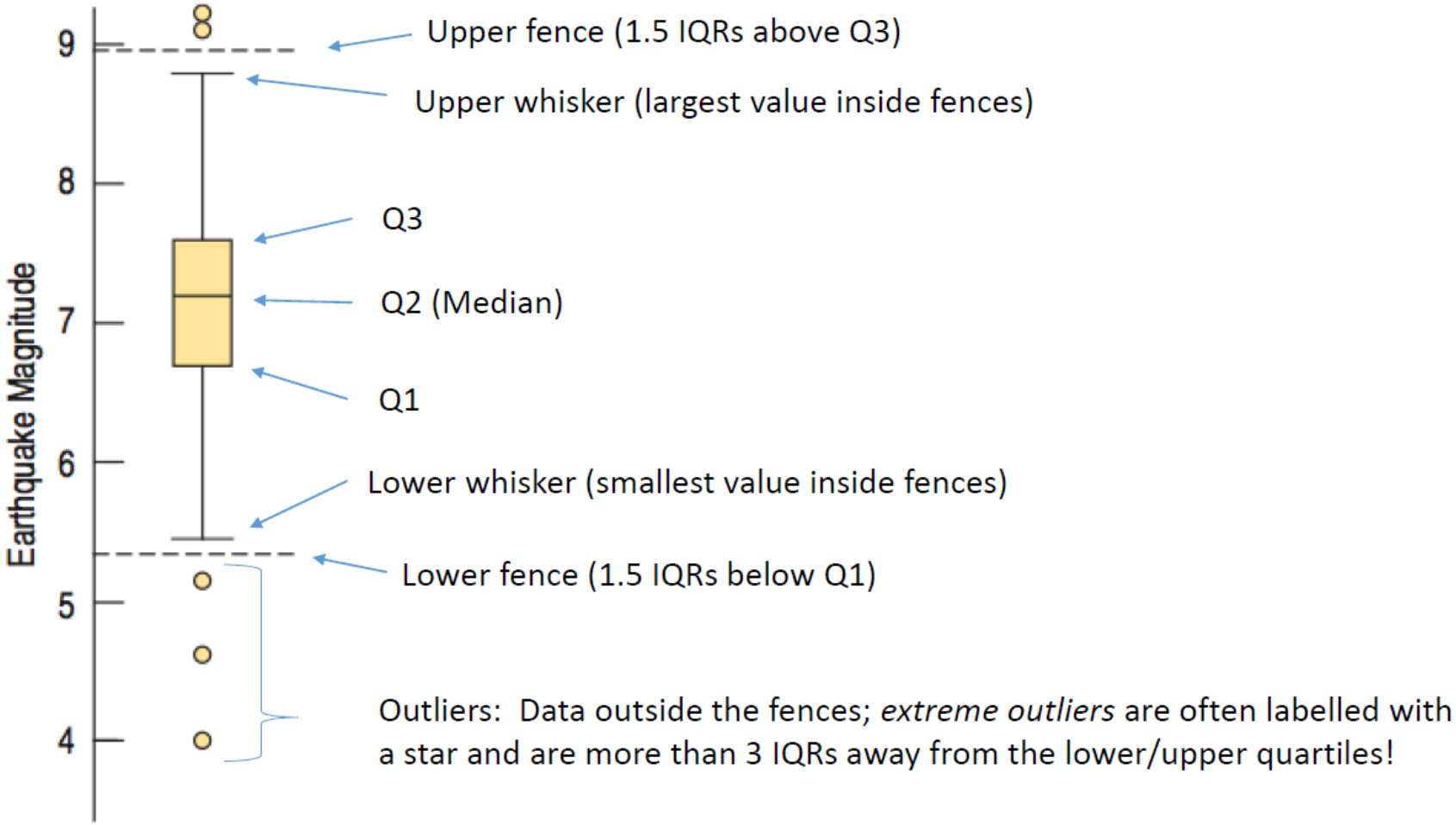
Middle quartile (Q2, Median): 7

Bringing It All Together: The Five Number Summary



The five number summary for the magnitudes of 221 earthquakes

Visualizing the Five-Number Summary: The Boxplot



Three Ideas for Measuring Spread

The sample **standard deviation** of $\mathbf{x} = (x_1, \ldots, x_n)$ is defined as:

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Three Ideas for Measuring Spread

The sample **standard deviation** of $\mathbf{x} = (x_1, \ldots, x_n)$ is defined as:

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Sum of squares: to get positive values
- Square root: to undo squaring action and have s with same units as the x_i's.
- $(n-1)$ instead of n: explained later in the course.

