MATH 1610/MATH 1553
7.1 Sampling Introduction

DEFINITIONS

Element – the entity on which data are collected

Population – the collection of all the elements of interest

Sample – subset of the population.

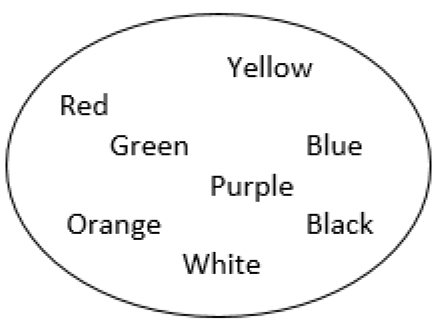
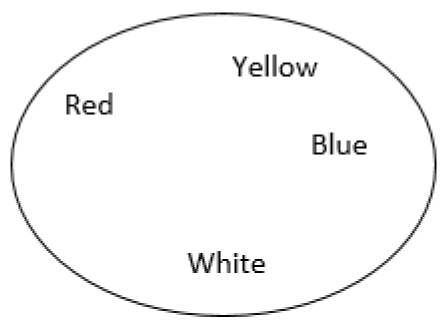
Sampled Population – the population from which the sample is drawn.

Frame – a list of the elements that the sample will be selected from

Parameters – numerical characteristics of a population

Descriptive Statistics – collecting, presenting and describing data

Inferential Statistics – drawing conclusions and/or making decisions concerning a population based only on sample data.

<table>
<tr><th>POPULATION</th><th>SAMPLE</th></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>All likely voters in the next election.</td><td>1000 voters selected at random for a poll.</td></tr>
<tr><td>All sales receipts for November.</td><td>Random receipts selected for an audit.</td></tr>
</table>

Reasons for Sample:	**Types of Sample Populations**:
1. collect data to make an inference 2. less costly and time consuming 3. answer a question about a population 4. obtain statistical results with sufficiently high precision	1. sample for infinite population 2. sample for finite population

<u>Example 1</u> Members of a political part in Texas were considering supporting a particular candidate for election to the U.S. Senate, and party leaders wanted to estimate the proportion of registered voters in the state favoring the candidate. A sample of 400 registered voters in Texas was selected, and 160 of the 400 voters indicated a preference for the candidate. Thus, an estimate of the proportion of the population of registered voters favoring the candidate is $\frac{160}{400} = 0.40$.

a. Identify the element(s)

b. Identify the population.

<u>Example 2</u> A study done on college drinking at a university reported the number of drinks consumed weekly by a female senior has a mean 2.3. We take a random sample of 33 female seniors to estimate the mean of the number of drinks consumed weekly by a female senior. Thus, an estimate of the mean number of drinks was 2.3 drinks.

a. Identify the element(s)

b. Identify the population.

SAMPLING REVIEW

- The reason we select a sample is to ________________________________to answer a research question about the ________________________________.
- The sample results only give ________________________________ of the true values of the population characteristics, since the sample only contains a __________________ of a population.
- With adequate ____________________________________, the sample results can provide a "good" estimate of the population characteristics.

MATH 1610/MATH 1553
7.2 Selecting a Sample

<u>Question</u>: How do we select a sample so that we can make statistical inferences about a population?

<u>Answer</u>: It is recommended to select a ***probability sample*** because it provides a tool for valid statistical inferences about the population.

<u>DEFINITIONS</u>

A **<u>Simple Random Sample</u>** of size n from a ***finite population*** of size N is a sample selected such that each possible sample of size n has the same probability of being selected.

<u>Finite Populations</u> – often defined by a list such as:

- membership roster of an organization
- credit card account numbers
- inventory product numbers

<u>Sampling with Replacement</u> – replacing each sample element before selecting a subsequent element

<u>Sampling without Replacement</u> - (used most often) once an element has been selected, it cannot be selected again

<u>Example of Finite Population</u>
St. Andrew's College received 900 applications for admission in the upcoming year from prospective students. The applications were numbered, from 1 to 900, as their applications arrived. The Director of Admissions would like to select a simple random sample of 10 applicants.

How could we select a simple random sample that would represent the population, and that ensures that each element has the same probability of being selected?

<u>Step 1</u> Assign a random number to each of the 900 applicants.
The random numbers generated by Excel's *RAND* function follow a uniform probability between 0 and 1.

