Understanding Sampling Variability

Roxy Peck
Cal Poly San Luis Obispo

Dangerous Equations!!

In his 2007 paper in *American Scientist* titled "The Most Dangerous Equation" Harold Wainer introduced the idea of a dangerous equation.

http://www.americanscientist.org/issues/pub/the-most-dangerous-equation/1

Two Kinds of Dangerous Equations

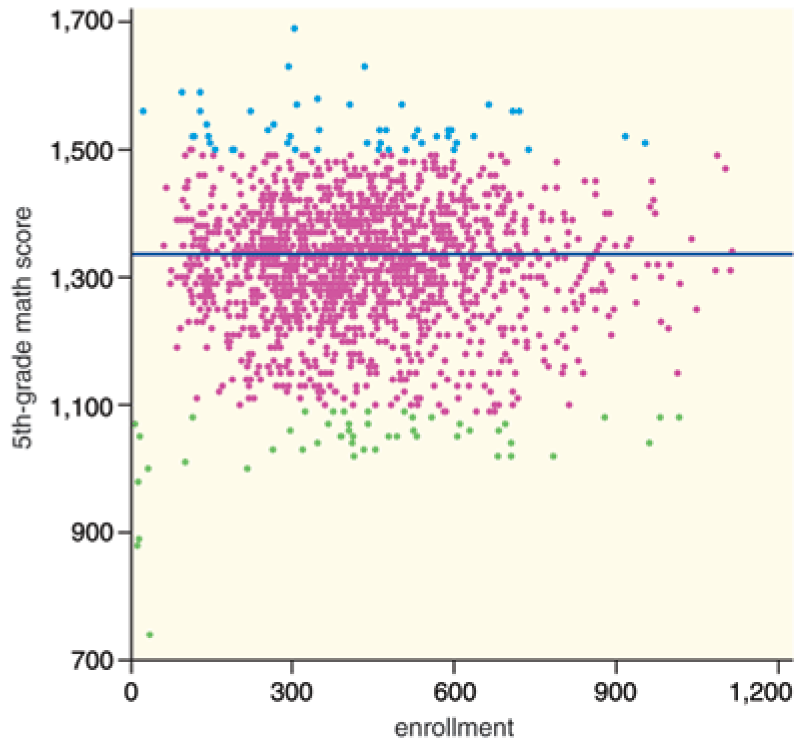
- Dangerous because we know them…
  - for example $e = mc^2$
    led to the development of the atomic bomb
- Dangerous because we don't know them…

What Did Wainer Think The Most Dangerous Equation Was???

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Danger Ahead!

- Small Schools Movement



Graph from Wainer paper. Based on 1,662 schools in Pennsylvania.

Danger ahead…

- Focus on high scoring schools. Notice that disproportionate number of high scoring schools were small.
- Lead to foundation funding for small schools movement. Many large schools were broken up into small schools (sometimes even under the same roof).
- And the outcome? No real gains in test scores. In fact some of the schools that initially scored well scored low in subsequent years.

Danger Ahead…

- What went wrong??
  - Forgot to look at the low scoring schools…they were also disproportionately small schools as well.
  - What is graphed is the average math score for the schools. Expect more variability in averages from year to year from small schools than from large schools. It's that darn square root of n thing!

More Danger!

- Other great examples can be found in the Wainer paper, including
  - Cancer rates by county.
  - “Safest” Cities—none of the safest cities are the largest cities. But then neither are the least safe cities!

So if these guys can't get it right…

- Understanding the difference between variability in individual observations and variability in statistics like means and proportions is IMPORTANT.
- But it is a challenging concept, and obviously there are lots of people (including policy makers) who don't get it, so it is not surprising that students struggle with this.

Implications for What and How We Teach This Concept

- Important concepts to develop:
  - Variability and the notion that distributions describe variability.
  - Size matters if we are talking about a sample.
  - Size doesn't matter (usually) when we are talking about a population.

First Key to Understanding: Distributions Describe Variability

Being able to think about data in terms of distributions and to distinguish between the different ways distributions are used (to describe the values in a population, the values in a sample or the values of a statistic for different possible samples) are key to understanding statistical inference.