Three Ideas for Measuring Spread

The sample **standard deviation** of $\mathbf{x} = (x_1, \ldots, x_n)$ is defined as:

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Sum of squares: to get positive values
- Square root: to undo squaring action and have s with same units as the x_i's.
- $(n-1)$ instead of n: explained later in the course.

- *Pros*: Very popular notion of spread; easy to handle mathematically
- *Cons*: Sensitive to outliers and skew; hard to explain to non-statisticians.

Explaining Your Data: SOCS

Shape: Modality (uniform, unimodal, bimodal, multimodal)
Symmetry (symmetric, skewed left, skewed right)

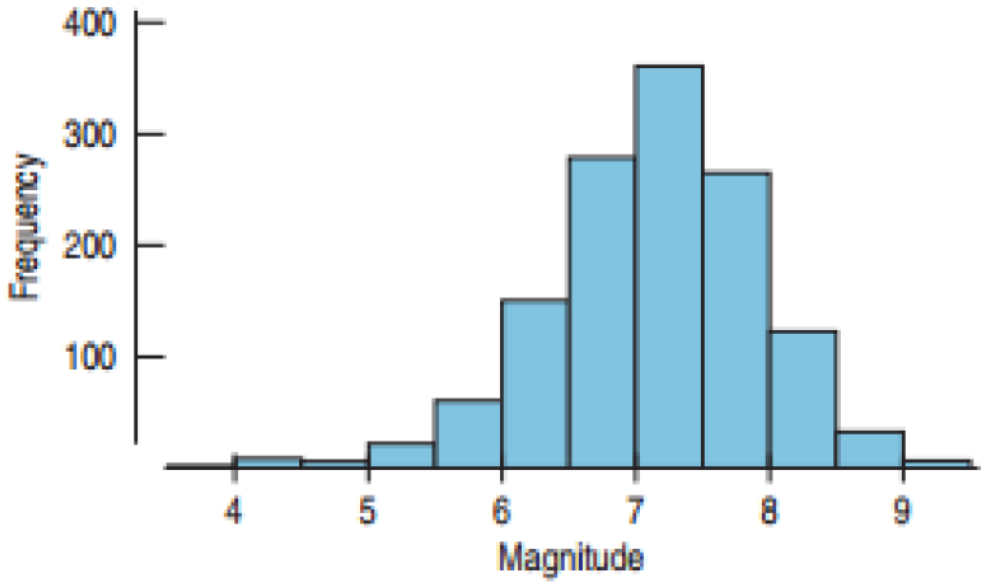
Outliers: Where are they? Do they have meaning?

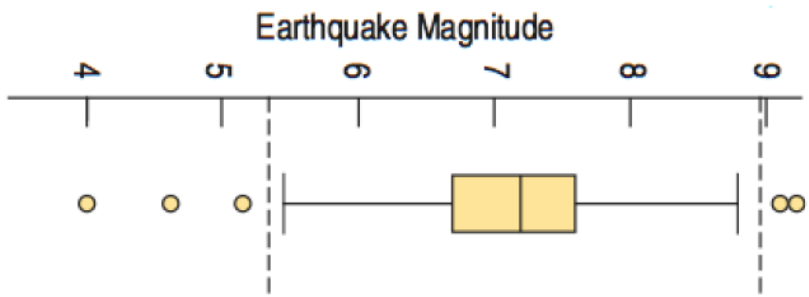
Center: Mean for unimodal, symmetric data with no outliers, median else

Spread: Standard deviation (SD) for unimodal, symmetric data with no outliers, IQR else

Practice

Describe Shape, Outliers, Center, and Spread of this dataset.