<u>Step 2</u> Select the first 10 random numbers. (See table 7.1 p. 300)

63271	59986
88547	09896
55957	57243
46276	87453
55363	07449

REVIEW CHAPTER 3

POPULATION FORMULAS	SAMPLE FORMULAS
Population Mean $\boldsymbol{\mu = \frac{\sum_{i=1}^{N} x_i}{N}}$ where N is the population size (equally weighted)	**Sample Mean** $\boldsymbol{\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}}$ where n is the sample size (equally weighted)
Population Variance $\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$	**Sample Variance** $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$
Population Standard Deviation $\sigma = \sqrt{\sigma^2}$	**Sample Standard Deviation** $s = \sqrt{s^2}$

1. (3.2 Exercise 25) Consider a sample data with values of 27, 25, 20, 15, 30, 34, 28, and 25. Compute the mean, variance, and standard deviation.

2. (3.2 Exercise 24) Consider a sample data with values of 10, 20, 12, 17, and 16. Compute the mean, variance, and standard deviation.

MATH 1610/MATH 1553
7.3 Point Estimation

Point Estimation – a form of statistical inference
- we use the data from the sample to compute a sample statistic that serves as an estimate of a population parameter

Population Statistic	**Sample Statistic (Point Estimators)**
Mean ($\boldsymbol{\mu}$)	Mean ($\bar{x}$)
Standard Deviation (σ)	Standard Deviation (s)
Proportion $P = \frac{x}{N}$ where N= population	Proportion $\bar{p} = \frac{x}{n}$ where n = sample size

Example 1

St. Andrew's College received 900 applications from prospective students. The application from contains different information including SAT scores and whether or not an individual desires to live on campus.

At a meeting in a few hours, the Director of Admissions would like to announce the average SAT score and the proportion of applicants that want to live on campus, for the population of 900 applicants. However, not all of the necessary data for each applicant has been entered in the college database. Hence, the Director decides to estimate the parameters of interest based on sample statistics.

Let's suppose that a sample of 30 applicants is selected using computer generated random numbers. Using data from these random numbers the point estimators are:

$\bar{x}$ is a point estimator of μ:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{32{,}900}{30} = 1097$$

s as a point estimator of σ:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{163{,}996}{29}} = 75.2$$

$\bar{p}$ is a point estimator of p: (assume that 20 of the 30 students want to live on campus)

$$\bar{p} = \frac{x}{n} = \frac{20}{30} \approx 0.67$$

****NOTE**: Different random numbers would yield different sample values and therefore resulted in different point estimators

Example 1 (con't)

Once all of the data for the 900 applicants were entered in the college's database, the values of the parameters of interest were calculated.

Population Statistic	Sample Statistic (Point Estimators)
$\mu = \frac{\sum x_i}{N} = \frac{981{,}000}{900} = 1090$	$\bar{x} = \frac{\sum x_i}{n} = \frac{32{,}900}{30} = 1097$
$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}} = \sqrt{\frac{5{,}760{,}000}{900}} = 80$	$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{163{,}996}{29}} = 75.2$
$p = \frac{x}{N} = \frac{648}{900} \approx 0.72$	$\bar{p} = \frac{x}{n} = \frac{20}{30} \approx 0.67$

Compare the values in the Population Statistic column with the values in the Sample Statistic column, what can you conclude?

DEFINITIONS

Target population – the population about which we want to make inference

Sample population – the population from which the sample is actually taken

Whenever a sample is used to make inferences about a population, we should make sure that the targeted population and the sampled population are in close agreement.

<u>Example 1</u> The following data are from a ***simple random sample***:

3 5 13 4 11 6

a. What is the point estimate of the population mean?

b. What is the point estimate of the population standard deviation?

<u>Example 2</u> (7.3 exercise 14) Morningstar publishes ratings data on 1208 company stocks. A sample of 40 of these stocks is contained in the data set on the following page. Use the data to answer the following:

a. Develop a point estimate of the proportion of the stocks that receive Morningstar's highest rating of 5 stars.

b. Develop a point estimate of the proportion of the Morningstar stocks that are rated Above Average with respect to business risk.

c. Develop a point estimate of the proportion of the Morningstar stocks that are rated 2 stars or less.