Life would be simple if…

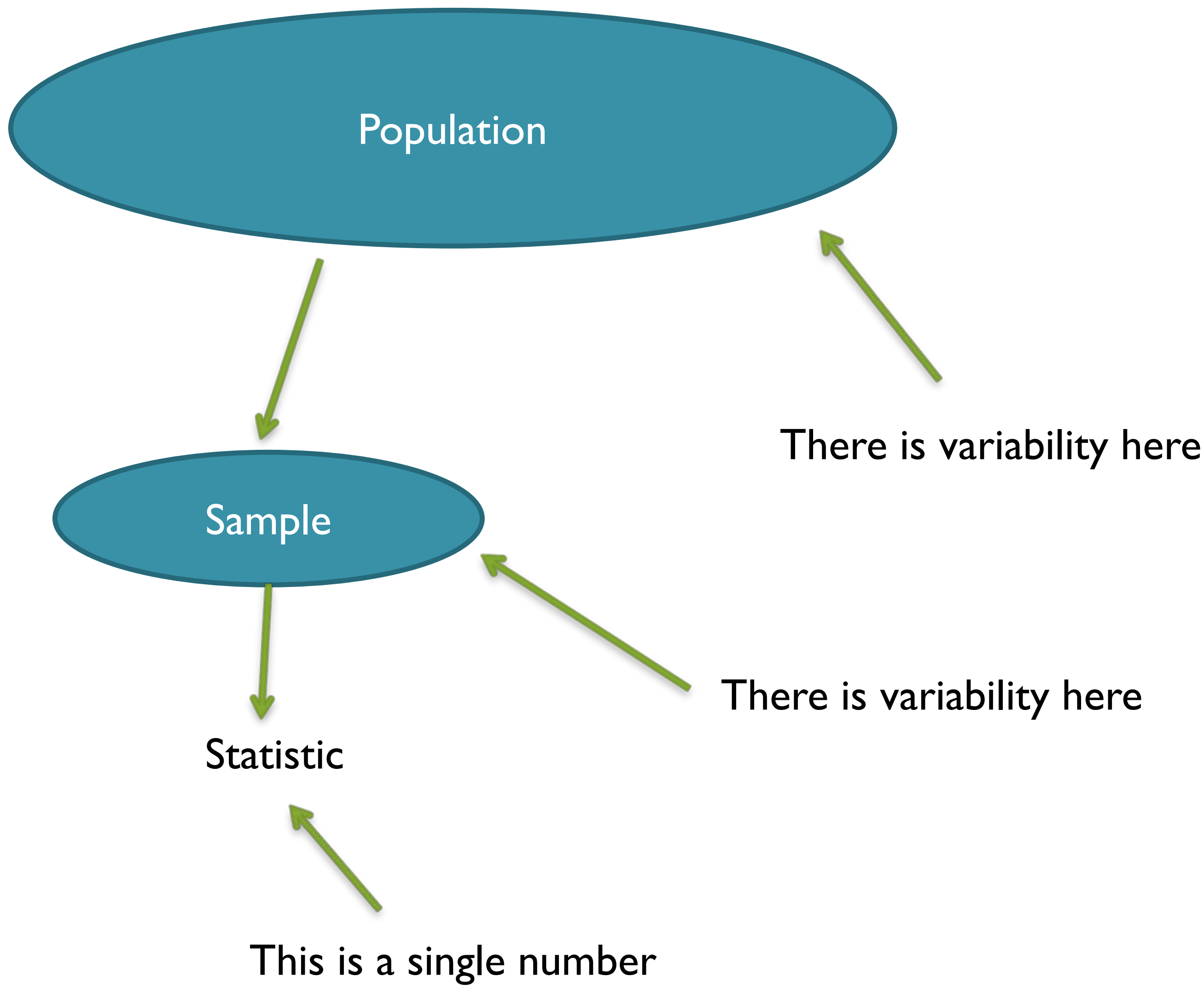
- there was no variability!
  - Simple because it would be easy to draw conclusions based on data.
  - Simple, but BORING!
- As Statistics people we
  - LOVE variability
  - Are OK with being wrong 5% of the time!