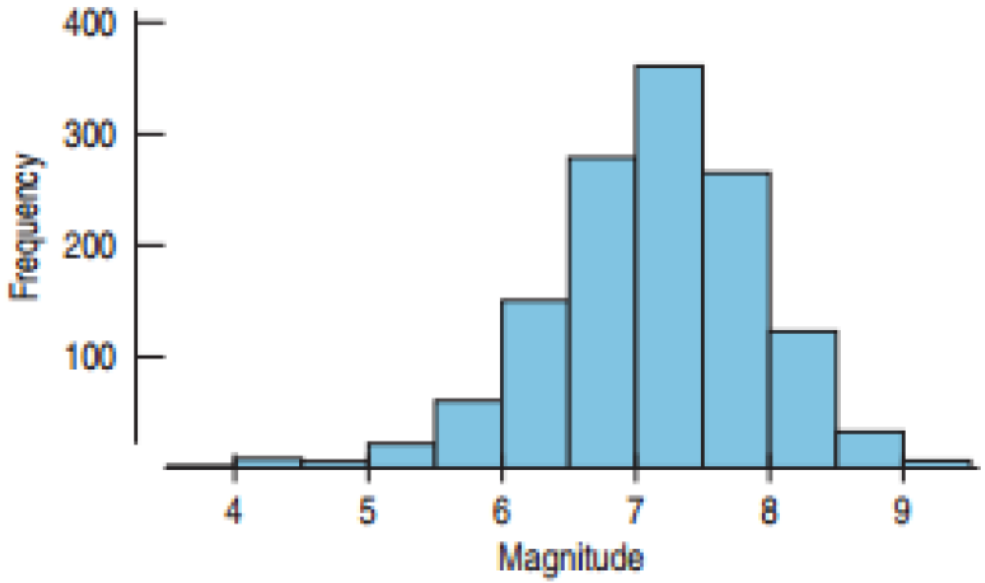




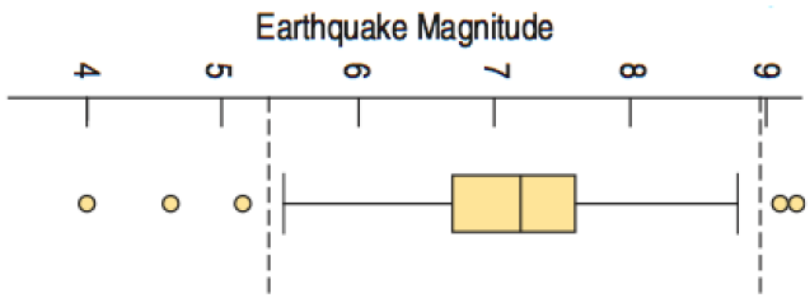
Max	9.1
Q3	7.6
Median	7.2
Q1	6.7
Min	4.0

Practice

Describe Shape, Outliers, Center, and Spread of this dataset.



Max	9.1
Q3	7.6
Median	7.2
Q1	6.7
Min	4.0



Answer: Unimodal and slightly skewed left, no outliers, width median $Q_2 = 7.2$ and spread $IQR = 7.6 - 6.7 = 0.9$.