MORNING STAR RATINGS DATA (7.3 exercise 14, Oct. 24, 2012)

Company	Morningstar Rating	Business Risk
Ameren Corp	3 Star	Average
Apple Inc	4 Star	Average
At&T Inc	2 Star	Average
Bank of New York Mellon Corp	4 Star	Above Average
BlackRock Inc	4 Star	Average
Buckeye Partners	4 Star	Below Average
Cardinal Health Inc	3 Star	Average
CenturyLink Inc	3 Star	Above Average
Cintas Corporation	2 Star	Average
Colgate-Palmolive Company	2 Star	Below Average
Corning Inc	4 Star	Above Average
Darden Restaurants Inc	3 Star	Average
Discover Financial Services	3 Star	Above Average
eBay Inc	3 Star	Average
Endo Health Solutions	4 Star	Above Average
Fair Isaaac Corp	2 Star	Average
Fluor Corp	3 Star	Above Average
Gentex Corporation	5 Star	Average
Group 1 Automotive Inc	2 Star	Above Average
Hess Corp	4 Star	Average
Icon PLC	4 Star	Above Average
International Flavors & Fragrances	3 Star	Average
John Wiley & Sons, Inc.	4 Star	Average
Knight Transportation, Inc.	4 Star	Average
Life Technologies corp	4 Star	Average
Marvell Technology Group, Ltd.	4 Star	Above Average
Methanex Corporation	3 Star	Above Average
Mylan Inc	3 Star	Above Average
New York Times Company	1 Star	Above Average
OfficeMax Inc	1 Star	Above Average
Owens-Corning, Inc.	3 Star	Above Average
Polaris Industries, Inc.	3 Star	Above Average
Rio Tinto PLC	5 Star	Average
Sempra Energy	2 Star	Average
Stifel Financial Corp.	3 Star	Above Average
Texas Instruments, Inc.	4 Star	Average
United Technologies Corp	4 Star	Average
Weight Watchers International, Inc.	4 Star	Above Average
WR Berkley Corp	4 Star	Average
Yum Brands Inc	3 Star	Average

MATH 1610/MATH 1553

7.4 & 7.5 Sampling Distribution of $\bar{x}$

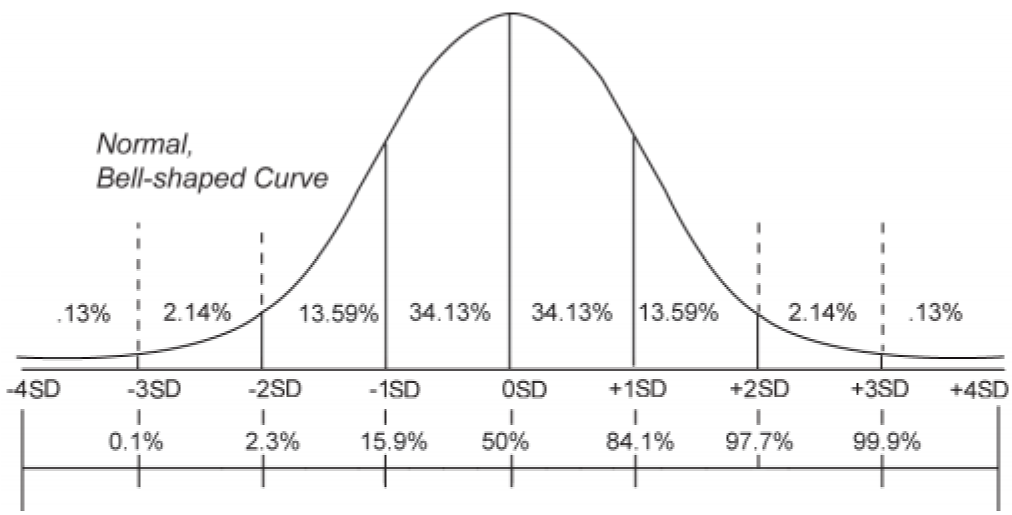
In ***inferential statistics***, we want to use characteristics of the sample (ie a ***statistic***) to estimate the characteristics of the population (ie a ***parameter***).

Example

Consider the population of students on the university campus. If each of you asked 10 students for their height and calculated the average of those heights, would all of you come back with the same average of your sample? How do you suppose your sample averages would compare?

Probability plays an important role in inferential statistics; however, variability exists.

Where do we expect most of the average sample heights to occur compared to the average of the average height of the student population?



SAMPLING DISTRIBUTIONS

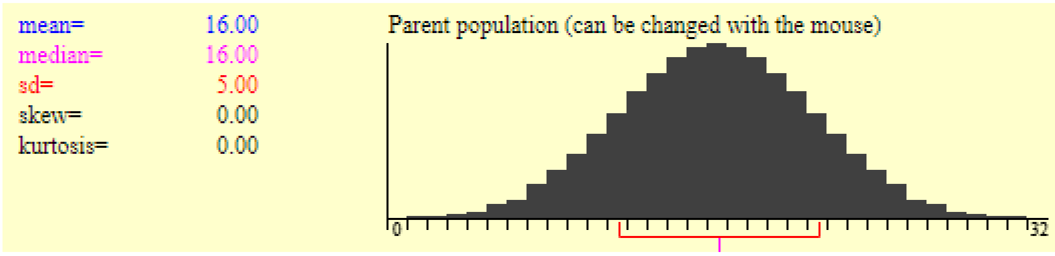
If we consider the process of selecting a simple random sample as an experiment, then $\bar{x}$ is the outcome of the experiment, which we can consider the random variable of the experiment. This means that $\bar{x}$ has an expected value, standard deviation, and probability distribution.

