Stat 201: Introduction to Statistics

Standard 5-6: Numerical Summaries

Chapter Two

Summaries

From *Naked Statistics:* Descriptive Statistics

- "Descriptive Statistics are the numbers and calculations we use to summarize raw data"

- "Descriptive statistics give us insight into phenomena that we care about."

- "…Raw data would take a while to digest, given that Jeter has played seventeen seasons (9,868 at bats.) … a career batting average of .313 is a descriptive statistic or a 'summary statistic'"

From *Naked Statistics:* Descriptive Statistics

- “My quick calculation is technically correct and yet totally wrong in terms of the question I set out to answer”

- “A spreadsheet with the name and income history of every American would contain all the information we could ever want about the economic health of the country – yet it would also be so unwieldy as to tell us nothing at all… So we simplify”

From *Naked Statistics:* Descriptive Statistics

- “The good news is that these descriptive statistics give us a manageable and meaningful summary of the underlying phenomenon… The bad news is that any simplification invites abuse. Descriptive statistics can be like online dating profiles: technically accurate and yet pretty darn misleading.”

From *Naked Statistics: Descriptive Statistics*

- "The standard deviation is the descriptive statistic that allows us to assign a single number to this dispersion around the mean."

From *Naked Statistics: Descriptive Statistics*

- “The Democrats, who engineered this tax increase, pointed out (correctly) that the state income tax rate was increased by 2 percentage points (from 3 percent to 5 percent.) The Republicans pointed out (also correctly) that the state income tax rate had been raised by 67 percent… the Republican description more accurately conveys the impact of the tax change, since what I’m going to have to pay to the government – the amount that I care about, as opposed to the way it is calculated –really has gone up by 67 percent”

From *Naked Statistics: Deceptive Statistics*

- To anyone who has ever contemplated dating, the phrase ‘he’s got a great personality’ usually sets off alarm bells, not because the description is necessarily wrong, but for what it may not reveal, such as the fact that the guy has a prison record or that his divorce is ‘not entirely final.’”
- “The statement is not a lie per se, meaning that it wouldn’t get you convicted of perjury, but it still could be so inaccurate as to be untruthful.”

From *Naked Statistics: Deceptive Statistics*

- “The descriptive statistic that we choose to use (or not to use) will have a profound impact on the impression that we leave. Someone with nefarious motives can use perfectly good facts and figures to support entirely disputable or illegitimate conclusions.”

Measures of Central Tendency

Measure	Computation	Interpretation	When to Use
Mean Statistic: $\bar{x}$ Parameter: μ	$\bar{x} = \frac{\sum x}{n}$	Center of Gravity	Use for quantitative data when the distribution is roughly symmetric
Median*	The point halfway through the data when it is arranged in ascending order.	The point which splits the data in half.	Use for quantitative data when the distribution is skewed
Mode	We report the observation with the highest frequency	Most frequent observation	When the most frequent observation is the desired measure or when data is qualitative.

* Denotes robustness to outliers – to be used when data is not bell-shaped

Measures of Dispersion

Measure	Computation	Interpretation
Range	Max – Min	The difference between the largest and smallest data point
Standard Deviation Statistic: s Parameter: σ	$\sqrt{Variance}$	The square root of the mean of squared deviations from the mean in the original units – this usually makes the standard deviation easier to interpret
Variance Statistic: s^2 Parameter: σ^2	$\frac{\sum(x-\overline{x})^2}{n-1}$	The square root of the mean of squared deviations from the mean in units squared
IQR*	$Q_U - Q_L$	The range of the middle 50%

* Denotes robustness to outliers – to be used when data is not bell-shaped

Walkthrough

Summarizing Qualitative Data: Frequencies

- A **Frequency Distribution** lists each category of the variable and the number or proportion of occurrences for each category of data.

Summarizing Qualitative Data: Frequencies

- **Class Frequency** is the number of occurrences for each class of variable of interest

- **Relative Frequency** is the proportion of observations of a class among all observations of the variable of interest

NOTE!

- **Relative Frequency** is the proportion of observations within a category and is found using the following formula

$$Relative\ Freq. = \frac{frequency}{sum\ of\ all\ frequencies}$$

Relative Frequency is also referred to as a **proportion, $\widehat{\boldsymbol{p}}$ or $\boldsymbol{\rho}$.** This will be really important later in the semester!

Example

- The 2012 South Carolina Republican Primary was held on January 21st. Newt Gingrich, Mitt Romney, Rick Santorum, Ron Paul, Herman Cain, Rick Perry, Jon Huntsman, Michele Bachmann and Gary Johnson were on the ballet for voters to choose from.

Example

Candidate Chosen	Class Frequency - the number of times candidate ‘x’ was voted for	Relative Frequency- the proportion of times candidate ‘x’ was voted for
Class = X = Bachmann	491	
Class = X = Cain	6,338	
Class = X = Gingrich	244,065	
Class = X = Huntsman	1,173	
Class = X = Johnson	211	
Class = X = Paul	78,360	
Class = X = Perry	2,534	
Class = X = Romney	168,123	
Class = X = Santorum	102,475	
TOTAL	603,770	