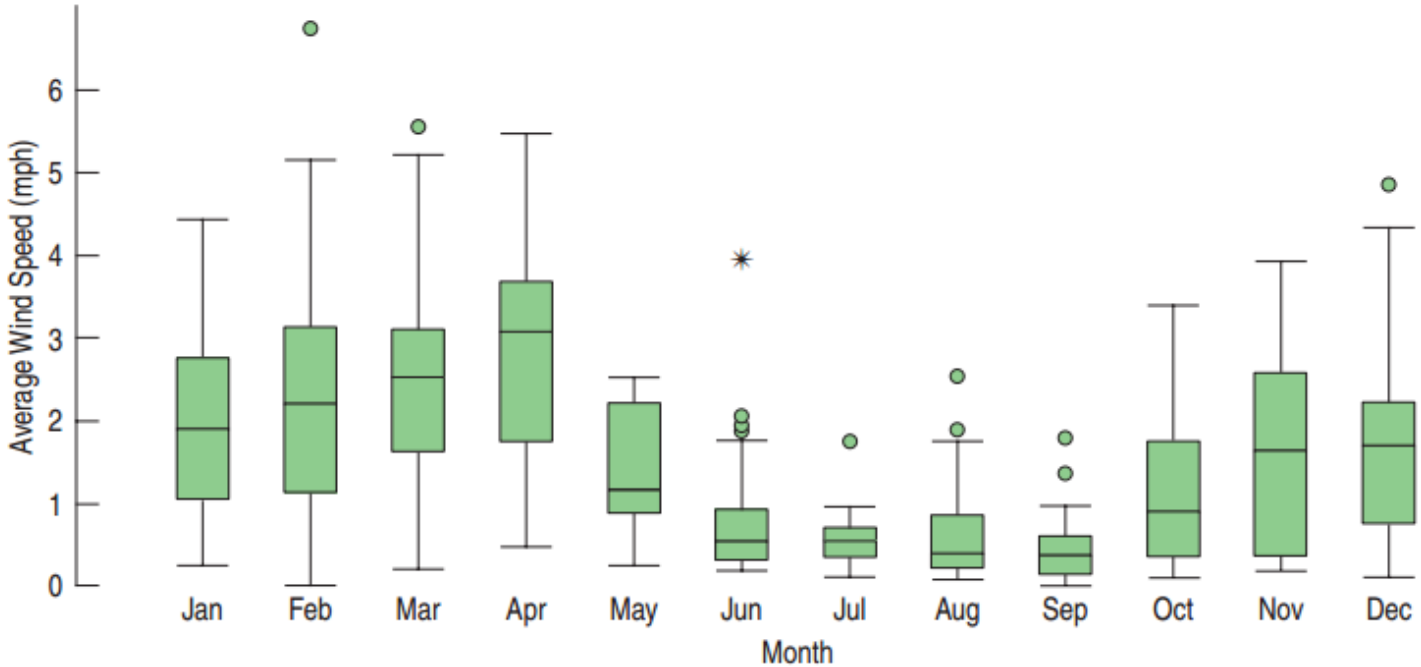
Comparing Distributions

Comparing Distributions

When Comparing Histograms:

- Make sure the scales are the same (both axes!)
- Discuss SOCS (Shape, Outliers, Center, Spread)
- Use the same measures of center and spread for both
- Also create five-number summaries
- Write your conclusions so they do not generalize beyond the sample

Comparing Boxplots: Tons of Info in One Place



Massachusetts Wind Speed Boxplots for Each Month in 2011

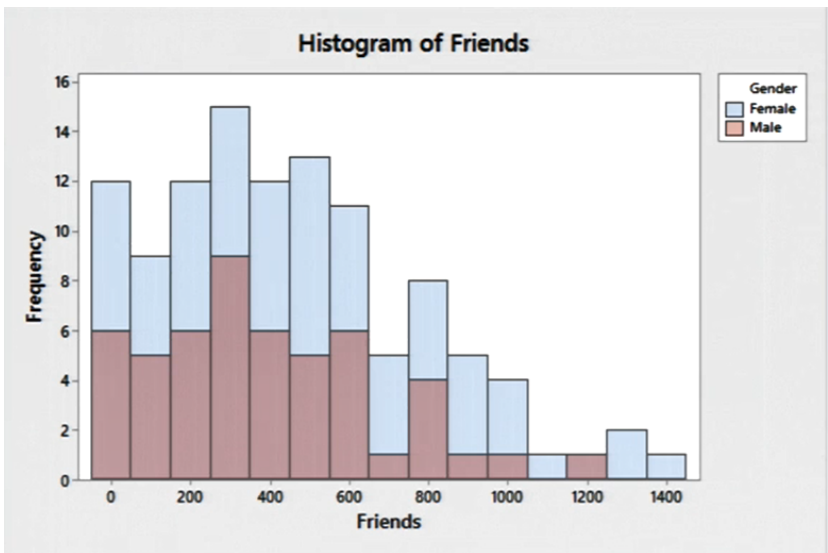
What to think about:

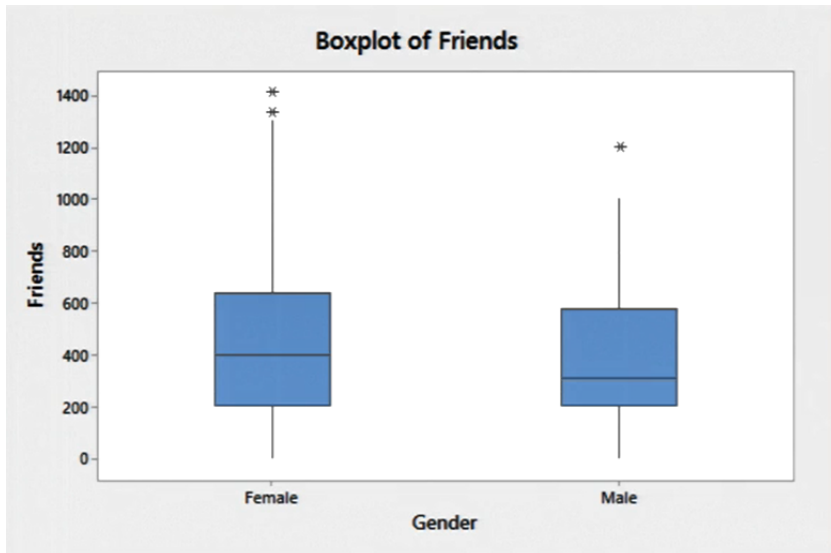
- Which groups have the highest medians, the greatest IQRs
- Where the middle 50% of data is for each group
- What things look like when outliers are minimized from our view
- The context-dependent nature of outliers