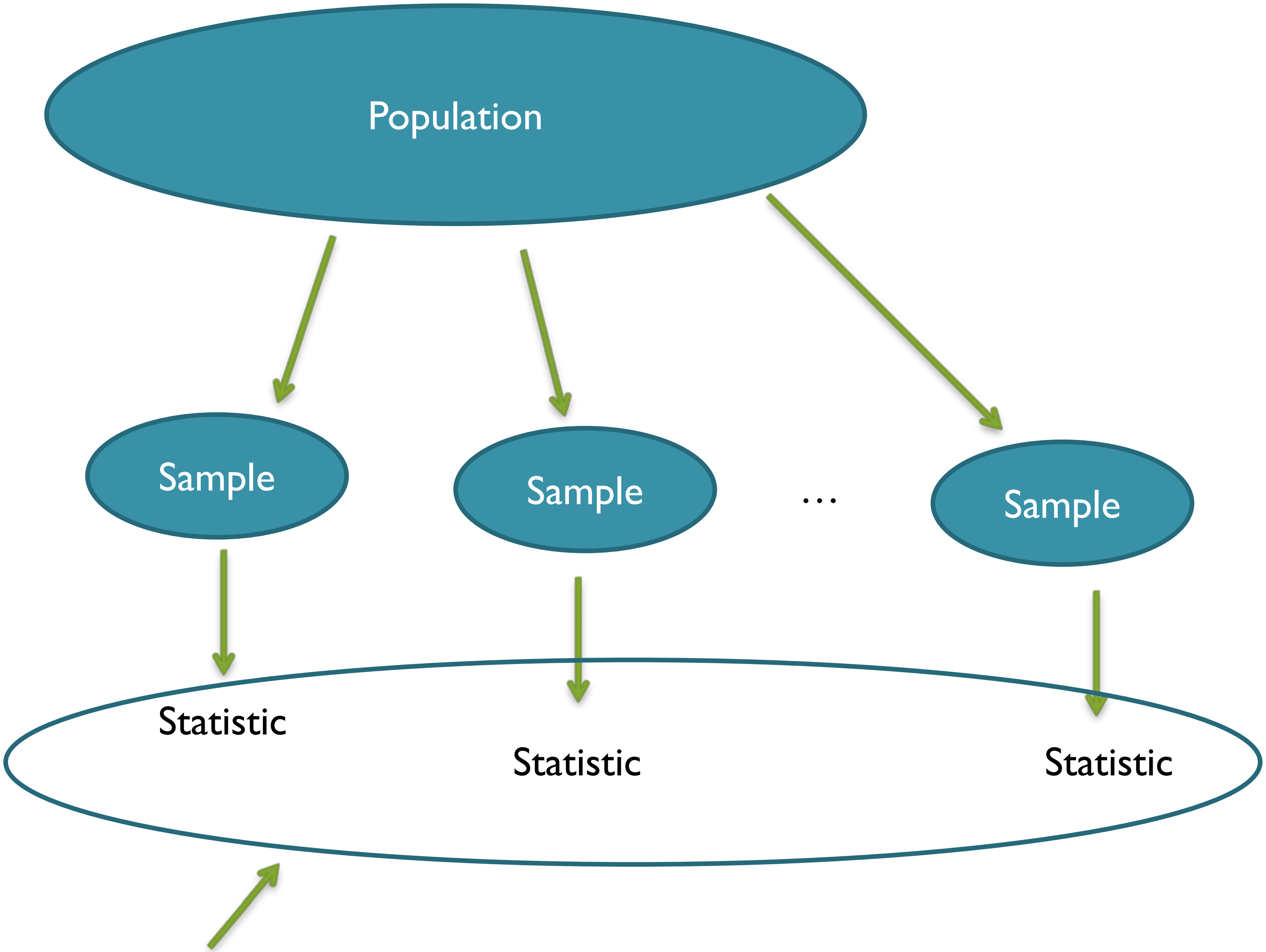
Most math folks don't get this!

Distributions Describe Variability

- Variability in a population
- Variability in a sample
- Sample-to-sample variability in the values of a statistic

- How do we get students to understand the difference between these three types of distributions at a conceptual level?