Example

Candidate Chosen	Class Frequency - the number of times candidate ‘x’ was voted for	Relative Frequency- the proportion of times candidate ‘x’ was voted for
Class = X = Bachmann	491	491/603,770 = .0008
Class = X = Cain	6,338	6,338/603,770 = .0105
Class = X = Gingrich	244,065	244,065/603,770 = .4042
Class = X = Huntsman	1,173	1,173/603,770 = .0019
Class = X = Johnson	211	211/603,770 = .0003
Class = X = Paul	78,360	78,360/603,770 = .1298
Class = X = Perry	2,534	2,534/603,770 = .0042
Class = X = Romney	168,123	168,123/603,770 = .2785
Class = X = Santorum	102,475	102,475/603,770 = .1697
TOTAL	603,770	~1

Example

Candidate Chosen	Class Frequency - the number of times candidate 'x' was voted for	Relative Frequency- the proportion of times candidate 'x' was voted for
Class = X = Bachmann	491	491/603,770 = .0008 = .08%
Class = X = Cain	6,338	6,338/603,770 = .0105 = 1.05%
Class = X = Gingrich	244,065	244,065/603,770 = .4042 = 40.42%
Class = X = Huntsman	1,173	1,173/603,770 = .0019 = .19%
Class = X = Johnson	211	211/603,770 = .0003 = .03%
Class = X = Paul	78,360	78,360/603,770 = .1298 = 12.98%
Class = X = Perry	2,534	2,534/603,770 = .0042 = .42%
Class = X = Romney	168,123	168,123/603,770 = .2785 = 27.85%
Class = X = Santorum	102,475	102,475/603,770 = .1697 = 16.97%
TOTAL	603,770	~100%

Categorical Summary: Frequency Table

- **StatCrunch Command:**

Stat→Tables→Frequency→Select your variable(s)→ Ensure Frequency and Relative Frequency are highlighted→Compute

The Greek Letter Sigma in Math

- Before the Sigma was famous for representing Greek organizations on campus it was used by those developing mathematics
- This is a mathematical operator just like +, -, etc.
- This weird looking E, capital sigma, is the notation for a summation – essentially it tells you to add everything up

$$\Sigma$$

The Greek Letter Sigma in Math

- X = {1,2,3,4,5,6,7,8,9)
- $\sum x$ = 1+2+3+4+5+6+7+8+9

  = 45

- This is easy, you could have learned this in first grade – don’t make it harder than it actually is
- You can add, I have faith in you

$$\Sigma$$

Quantitative Summary: Mean

- **Mean (Average) –** The mean is the sum of observations divided by the number of observations
  - **Properties:** Sensitive to outliers, pulled in direction of the longer tail of a skewed distribution

$$\overline{x} = \frac{\sum x}{n}$$

- X are the **variable** values for our sample
- n is the size of the sample

Quantitative Summary: Example

- X = {1,2,3,4,5,6,7,8,9)

- $\bar{x} = \frac{\sum x}{n} = \frac{1+2+3+4+5+6+7+8+9}{9} = \frac{45}{9} = 5$

Quantitative Summary: Median

- **Median** – the median is the midpoint of the observations when they are ordered from the smallest to largest
  - Properties: Resistant to outliers
  - In position .5(n+1) when the data is in ascending order

Is the position value a whole number	The Median
Yes	The number in that position
No	The average of the numbers in the above and below positions

Quantitative Summary: Example

- X = {0,1,2,3,4,5,6,7,8)
- Position = .5*(n+1) = .5*(9+1) = 5th position
- Median = 4

- X = {0,1,2,3,4,5,6,7,8,9)
- Position = .5*(n+1) = .5*(10+1) = 5.5th position
- Median = (4+5)/2 = 4.5

Quantitative Summary: Mode

- **Mode**– the mode is the observation that shows up the most in the data set.
  - We allow up to three ties, if there are more we say that there is no mode

Quantitative Summary: Example

- X = {1,2,3,4,5,6,7,8,9)
  - There is no mode; all observations are tied with one occurrence
- X ={1, 1, 2, 3, 4, 5, 5, 5, 5, 6, 10}
  - Mode = 5 because 5 is the observation that occurred most.
- X ={1, 1, 1, 2, 3, 4, 5, 5, 5, 6, 10, 10}
  - Mode = 5 and 1 because 5 and 1 are the observations that occurred most.
  - **We will allow up to three ties before we revert to the first answer – There is no mode.**

Quantitative Summary: Range

- **Range –** The range is the difference between the maximum and minimum observations
  - **Properties:** easy to calculate but relies on only two values, which may be outliers

Range = Maximum - Minimum

Quantitative Summary: Example

- X = {1,2,3,4,5,6,7,8,9)

- Range = max – min = 9 – 1 = 8

Quantitative Summary: Variance

- **Variance –** the average, squared deviation of each observation from the mean
  - The idea is that it measures the spread of the data about the mean
  - **Properties:** difficult to interpret because it's in squared units, cannot be negative and is only zero when all data points are equal

Variance = $\mathrm{s}^2 = \frac{\sum(\boldsymbol{x}-\overline{\boldsymbol{x}})^{\mathbf{2}}}{\boldsymbol{n}-\mathbf{1}}$

Quantitative Summary: Example

- X = {1,2,3,4,5,6,7,8,9)
- $\bar{x} = 5$

- **variance**= $s^2 = \frac{\boldsymbol{\sum(x-\overline{x})^2}}{\boldsymbol{n-1}}$

$$= \frac{\mathbf{60}}{\mathbf{9-1}} = \frac{\mathbf{60}}{\mathbf{8}} = \mathbf{7.5}$$