Let's look at some sampling distributions with an online simulator: onlinestatbook.com

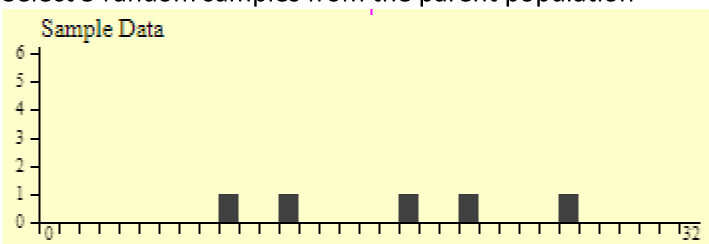
Simulations provided by:

Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.

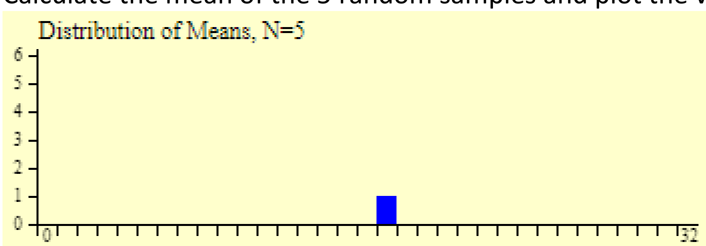
From the parent population 5 samples are randomly selected (n = 5)



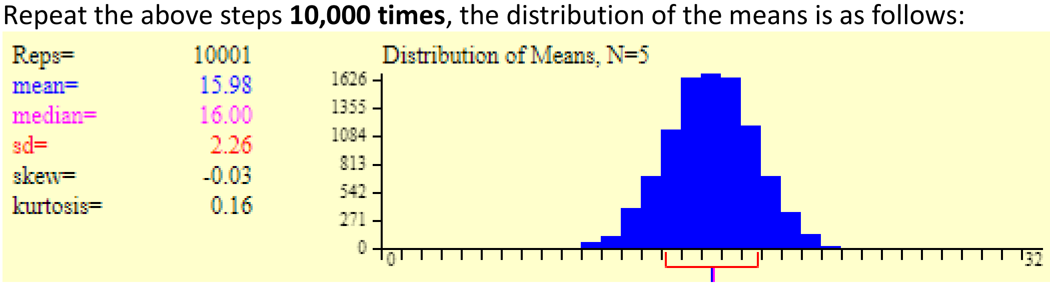
Step 1: Select 5 random samples from the parent population



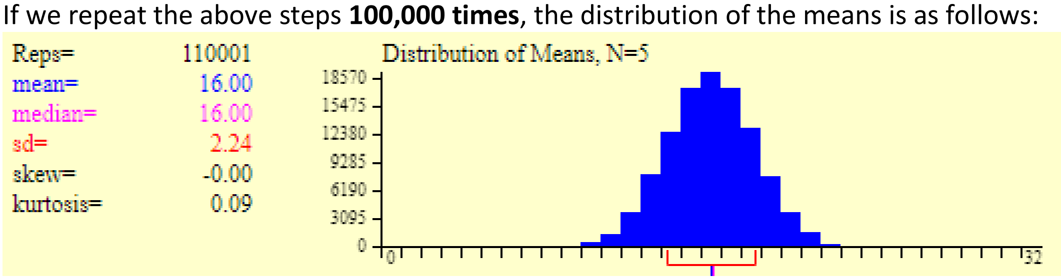
Step 2: Calculate the mean of the 5 random samples and plot the value.



Repeat the above steps **10,000 times**, the distribution of the means is as follows:

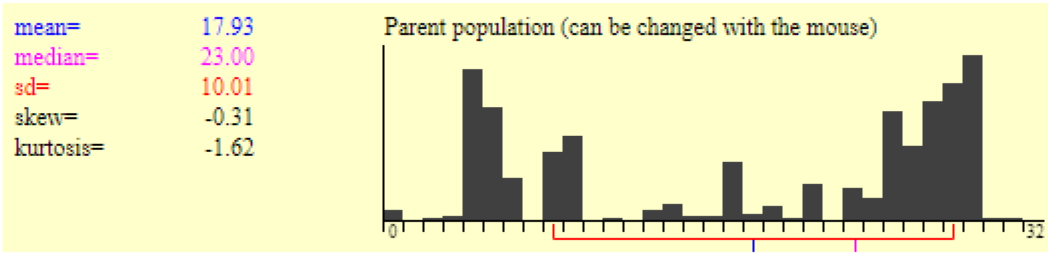


If we repeat the above steps **100,000 times**, the distribution of the means is as follows:



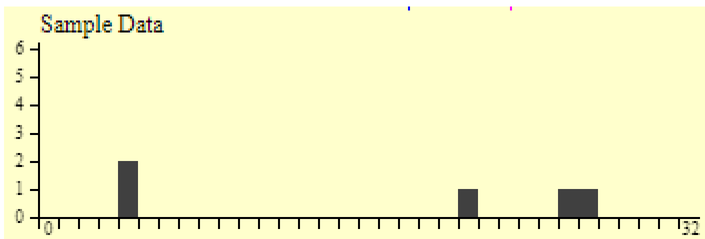
What do you notice about the mean? Standard deviation?

What happens if the parent population does not represent a normal distribution?

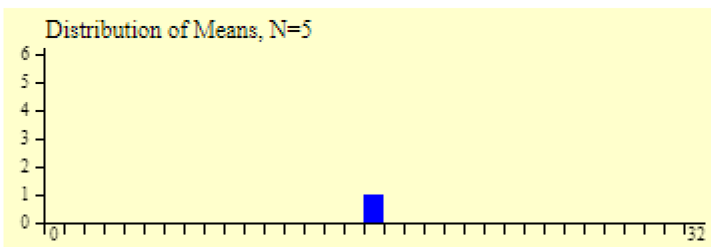


From the parent population 5 samples are randomly selected (n = 5)

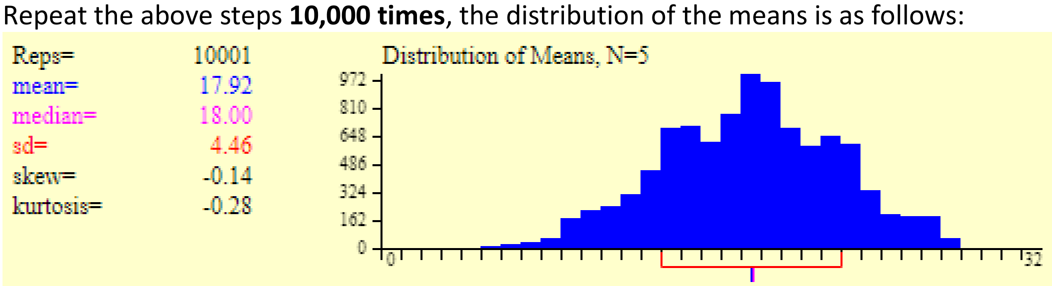
Step 1: Select 5 random samples from the parent population



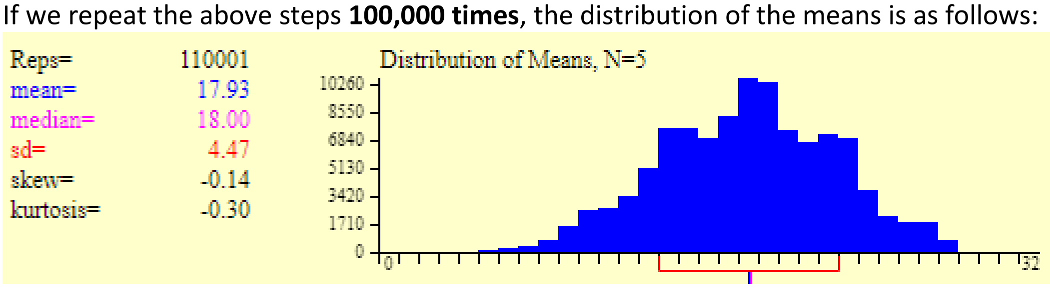
Step 2: Calculate the mean of the 5 random samples and plot the value.