Population
Sample
Statistic
There is variability here
There is variability here
This is a single number



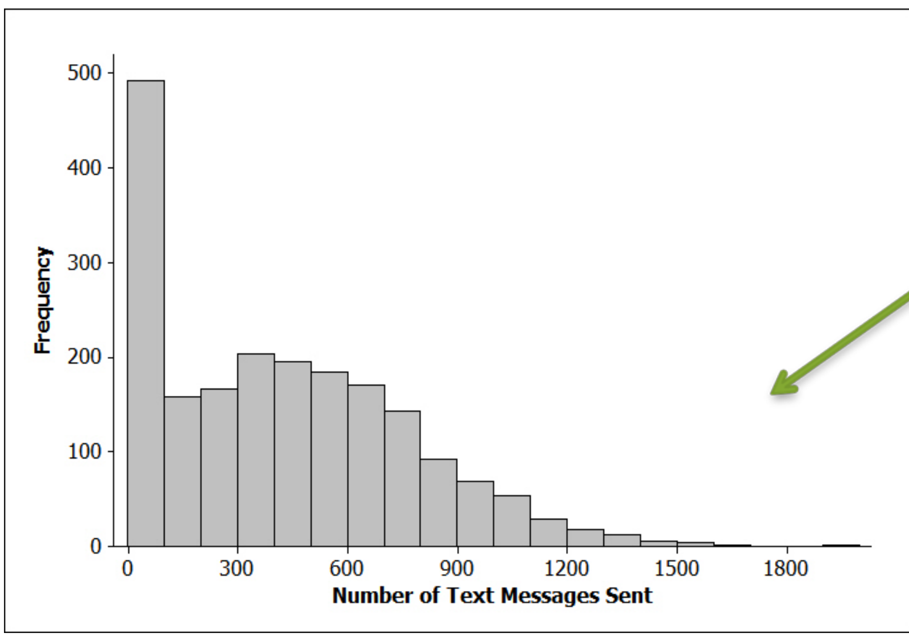
There is variability in these statistic values

Distributions Describe Variability

- Variability in a population
  - Population Distribution
- Variability in a sample
  - Sample Distribution
- Sample-to-sample variability in the values of a statistic
  - Sampling Distribution

Example

- Population of 2000 students
- Distribution of number of text messages sent in one month