Facebook Friends VS Gender

Descriptive Statistics: Friends

Variable	Gender	Mean	StDev	Minimum	Q1	Median	Q3	Maximum	Range	IQR
Friends	Female	447.3	326.8	0.0	200.0	400.0	640.0	1410.0	1410.0	440.0
	Male	385.9	282.9	0.0	200.0	310.0	579.0	1200.0	1200.0	379.0





Taking Outliers Seriously

- Try researching a particular outlier. Does it really belong?
- Some outliers are errors.
- Run analysis with and without your outliers.
- In the end, state what you are doing with outliers and why.

Taking Outliers Seriously

- Try researching a particular outlier. Does it really belong?
- Some outliers are errors.
- Run analysis with and without your outliers.
- In the end, state what you are doing with outliers and why.

Practice: Give an example of a quantitative variable and an outlier in that data.

Taking Outliers Seriously

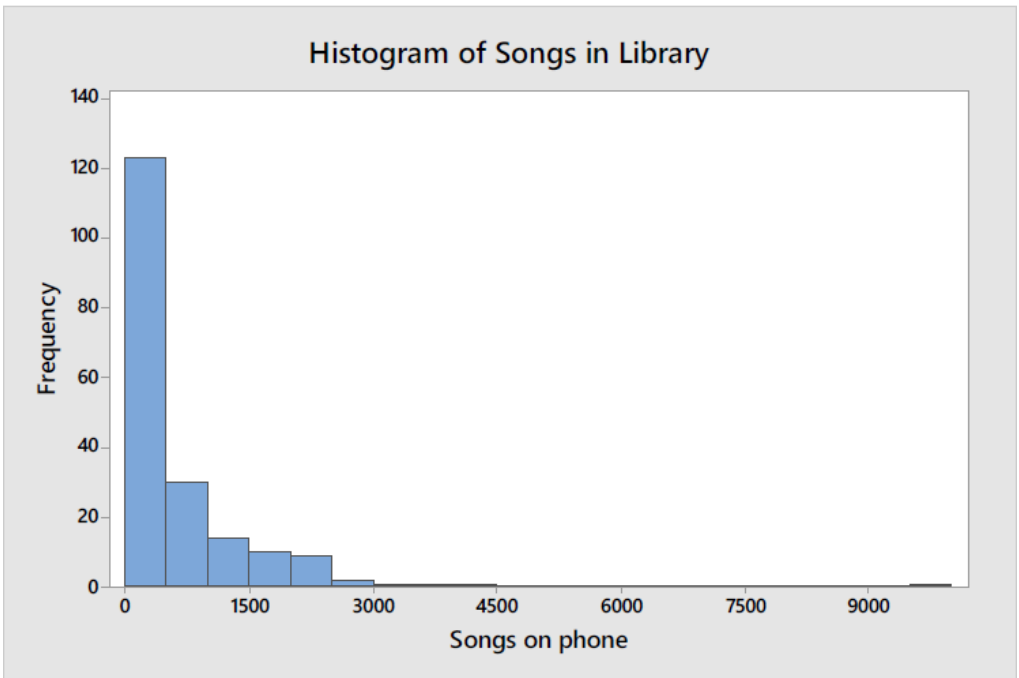
- Try researching a particular outlier. Does it really belong?
- Some outliers are errors.
- Run analysis with and without your outliers.
- In the end, state what you are doing with outliers and why.

Practice: Give an example of a quantitative variable and an outlier in that data.
An answer: Number of different types of cheese in a country, France

Re-expressing Highly Skewed Data

192 students were asked:

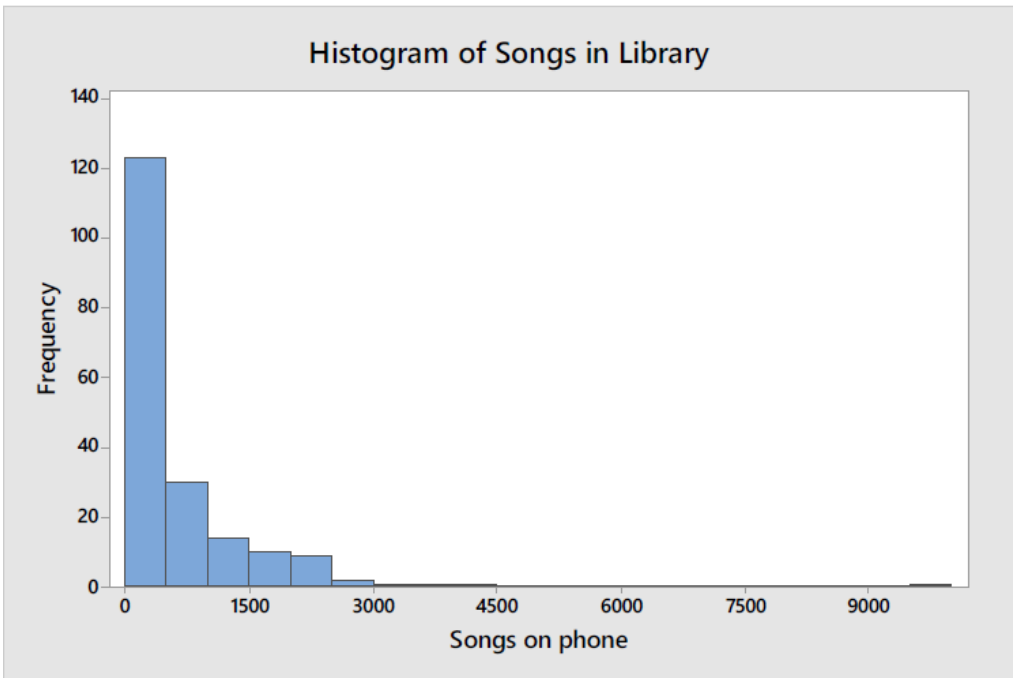
How many songs are in your digital music library?



Re-expressing Highly Skewed Data

192 students were asked:

How many songs are in your digital music library?



Examples where data are thought of on a re-expressed scale:

- Earthquake magnitudes

$$M = \log_e 10(A/A_0)$$

- pH of substance

$$pH = -\log(a_{H^+})$$

- Grades on a bad test

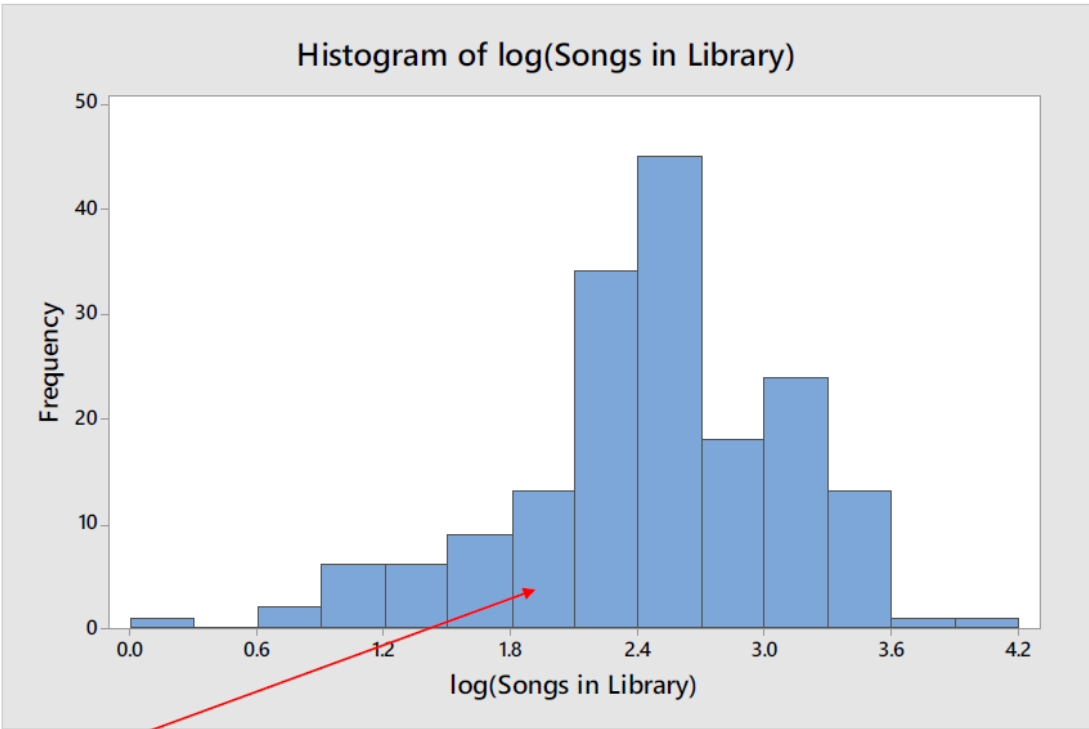
$$G = 10\sqrt{\text{Original grade}}$$

Re-expressing Highly Skewed Data

We display the new (made-up) variable

$$\log_{10}(\text{Song Count}).$$

Songs in Library	log(Songs in Library)
250	2.39794
5	0.69897
800	2.90309
500	2.69897
850	2.92942
430	2.63347
200	2.30103
600	2.77815
0	*
130	2.11394
1000	3.00000
1000	3.00000
400	2.60206
1500	3.17609
20	1.30103
1500	3.17609
872	2.94052



People with $1.8 \leq \log_{10}(\text{Song Count}) < 2.1$

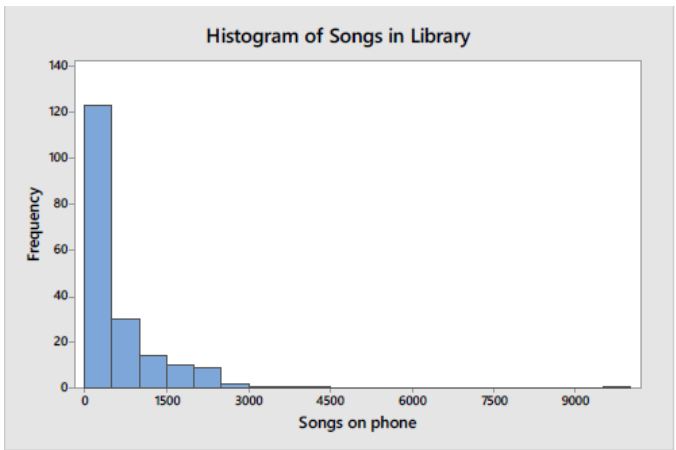
People with $10^{1.8} \leq \text{Song Count} < 10^{2.1}$

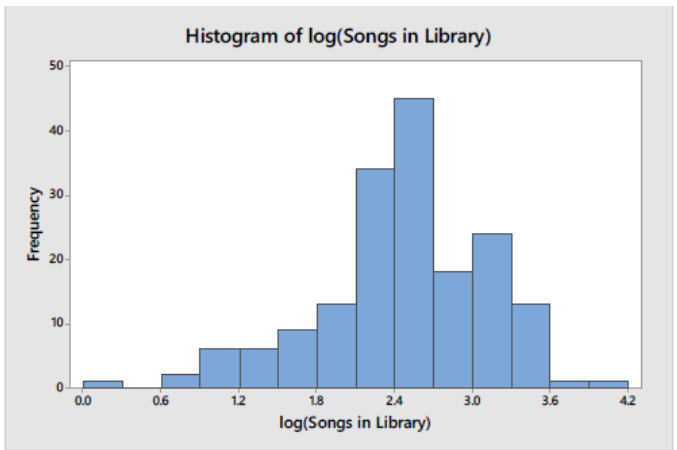
Why Re-express Skewed Data?

- To make it visually more appealing
- To create a more commonly-shaped histogram (the value of this is apparent later)
- To get the "lens of analysis correct"

Why Re-express Skewed Data?

- To make it visually more appealing
- To create a more commonly-shaped histogram (the value of this is apparent later)
- To get the "lens of analysis correct"