X	$(\overline{\boldsymbol{x}} - \boldsymbol{x})$	$(\overline{\boldsymbol{x}} - \boldsymbol{x})^{\mathbf{2}}$
1	(1-5)=-4	$(-4)^2 = 16$
2	(2-5)=-3	$(-3)^2 = 9$
3	(3-5)=-2	$(-2)^2 = 4$
4	(4-5)=-1	$(-1)^2 = 1$
5	(5-5)=0	$0^2 = 0$
6	(6-5)=1	$1^2 = 1$
7	(7-5)=2	$2^2 = 4$
8	(8-5)=3	$3^2 = 9$
9	(9-5)=4	$4^2 = 16$
	Total:	60

Quantitative Summary: Standard Deviation

- **Standard Deviation** – the standard deviation is an adjusted average deviation of each observations’ distance from the mean
  - The idea is that it measures the spread of the data about the mean
  - We prefer this to the variance because it isn’t in squared units.
  - **Properties:** The larger the value the more spread or variability in the data, influenced by outliers and it’s always positive.

Standard Deviation = $s = \sqrt{Variance} = \sqrt{\frac{\boldsymbol{\sum(x-\overline{x})^2}}{\boldsymbol{n-1}}}$

Quantitative Summary: Example

- X = {1,2,3,4,5,6,7,8,9)
- $\bar{x} = 5$
- **variance**= $s^2 = \frac{\sum(\boldsymbol{x}-\overline{\boldsymbol{x}})^{\mathbf{2}}}{\boldsymbol{n}-\mathbf{1}}$

  $= \frac{\mathbf{60}}{\mathbf{9}-\mathbf{1}} = \frac{\mathbf{60}}{\mathbf{8}} = \mathbf{7.5}$
- **Standard Deviation =** $\mathrm{s} = \sqrt{Variance}$

$= \sqrt{\frac{\sum(\boldsymbol{x}-\overline{\boldsymbol{x}})^{\mathbf{2}}}{\boldsymbol{n}-\mathbf{1}}} = \sqrt{\mathbf{7.5}} = \mathbf{2.7386}$

Interpreting the Standard Deviation

- Later we will talk about the Empirical Rule to show how valuable the standard deviation is

- This result is very powerful in the sense that they give us a good idea about how the data is spread out

From *Naked Statistics:* Central Tendency Example

- Let’s say ten people at a bar – the cast of Mixology – each make $35k/yr

From *Naked Statistics:* Central Tendency Example

- X=$\{35000, 35000, 35000, 35000, 35000, 35000, 35000, 35000, 35000, 35000\}$
- $\bar{x} = \frac{1}{10}(35000 + 35000 + 35000 + 35000 + 35000 + 35000 + 35000 + 35000 + 35000 + 35000) = \frac{1}{10} * 350000 = 35000$
- Median=35,000

From *Naked Statistics:* Central Tendency Example

- Range = 35,000 – 35,000 = 0
- IQR = 35,000 – 35,000 = 0
- Variance = 0 (Via StatCrunch)
- Standard Deviation = $\sqrt{0} = 0$

From *Naked Statistics:* Central Tendency Example

- Let’s add Bill Gates to the mix who, we’ll assume, make $1B/Year

From *Naked Statistics:* Central Tendency Example

- $\mathrm{X}=\{35000, 35000, 35000, 35000, 35000, 35000, 35000, 35000, 35000, 35000, 1000000000\}$
- $\bar{x}=\frac{1}{11}(35000+35000+35000+35000+35000+35000+35000+35000+35000+35000+1000000000)$
  $$=\frac{1}{11}*1000350000=\$90{,}940{,}909.09$$
- Median=35,000
- This outliers inflate the mean but not the median.

From *Naked Statistics:* Central Tendency Example

- Range = 1,000,000,000 – 35,000 = 999,965,000

- IQR = 35,000 – 35,000 = 0

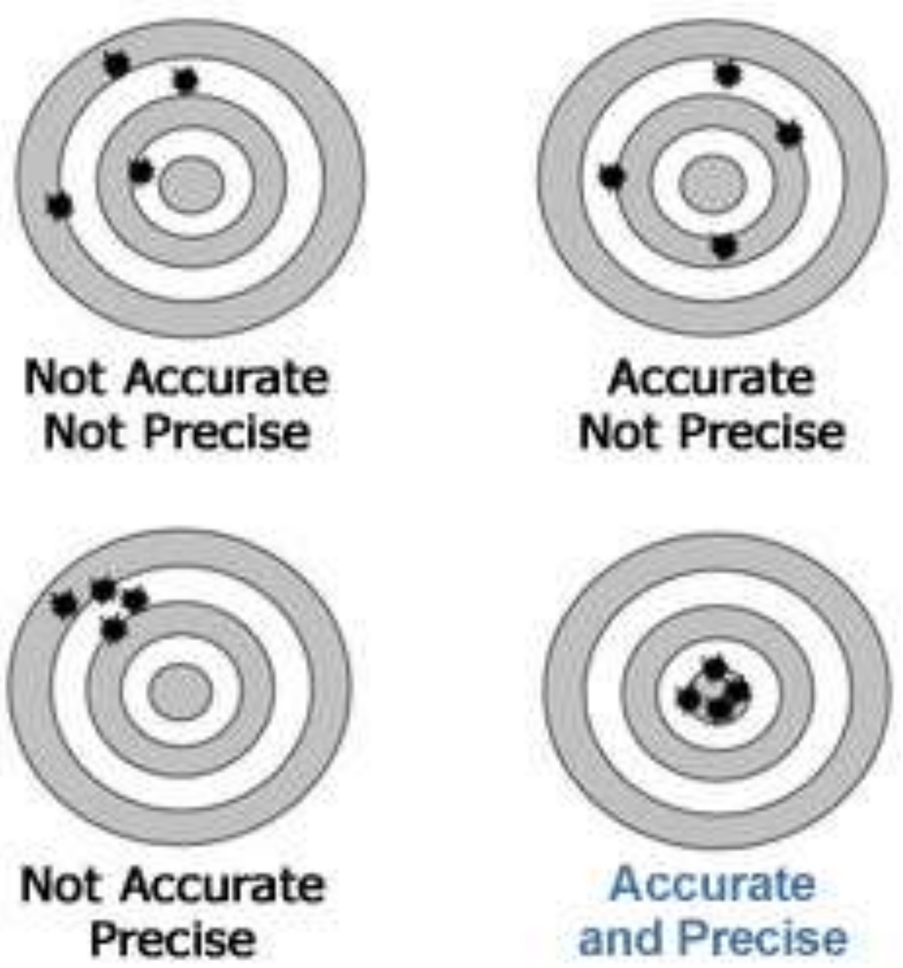
- Variance = 90,902,727,000,000,000(Via StatCrunch)