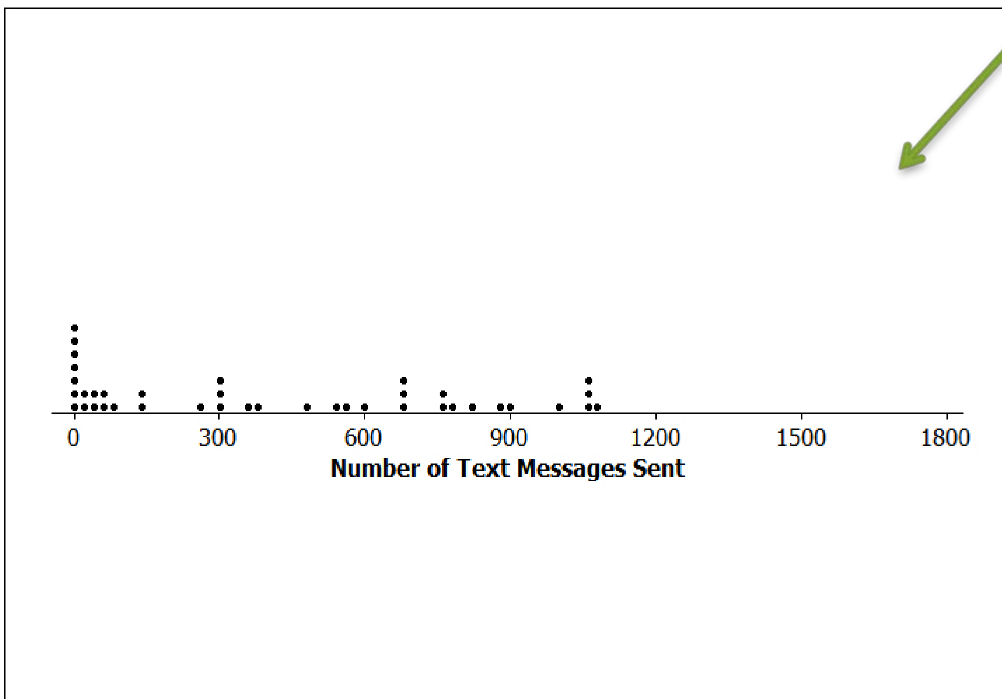


This is a population distribution

Example continued

- Random sample of 40 students

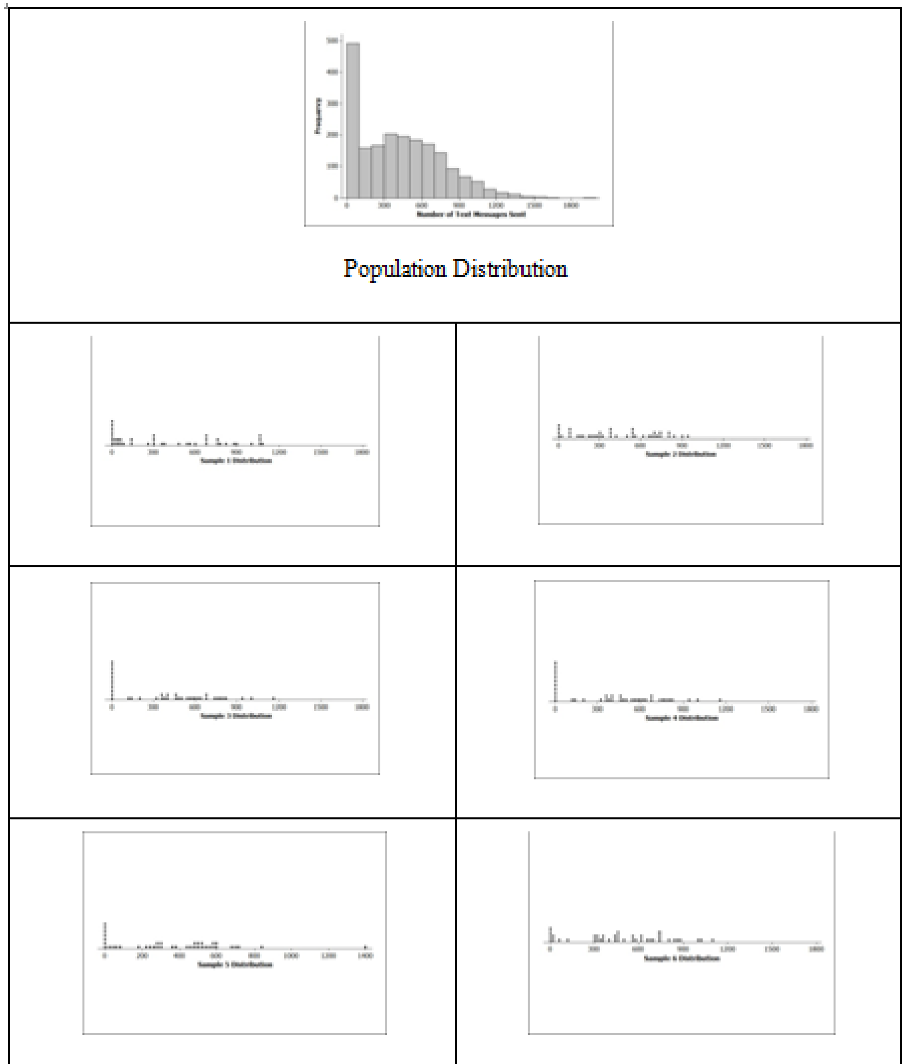
This is a sample distribution



Example continued

IF sample is selected at random from the population, we expect the sample distribution to resemble the population distribution