Repeat the above steps **10,000 times**, the distribution of the means is as follows:



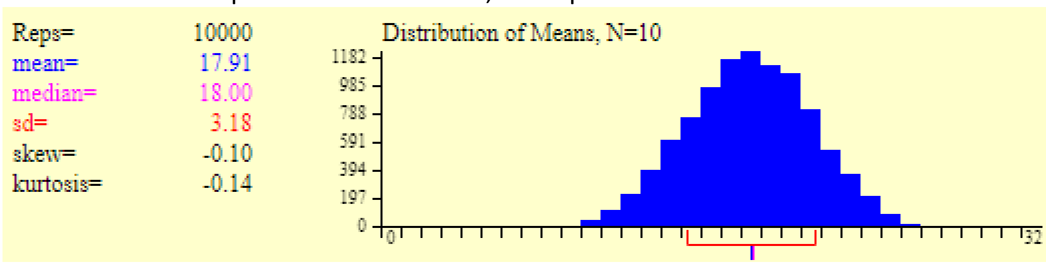
If we repeat the above steps **100,000 times**, the distribution of the means is as follows:



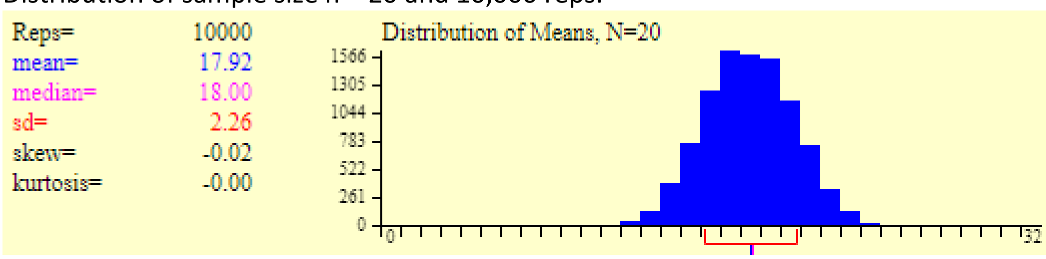
What do you notice about the mean? Standard deviation?

What happens if we increase our sample size?
We have only been picking 5 random samples, but what if we pick 10, 20, or 25 samples?

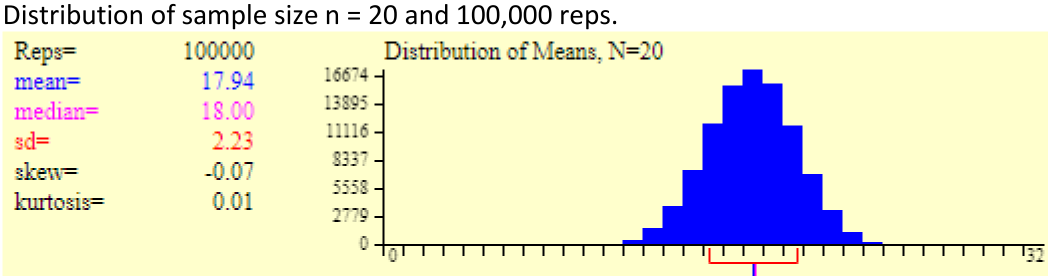
Distribution of sample size n = 10 and 10,000 reps.



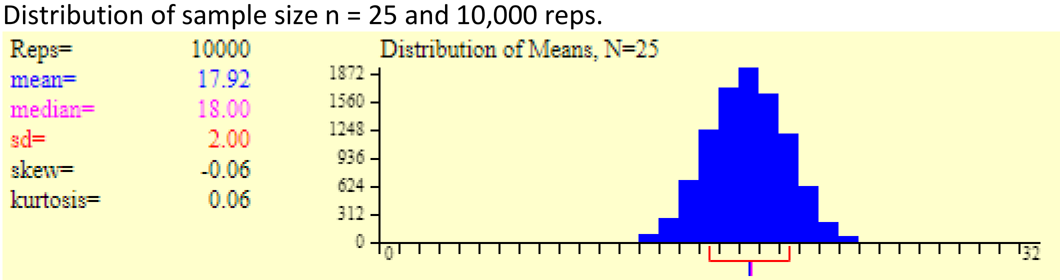
Distribution of sample size n = 20 and 10,000 reps.



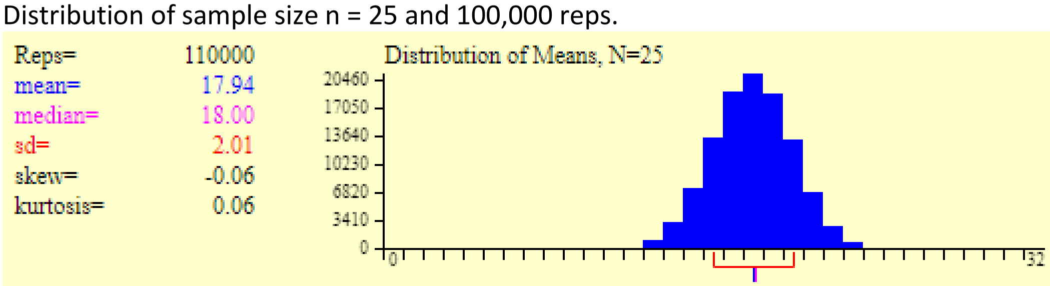
Distribution of sample size n = 20 and 100,000 reps.