- Standard Deviation = $\sqrt{var} = 301500790$

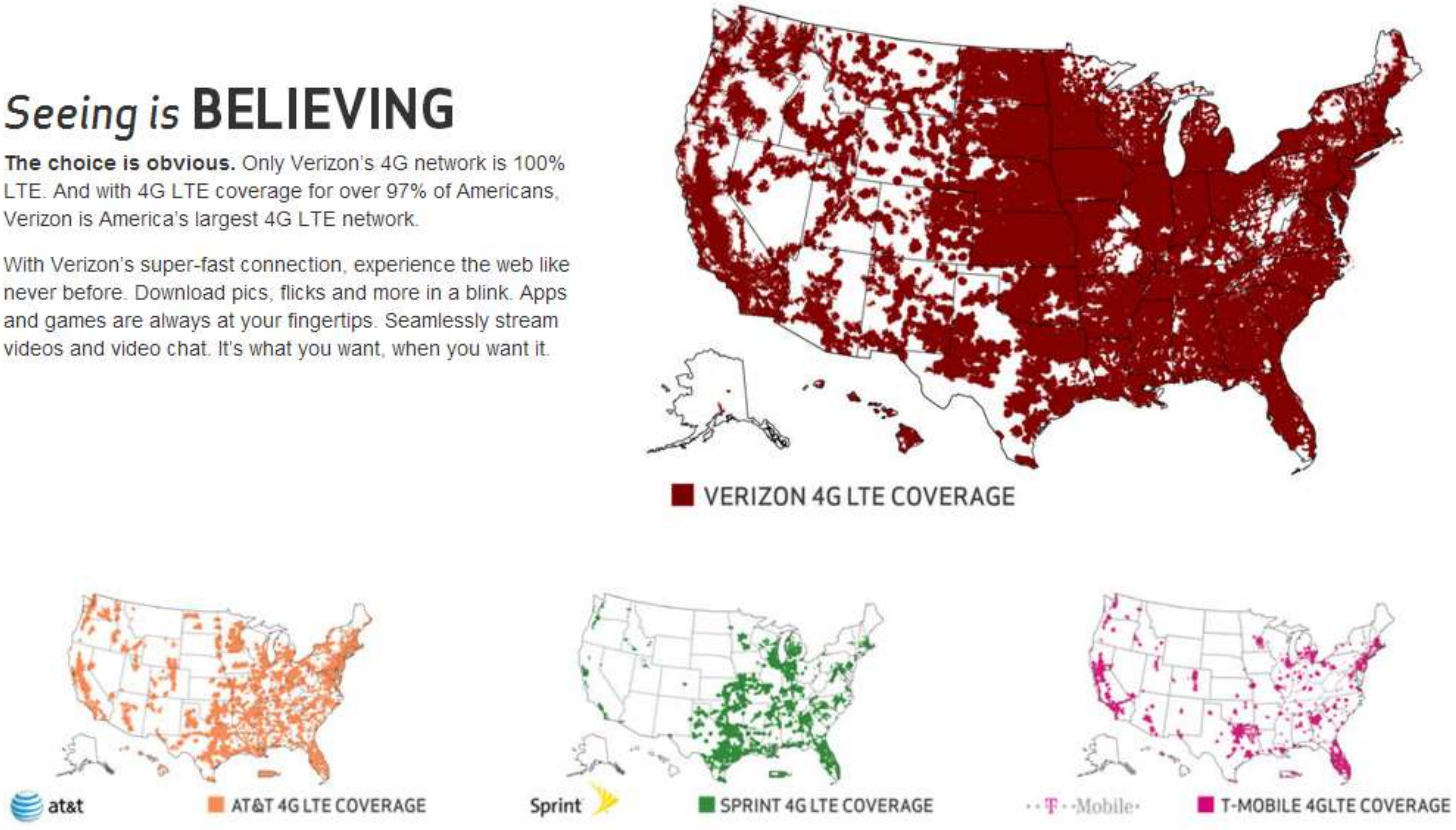
- This outliers inflates everything but the IQR.

From *Naked Statistics: Deceptive Statistics*

- "Precision reflects the exactitude with which we can express something."
- "Accuracy is a measure of whether a figure is broadly consistent with the truth."
- "These words are not interchangeable"



From *Naked Statistics:* *Deceptive Statistics*



Seeing is **BELIEVING**

The choice is obvious. Only Verizon's 4G network is 100% LTE. And with 4G LTE coverage for over 97% of Americans, Verizon is America's largest 4G LTE network.