Sample-to-Sample Variability

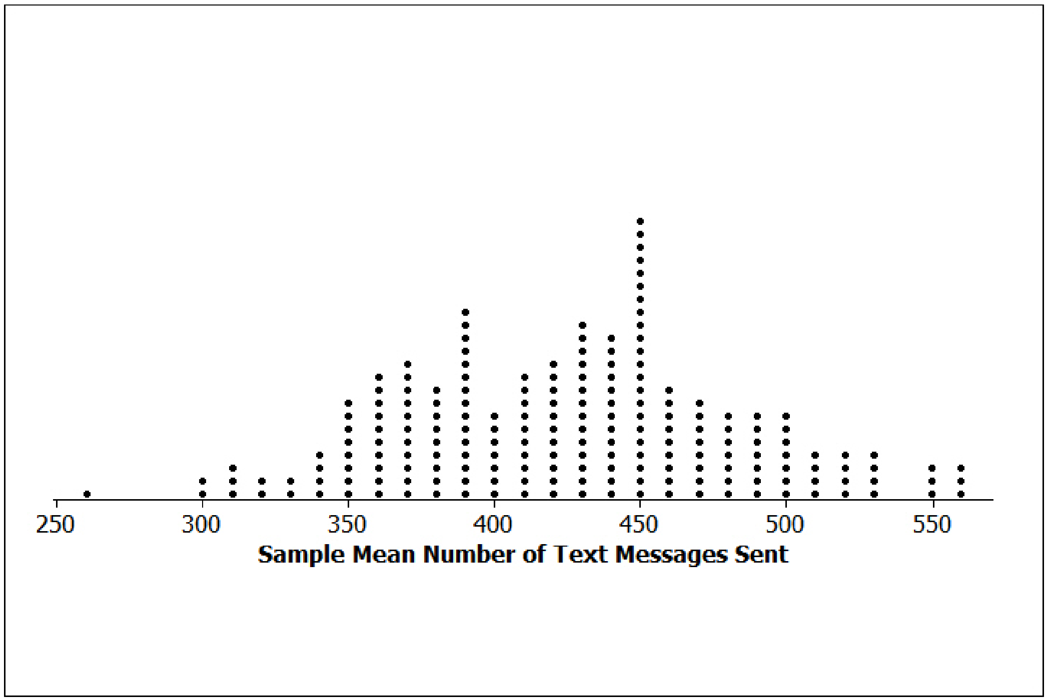


Example Continued

For this population and for these six samples:

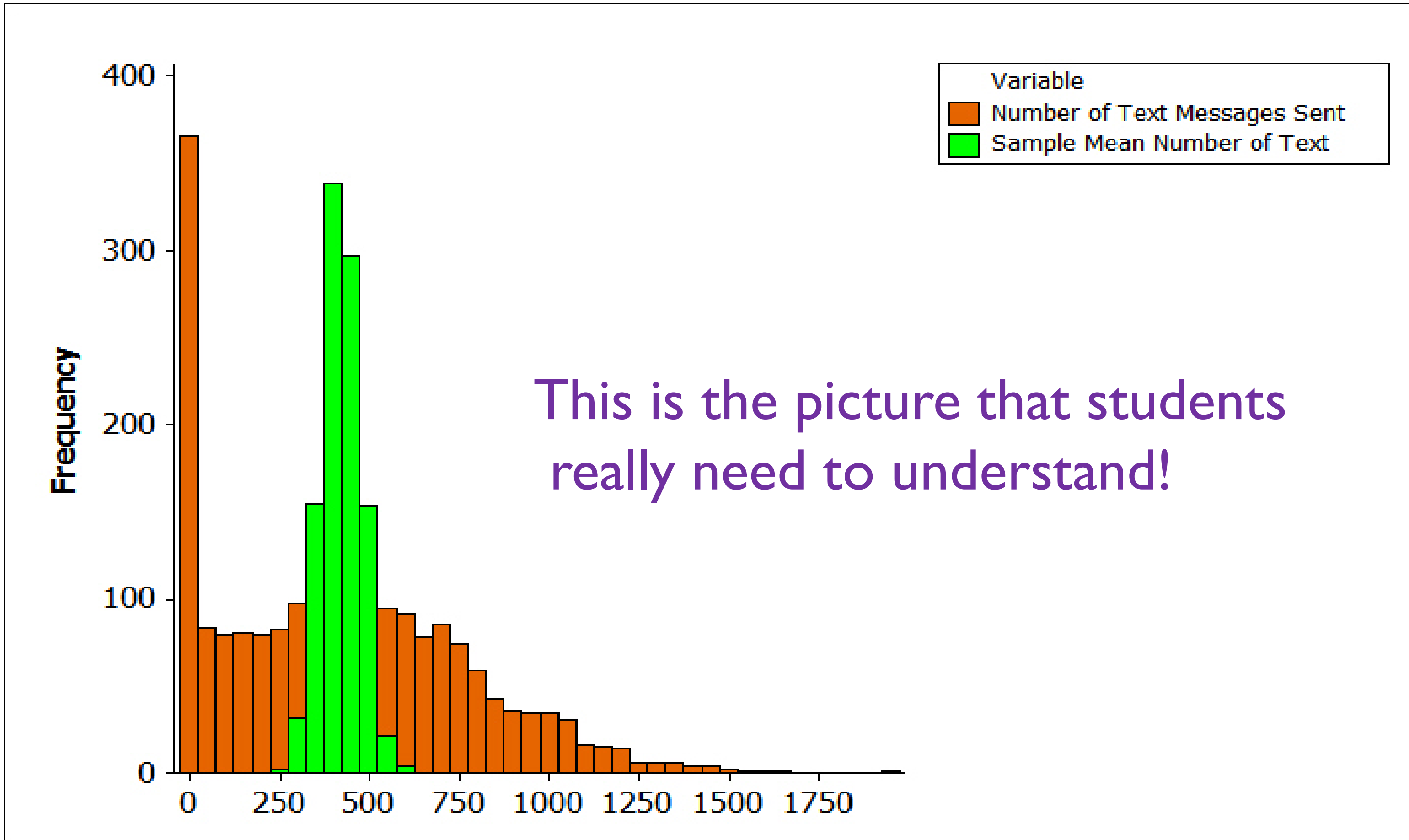
Population mean	**420.6**
Sample 1 mean	422.2
Sample 2 mean	418.2
Sample 3 mean	411.7
Sample 4 mean	564.2
Sample 5 mean	367.2
Sample 6 mean	484.7