Distribution of sample size n = 25 and 10,000 reps.



Distribution of sample size n = 25 and 100,000 reps.



SAMPLING DISTRIBUTION OF $\bar{x}$

The sampling distribution of $\bar{x}$ is the probability distribution of all possible values of the sample mean $\bar{x}$.

EXPECTED VALUE of $\bar{x}$ = $E(\bar{x})$

$$E(\bar{x}) = \mu \text{ where } \mu \text{ is the population mean}$$

When the expected value of the point estimator equals the population parameter, we say the point estimator is <u>unbiased</u>.

STANDARD DEVIATION of $\bar{x}$

Notation used to define the standard deviation of the sampling distribution of $\bar{x}$

- $\sigma_{\bar{x}}$ = the standard deviation of $\bar{x}$
- σ = the standard deviation of the population
- n = the sample size
- N = the population size

<u>Finite Population</u>:

$$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Called the finite population correction factor.
Use only if $\frac{n}{N} > 0.5$, otherwise use $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

<u>Infinite Population</u>
** used most often**

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

$\sigma_{\bar{x}}$ is referred to as the standard error of the mean.

When the population from which we are selecting a random sample does NOT have a normal distribution, the Central Limit Theorem helps identify the shape of the sampling distribution of $\bar{x}$.

CENTRAL LIMIT THEOREM

In selecting random samples of size n from a population, the sampling distribution of the sample mean $\bar{x}$ can be approximated by a normal distribution as the sample size becomes large.

In general, regarding the standard deviation of $\bar{x}$:

- When the population has a ***normal distribution***, the sampling distribution of $\bar{x}$ is normally distributed for any sample size.
- Usually, the sampling distribution of $\bar{x}$ can be approximated by a normal distribution whenever the sample is size 30 or more.
- When the population is highly skewed or outliers are present, larger samples (50 or more) may be needed.
- The sampling distribution of $\bar{x}$ can be used to provide probability information about how close the sample mean $\bar{x}$ is to the population mean $\boldsymbol{\mu}$.

Question: How do we use the sampling distribution of $\bar{x}$ to solve problems?

St. Andrew's College Example

Given the sampling distribution of $\bar{x}$ for SAT scores as:

$$E(x) = \mu = 1090 \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{80}{\sqrt{30}} = 14.6059$$

What is the probability that a simple random sample of 30 applicants will provide an estimate of the population mean SAT scores that is within ±10 of the actual population mean μ?
In other words, what is the probability that $\bar{x}$ will be between 1080 and 1100?

Step 1: Calculate the z-value at the ***upper*** endpoint of the interval.

$$z = \frac{1100-1090}{14.6059} = 0.68$$

Step 2: Find the area under the curve to the left of the ***upper*** endpoint.

$$P(z \le 0.68) = 0.7517$$

Step 3: Calculate the z-value at the ***lower*** endpoint of the interval.

$$z = \frac{1080-1090}{14.6059} = -0.68$$

Step 4: Find the area under the curve to the left of the ***lower*** endpoint.

$$P(z \le -0.68) = 0.2483$$

Step 5: Calculate the area under the curve between the upper & lower endpoints.

$$P(-0.68 \le z \le 0.68) = 0.7517 - 0.2483 = 0.5034$$

Therefore, the probability that the sample mean SAT score will be between 1080 and 1100 is:

$$P(1080 \le \bar{x} \le 1100) = 0.5034$$

Another Example

In February 2006, the mean annual cost of auto insurance was $\mu = \$939$ with a standard deviation of $\sigma = \$245$. What is the probability that a simple random sample of auto insurance policies will have a sample mean within $25 of the population mean if the sample size is the following:

(a) 30 (b) 50 (c) 100 (d) 400

After performing each calculation, what can you conclude about the standard error $\sigma_{\bar{x}}$ as the value of n increases?

Solution:

Solution (con't)

Math 1610/Math 1553
Worksheet Section 7.5

1. An automotive repair shop has determined that the average service time on an automobile is 2 hours with a standard deviation of 32 minutes. A random sample of 64 services is selected.

   a. What is the probability that the sample of 64 will have a mean service time greater than 114 minutes?

   b. Assume the population consists of 400 services. Determine the standard error of the mean.

2. There are 8,000 students at the University of Tennessee at Chattanooga. The average age of all the students is 24 years with a standard deviation of 9 years. A random sample of 36 students is selected.

a. Determine the standard error of the mean.

b. What is the probability that the sample mean will be larger than 19.5?

c. What is the probability that the sample mean will be between 25.5 and 27 years?