With Verizon's super-fast connection, experience the web like never before. Download pics, flicks and more in a blink. Apps and games are always at your fingertips. Seamlessly stream videos and video chat. It's what you want, when you want it.

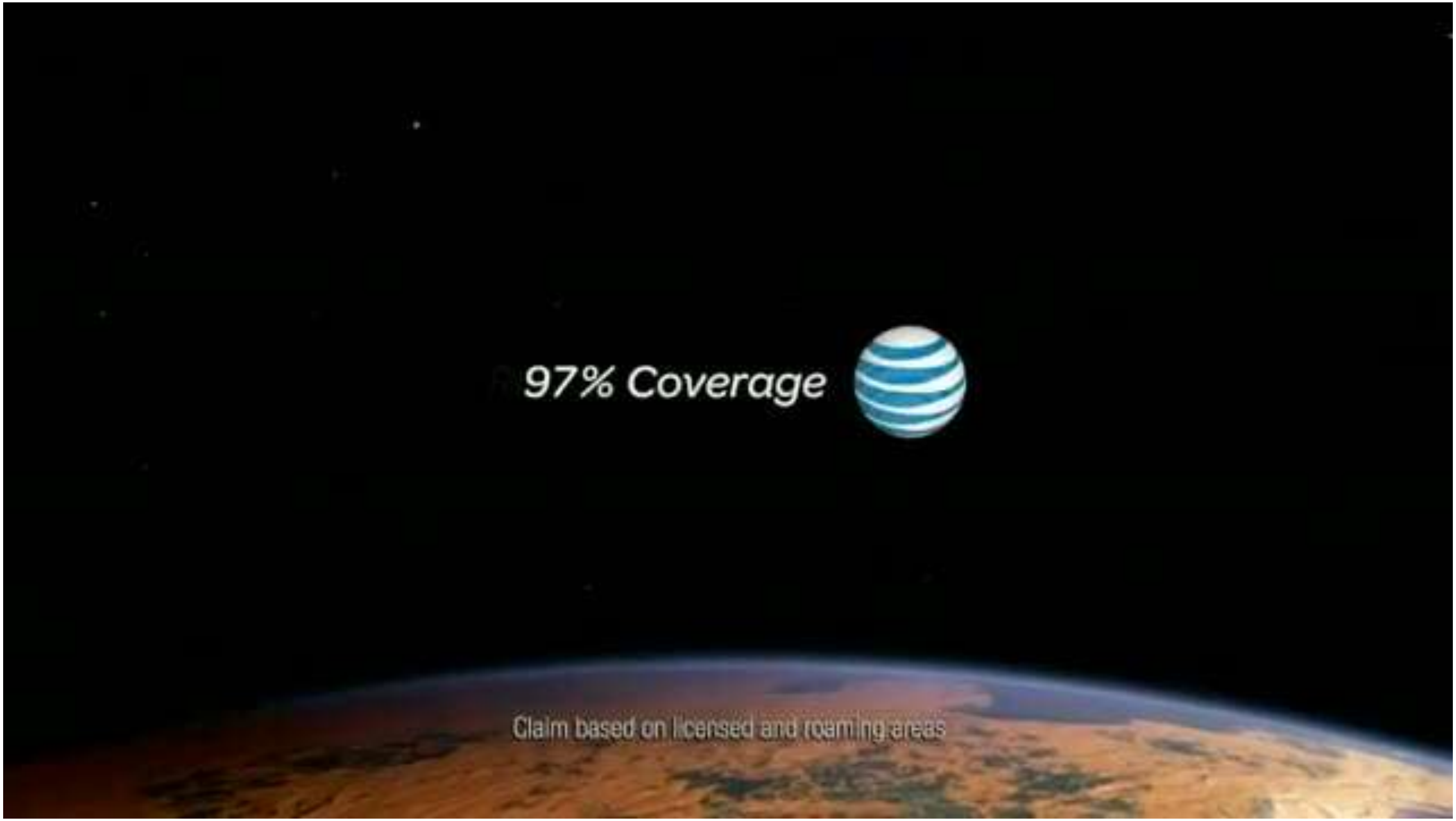
■ VERIZON 4G LTE COVERAGE

at&t ■ AT&T 4G LTE COVERAGE

Sprint ■ SPRINT 4G LTE COVERAGE

T-Mobile ■ T-MOBILE 4GLTE COVERAGE

From *Naked Statistics: Deceptive Statistics*



From *Naked Statistics: Deceptive Statistics*

- “The unit of analysis chosen by Verizon is geographic area covered – because the area has more of it.”
- “AT&T countered by launching a campaign that changed the unit of analysis. Its billboards advertized that ‘AT&T covers 97 percent of Americans”
- “Note the use of the word ‘Americans’ rather than ‘America’”