For 200 Random Samples



Start of a Sampling Distribution

For 1000 Random Samples



Second Key to Understanding: Size Matters (well sometimes)!

- Sample size matters.
  - Why? Because of

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- Population size doesn't (usually) matter.
  - Why? Because of
  - But sometimes (not often)

Finite population correction factor
Usually very close to 1

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

Classroom Activities

- Demonstration similar to the text messages example.
- Activity 1: Is it s or is it $\frac{s}{\sqrt{n}}$?
- Activity 2: Matching activity

Activity 1

The weights of full term babies born in the U.S. are normally distributed with a mean of 3500 grams and a standard deviation of 600 grams.

Would it be surprising to see a full term baby with a birth weight of 4000 grams?

Would it be surprising if the average weight of babies born during 2015 at a large urban hospital was 4000 grams?

Would it be surprising if the average weight of babies born on October 10, 2015 at a large urban hospital was 4000 grams?

Would it be surprising if the average weight of babies born on October 10, 2015 at a small rural hospital was 4000 grams?

Activity 1

Suppose that you come to class unprepared and find out that there is a quiz over the material in the assigned reading. The questions on the quiz are true false and are scored as 0 points for an incorrect answer and 1 point for a correct answer. The score on the quiz is the average of the scores assigned (the proportion correct). To pass the quiz, you need to have a score of 0.7 or higher. Would you prefer that the quiz have 10 questions or that the quiz have 30 questions?

Activity 1

Suppose that there is a lot of variability in the actual weight of gold coins produced at a particular mint. The average weight of the coins produced is 1 ounce, and the standard deviation of the weights is 0.2 ounces. (This is not a very good mint!). All coins are the same price and are based on the price of 1 ounce of gold. If you get a coin that weighs less than an ounce, you will lose money, but if you get a coin that weighs more than one ounce, you actually come out ahead. You can buy either one coin or 20 coins.

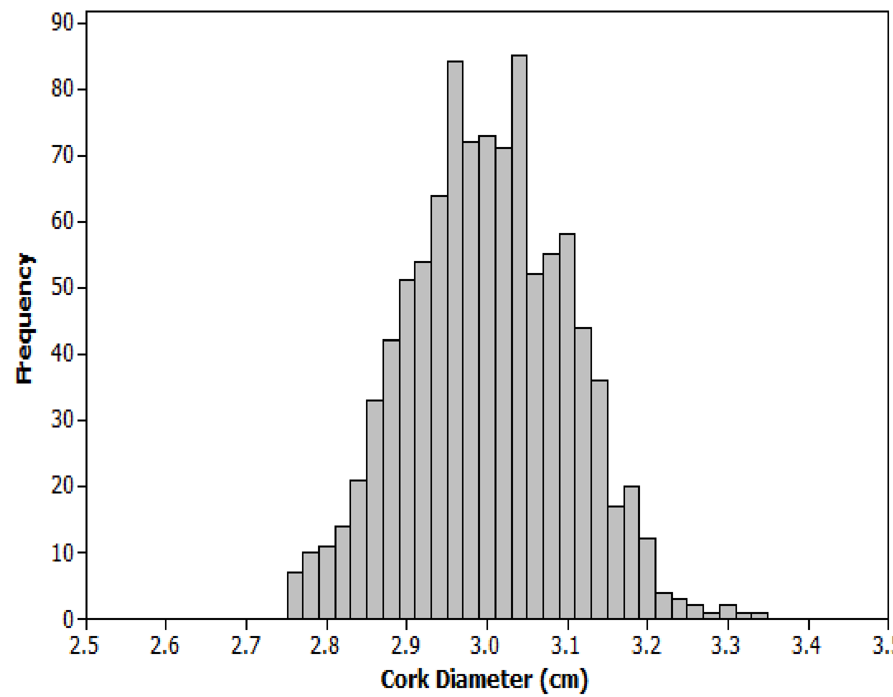
Describe one reason why you might choose to only buy one coin.