Math 1553 Business Stats Name ___________________________

Section 7.5 (Extra Practice) Date_____________________________

1. A professor at a local community college noted that the latest test grades of his students were normally distributed with a mean of 74 and a standard deviation of 10.
The professor informed his class that 6.3% of his students earned As while only 2.5% earned Fs.

a. What is the minimum score needed to earn an A on the test?

b. What is the maximum score among the students who earned a F?

c. If there were 5 students who earned Fs, how many students took the test?

MATH 1610/MATH 1553

7.6 Sampling Distribution of $\bar{p}$

In many cases, we will be interested in the proportion of a population that satisfies a certain characteristic and the corresponding sample statistics, the sample proportion.

For example, if we have a population in which 66% of all adults use the Internet, then we say p=0.66. In general, the proportion of a population that satisfies a certain characteristic is:

$$p = \frac{x}{N}$$ where x = # of elements satisfying the characteristic
N = # of elements in the population

Example Suppose we want to know what percent of the U.S. population is left-handed.
It would be impossible to check every single person, so we could take a sample of 500 and determine the sample proportion. This would give us a good idea of how many people in the U.S. are left-handed.

Suppose that 75 people from the 500 people sampled are left-handed.

Therefore, the sample proportion is: $\bar{p} = \frac{x}{n}$

$$\bar{p} = \frac{75}{500}$$

$$\bar{p} = 0.15$$

Depending on the group of people selected, the sample proportion could result in a value higher than 0.15 or lower than 0.15.

Therefore, similar to $\bar{x}$, we could consider the probability distribution of the sample mean $\bar{p}$ called the **sampling distribution of $\bar{p}$.**

The sampling distribution of $\bar{p}$ can be approximated by a normal distribution whenever $np \geq 5$ and $n(1-p) \geq 5$

Sample Proportion $\bar{p}$

The Sample Proportion $\bar{p}$ is the point estimator of the population proportion p

$$\bar{p} = \frac{x}{n}$$

where x = the number of elements in the sample that has the characteristic of interest
n = sample size

Sampling Distribution of $\bar{p}$

The sampling distribution of $\bar{\boldsymbol{p}}$ is the probability distribution of all possible values of the sample mean $\bar{\boldsymbol{p}}$.

Expected Value of $\bar{p}$

The expected value of $\bar{\boldsymbol{p}}$ is the mean of all possible values of $\bar{\boldsymbol{p}}$ and equals the population proportion.

$$\boldsymbol{E(\bar{p}) = p}$$

where $E(\bar{p})$ = the expected value of $\bar{p}$

p = the proportion of the population

Recall when the expected value of a point estimator equals the population parameter, we say the point estimator is unbiased, so $\bar{p}$ is an ***unbiased*** estimator of p.

Standard Deviation of the Sampling Distribution of $\bar{p}$

Finite Population

$$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}} \cdot \sqrt{\frac{p\,(1-p)}{n}}$$

The finite population correction factor
Use only if $\frac{n}{N} > 0.5$

Infinite Population

$$\sigma_{\bar{p}} = \sqrt{\frac{p\,(1-p)}{n}}$$

where p = the population proportion
n = the sample size

SUMMARY

Parameter	Population distribution	Sample	Sampling Distribution of $\bar{p}$
Mean	$\mu = np$	$\bar{p} = \frac{x}{n}$	$\bar{p}$ and $E(\bar{p}) = p$
Standard Deviation	$\sigma = \sqrt{np(1-p)}$		$\sigma_{\bar{p}} = \sqrt{\frac{p\,(1-p)}{n}}$

Example 1

Suppose we know that 66% ($p = 0.66$) of adults in the U.S. use the Internet. Consider a sample of 300 adults. What is the probability that the sample proportion of adults using the Internet will be within ±0.04 of the population proportion?

Example 2

Of the prospective students applying to St. Andrew's College, 72% desire on-campus housing. What is the probability that a simple random sample of 30 applicants will provide an estimate of the population proportion of applicants desiring on-campus housing that is within ±0.05 of the actual population proportion?

MATH 1610/1553 Name______________________________

Review 7.5 & 7.6 WS

1. The average weekly salary for computer programmers in a city is $1750 with a population standard deviation of $400.

a. What percent of the programmers make more than $1850 a week? Answer with a sentence.

b. From a sample of 60 programmers, what is the probability that the average will be over $1850 a week? Answer with a sentence.

2. At a university 40% of the students live in the dorms. If a random sample of 80 students is used in a study, what is the probability that the sample proportion of students living is the dorms is between 0.3 and 0.45?