From *Naked Statistics: Deceptive Statistics*

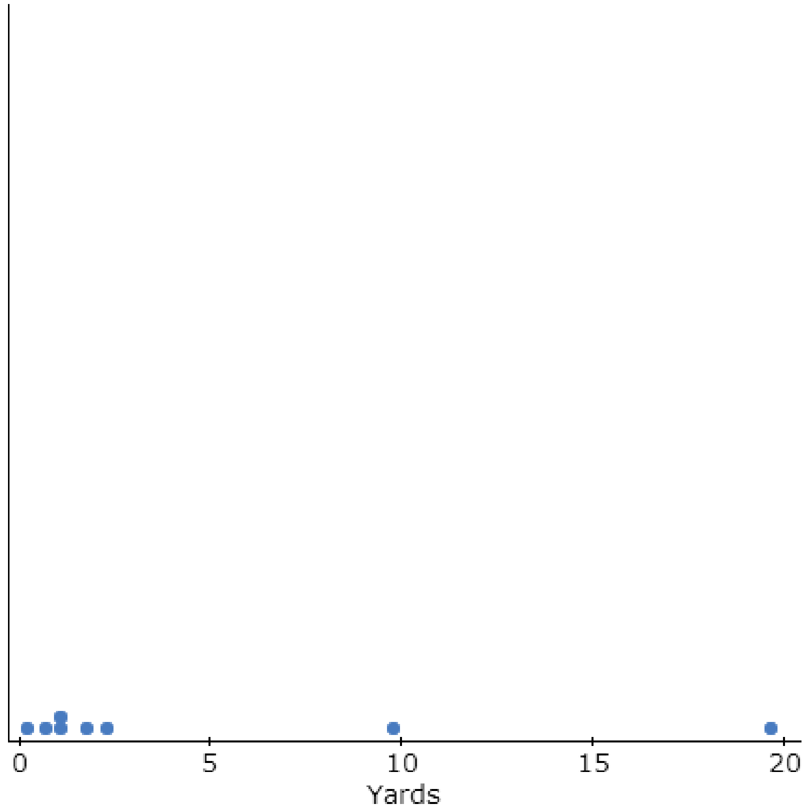
- “The fact that you’ve calculated the mean correctly will not alter the fact that the median is a more accurate indicator. Judgement and integrity turn out to be surprisingly important. A detailed knowledge of statistics does not deter wrongdoing any more than a detailed knowledge of the law averts criminal behavior. With both statistics and crime, the bad guys often know exactly what they’re doing”

Quantitative Summary: Example

- X = yards per carry for Marcus Lattimore = {.2, .7, 1.1, 1.2, 1.8, 2.3, 9.8, 19.7}
- What kind of **data** is this?
  - We know that distance or length is a **Continuous Quantitative** variable but we measure it discretely here by tenths of a yard
  - What type of graphs would be appropriate?
    - **Dot plot**, **Box plot**, **steam and leaf plot**, or a **histogram**

Quantitative Summary: Example

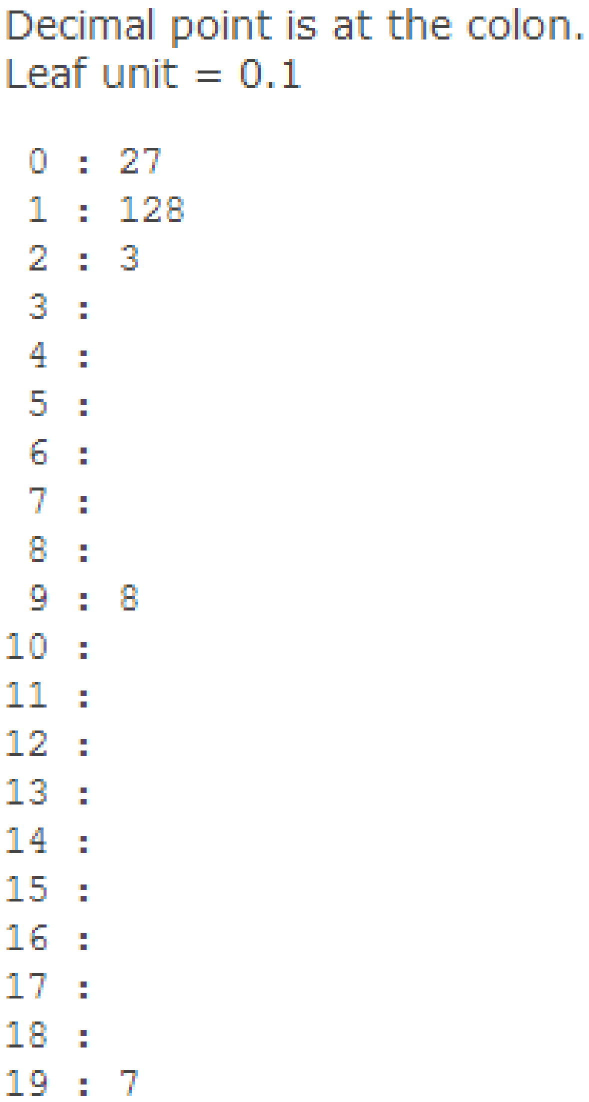
- Let’s try a **dot plot**!
- Our gaps/possible outliers are clear because it is far away but the graph is awkward and hard to read in general



Quantitative Summary: Example

- Let’s try a **Stem and Leaf Plot**!