Describe one reason why you might choose to buy 20 coins.

Activity 2

A machine that cut corks for wine bottles produces corks that have diameters with a mean of 3 cm and a standard deviation of 0.1 cm. Corks with diameters that are smaller than 2.9 cm or larger than 3.1 cm are considered defective (a cork that is too small will result in a leaky bottle and a cork that is too big will not fit in the bottle. Below is a histogram of the distribution of the diameters of the 1000 corks produced by this machine in one day.

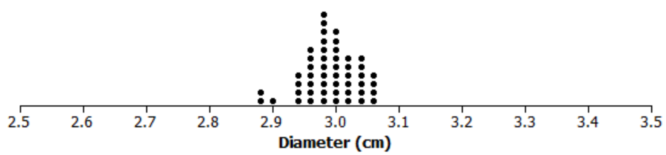
Activity 2



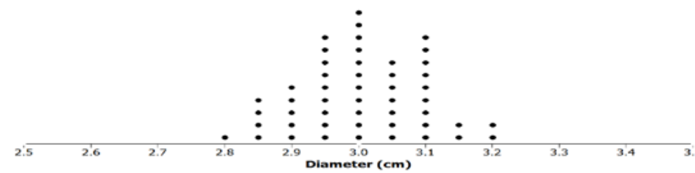
Frequency
90
80
70
60
50
40
30
20
10
0
2.5
2.6
2.7
2.8
2.9
3.0
3.1
3.2
3.3
3.4
3.5
Cork Diameter (cm)

Activity 2

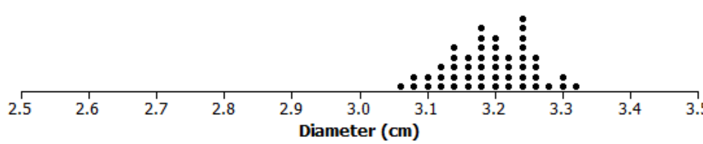
Dotplot 1



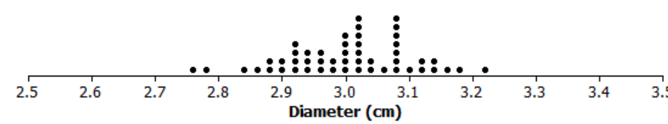
Dotplot 2



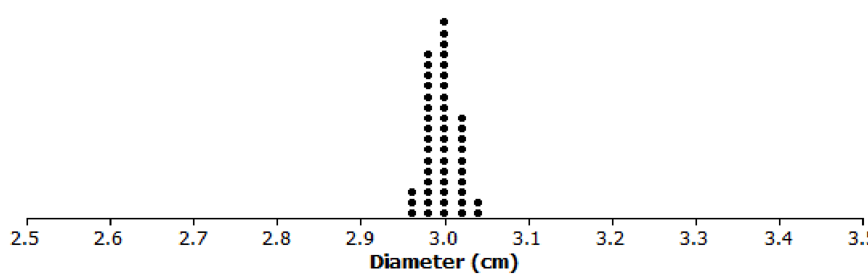
Dotplot 3



Dotplot 4



Dotplot 5



Activity 2

Two of the dotplots are dotplots of the diameters of corks in random samples of size 50 from this population. Which two dotplots are these?

Two of the dotplots are dotplots of sample means from 50 random samples from this population. Which two dotplots are these?

Which of the two dotplots that you identified as a dotplot of sample means is based on the larger sample size?

Which dotplot is not a dotplot of a sample from this population and is also not a dotplot of sample means for random samples from this population? How can you tell?

In Conclusion…

- Need to help students develop understanding of sampling variability.
- Be on the lookout for situations that require "the dangerous equation".
- Focus on
  - Variability and the notion that distributions describe variability.
  - Size matters if we are talking about a sample.
  - Size doesn't matter (usually) when we are talking about a population.

Thanks for attending this session.

- Questions?

- Contact information
  - Roxy Peck
  - Cal Poly
  - rpoeck@calpoly.edu