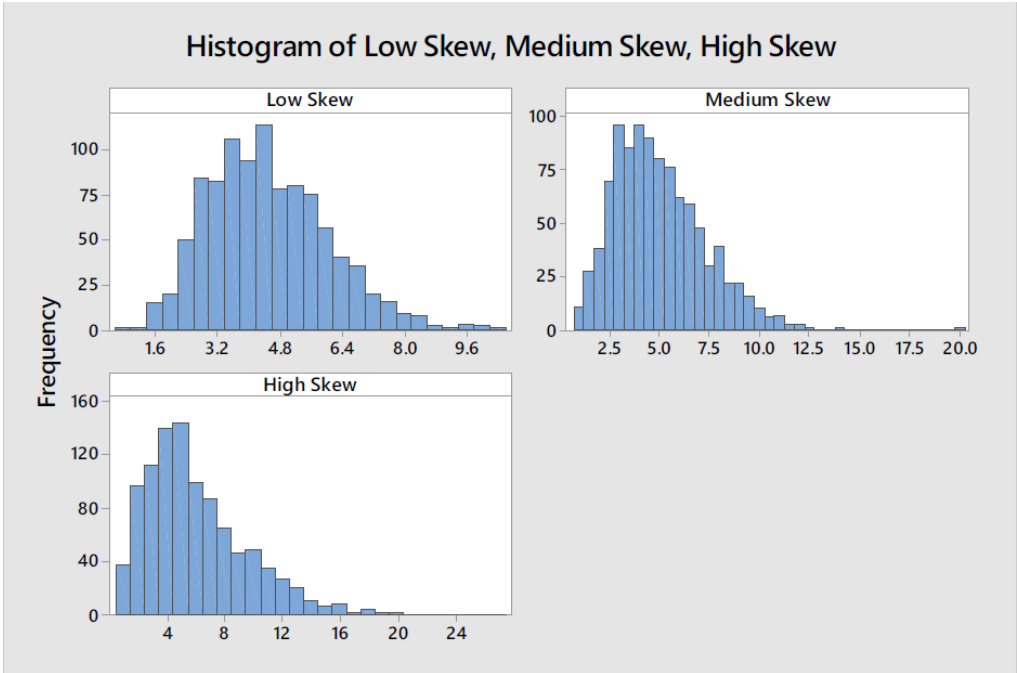




"How many songs?" (1, 2, 3 ...)

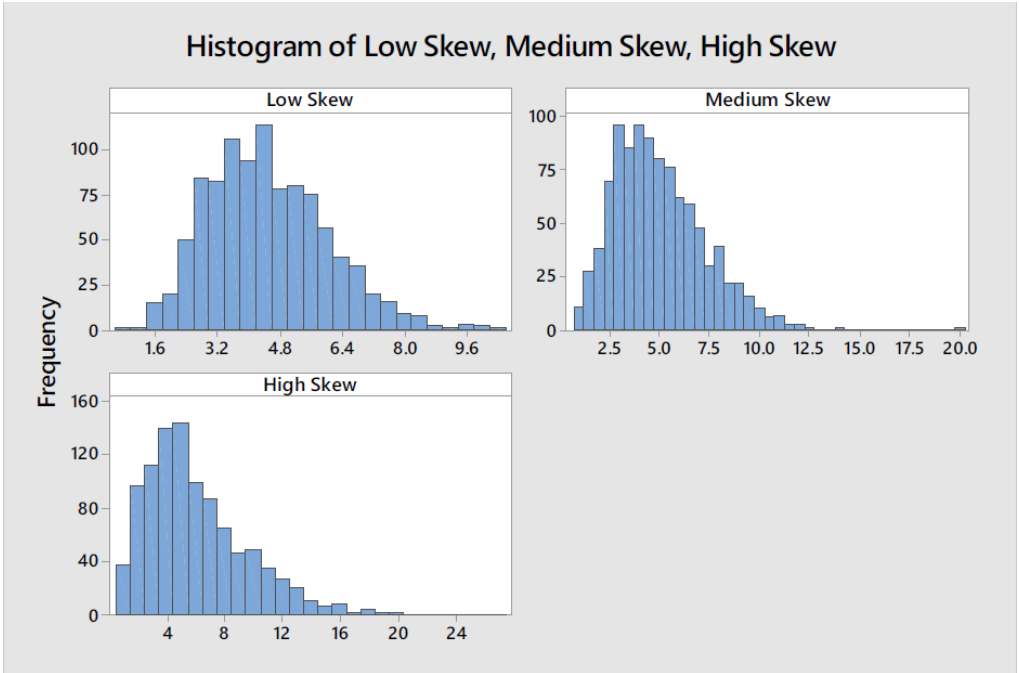
"What order of magnitude?" (1, 10, 100, ...)

Other Thoughts on Skewed Data



Discuss with a classmate: Will the mean be less than, about equal to, or greater than the median in each distribution?

Other Thoughts on Skewed Data



Discuss with a classmate: Will the mean be less than, about equal to, or greater than the median in each distribution?

Variable	Mean	Median
Low Skew	4.5161	4.4005
Medium Skew	5.0217	4.6575
High Skew	6.089	5.334

As the graph skews to the right, the mean becomes larger than the median: the mean is pulled right by the larger values in the data set.

What you Should Do After the Lecture

- Buy textbook
- Start Homework 1 (Due next Tuesday in class)
  Print it, and make sure to write down your full name, the PID and your section.
- Install Minitab
- Read the course syllabus and look at the calendar