- Our gaps/possible outliers are clear because it is far away but the graph is awkward and hard, or at least annoying, to read



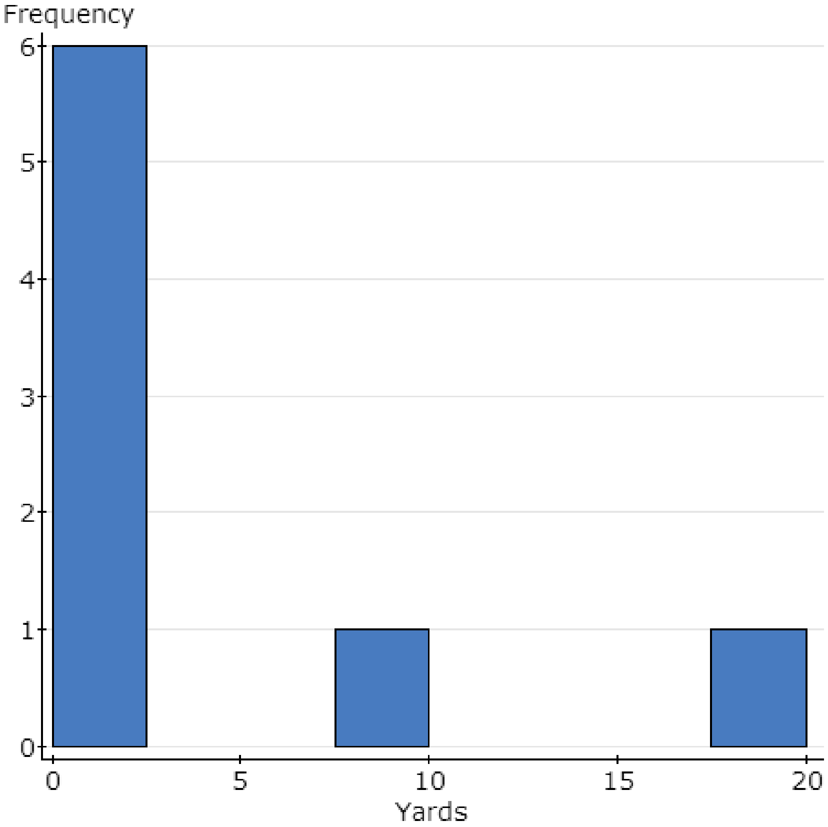
```
Decimal point is at the colon.
Leaf unit = 0.1

 0 : 27
 1 : 128
 2 : 3
 3 :
 4 :
 5 :
 6 :
 7 :
 8 :
 9 : 8
10 :
11 :
12 :
13 :
14 :
15 :
16 :
17 :
18 :
19 : 7
```

Quantitative Summary: Example

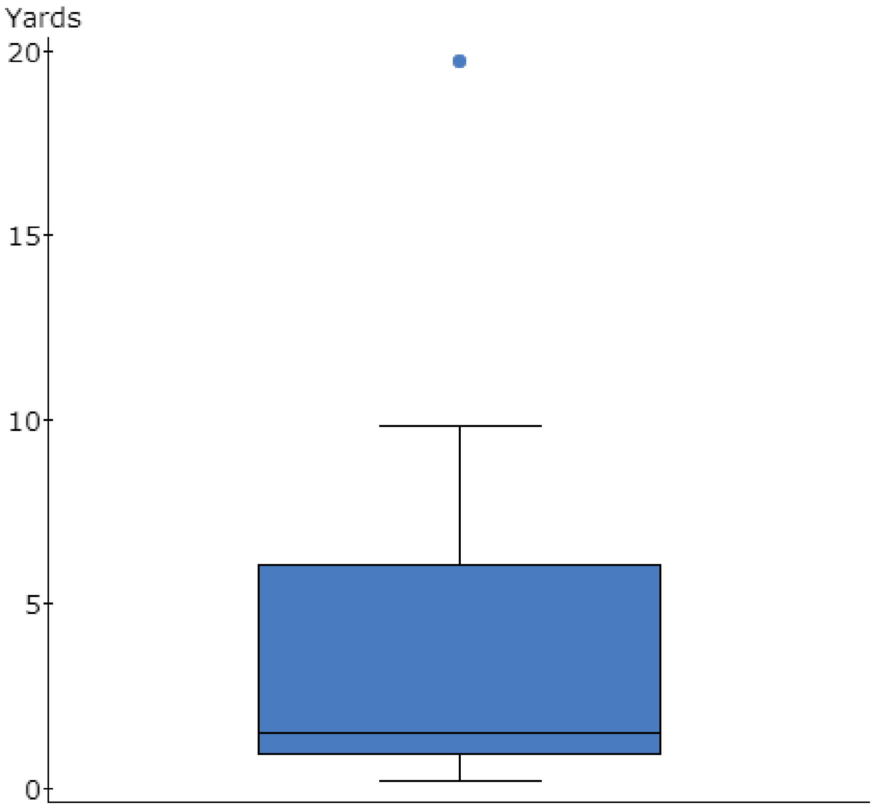
- Let’s try a **histogram**

- This is better, but we still have some awkwardness with the gaps/possible outliers aren’t as obvious here



Quantitative Summary: Example

- Let’s try a **box plot**
- This is really the best choice
  - An outlier is very clearly shown
  - The rest of the graph is readable and not nearly as awkward as the others

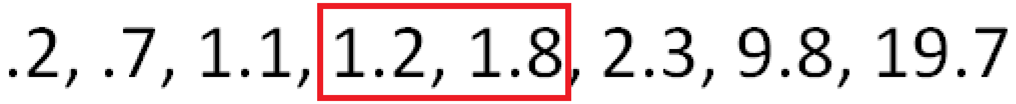


Quantitative Summary: Example

- **Mean:** $\bar{x} = \frac{\sum x}{n}$

  = (.2 + .7 + 1.1 + 1.2 + 1.8 + 2.3 + 9.8 + 19.7) **/** 8

  = 4.6

- **Median:** 

  .2, .7, 1.1, 1.2, 1.8, 2.3, 9.8, 19.7

  - **Position** = .5(8+1) = 4.5th position

  = (1.2 + 1.8) / 2 We take the average of the two

  = 1.5

- **Mode:** there is no mode

Quantitative Summary: Example

- X = yards per carry for Marcus Lattimore = {.2, .7, 1.1, 1.2, 1.8, 2.3, 9.8, 19.7}

- **Mean:** $\bar{x} = \frac{\sum x}{n}$ = 4.6
- **Median** = 1.5
- **Mode:** there is no mode

- Do you get the same answers in the calculator?

Quantitative Summary: Example

After removing the outlier,

- **Mean:** $\bar{x} = \frac{\sum x}{n}$

  = (.2 + .7 + 1.1 + 1.2 + 1.8 + 2.3 + 9.8) **/** 7

  = 2.442857

- **Median:** $.2, .7, 1.1, \boxed{1.2}, 1.8, 2.3, 9.8$
  - **Position =** .5(7+1) = 4

  = 1.2

Quantitative Summary: Example

- X = yards per carry for Marcus Lattimore = {.2, .7, 1.1, 1.2, 1.8, 2.3, 9.8, 19.7}

- **Mean:** $\bar{x} = \frac{\sum x}{n}$ = 2.442857
- **Median** = 1.2
- **Mode:** there is no mode

- Do you get the same answers in the calculator?

Quantitative Summary: Example

- Before Removing Outlier: **Mean =** 4.6
  Median = 1.5
- After Removing Outlier: **Mean =** 2.442857
  Median = 1.2
- Notice that the mean changes much more than the median. Remember that the median is resistant to outliers and the mean is not.
- Notice the **mean > median** so it is **right skewed** in both cases!

Quantitative Summary: Example

- X = yards per carry for Marcus Lattimore = {.2, .7, 1.1, 1.2, 1.8, 2.3, 9.8, 19.7}

- **Range** = Maximum – Minimum

  = 19.7 - .2

  = 19.5

Quantitative Summary: Example

- X={.2, .7, 1.1, 1.2, 1.8, 2.3, 9.8, 19.7}
- **Variance =** $\frac{\sum(\boldsymbol{x}-\overline{\boldsymbol{x}})^{2}}{\boldsymbol{n}-\mathbf{1}} = \frac{326.56}{8-1} = 46.6514 \text{ yds}^2$

x	(x - mean)		(x - mean) ^2	
0.2	.2 - 4.6 =	-4.4	(-4.4)^2 =	19.36
0.7	.7 - 4.6 =	-3.9	(-3.9)^2 =	15.21
1.1	1.1 - 4.6 =	-3.5	(-3.5)^2 =	12.25
1.2	1.2 - 4.6 =	-3.4	(-3.4)^2 =	11.56
1.8	1.8 - 4.6 =	-2.8	(-2.8)^2 =	7.84
2.3	2.3 - 4.6 =	-2.3	(-2.3)^2 =	5.29
9.8	9.8 - 4.6 =	5.2	(5.2)^2 =	27.04
19.7	19.7 - 4.6 =	15.1	(15.1)^2 =	228.01
			Total	326.56

Quantitative Summary: Example

- X = {.2, .7, 1.1, 1.2, 1.8, 2.3, 9.8, 19.7}

- Standard Deviation =$\sqrt{Variance}$

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

$$= \sqrt{46.6514}$$

$$= 6.8302$$

Quantitative Summary: A Tricky One

- Scores for Class A: 30, 65, 70, 76, 93, 99
- Scores for Class B: 68, 72, 73, 73, 74, 77

Class	n	Mean	Median
Class A	6	72.1667	73
Class B	6	72.8333	73

- Now, these are very similar. Would you say the students in each class performed the same?
  - Yes, the **mean** and **median** are almost identical

Quantitative Summary: A Tricky One

- A more complete summary will include a measure of spread

Class	n	Mean	Median	Variance	St. Dev
Class A	6	72.1667	73	600.5667	24.5065
Class B	6	72.8333	73	8.5667	2.9269

- Note, now we can say that although the mean and median were almost identical, the scores of Class A were more varied.

Getting Descriptive Statistics on our TI

1. Press STAT
2. Press ENTER with ‘Edit’ highlighted
3. Enter the data into the L1 column
4. Press STAT
5. Press → to CALC
6. Press ENTER with ‘1-Var Stats’ highlighted
7. Press 2nd
8. Press 1
9. Press Enter

Getting Descriptive Statistics on our TI

- $\bar{x}$ = the sample mean
  - This is the population mean, μ, when we've entered the population

- $\sum x$ = the sum of our x variables
  - We divide this by n to get $\bar{x}$

Getting Descriptive Statistics on our TI

- $\sum x^2$ = the sum of our x^2 variables
  - All x values are squared before the are added

- s_x = the sample standard deviation
  - Use this when we’ve entered a sample

Getting Descriptive Statistics on our TI

- σ_x = the population standard deviation
  - Use this when we've entered the population

- n = the sample population size
  - This is the population size when we've entered the population

Getting Descriptive Statistics on our TI

- Scroll down using ↓ to see the five number summary:
- minX= the minimum of our data set
- Q1 = the first quartile of our data set
- Med = the median of our data set
- Q3 = the third quartile of our data set
- MaxX = the maximum of our data set

Getting Descriptive Statistics

- **StatCrunch Command:**

Stat→Summary Stats→Columns→Select the variable(s)→Compute

- **StatCrunch Command to compare summaries across groups:**

Stat→Summary Stats→Columns→Select the variable(s)→Group by the group variable→Compute