STT 315, Summer A, 2019
Lectures 7 and 8

Arnab Bhattacharjee

arnab@msu.edu

Chapter 5
Sampling Distributions

5 Sampling and Sampling Distributions

- Using Statistics
- Sample Statistics as Estimators of Population Parameters
- Sampling Distributions
- Estimators and Their Properties
- Degrees of Freedom

5 LEARNING OBJECTIVES

After studying this chapter you should be able to:

- Take random samples from populations
- Distinguish between population parameters and sample statistics
- Apply the central limit theorem
- Derive sampling distributions of sample means and proportions
- Explain why sample statistics are good estimators of population parameters
- Judge one estimator as better than another based on desirable properties of estimators
- Apply the concept of degrees of freedom
- Identify special sampling methods
- Compute sampling distributions and related results

5-1 Using Statistics

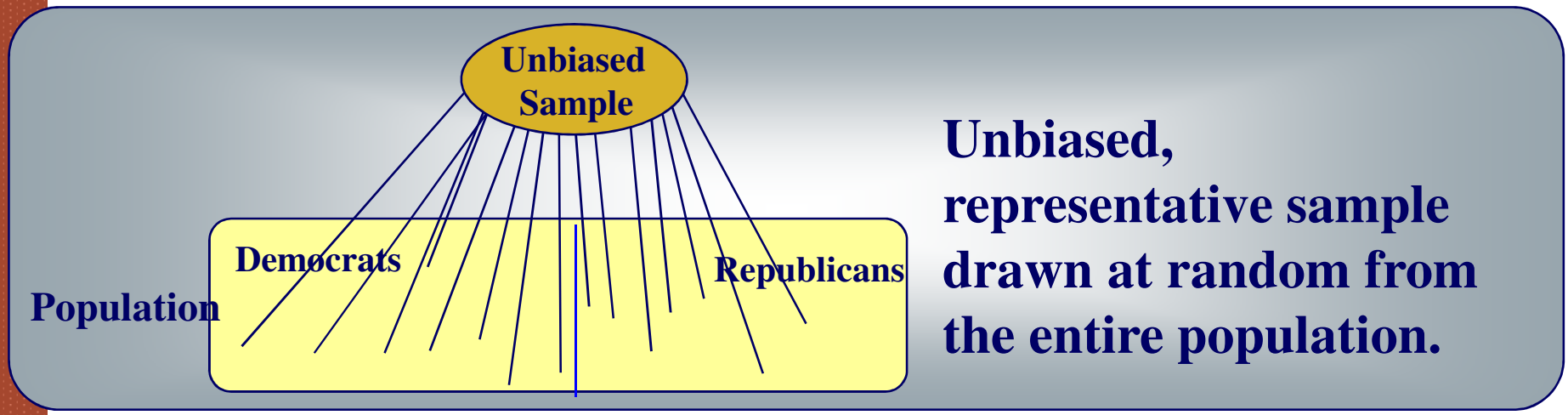
- Statistical Inference:
  - ✓ Predict and forecast values of ***population parameters...***
  - ✓ Test hypotheses about values of population parameters...
  - ✓ Make decisions...

On basis of ***sample statistics*** derived from limited and incomplete sample information.

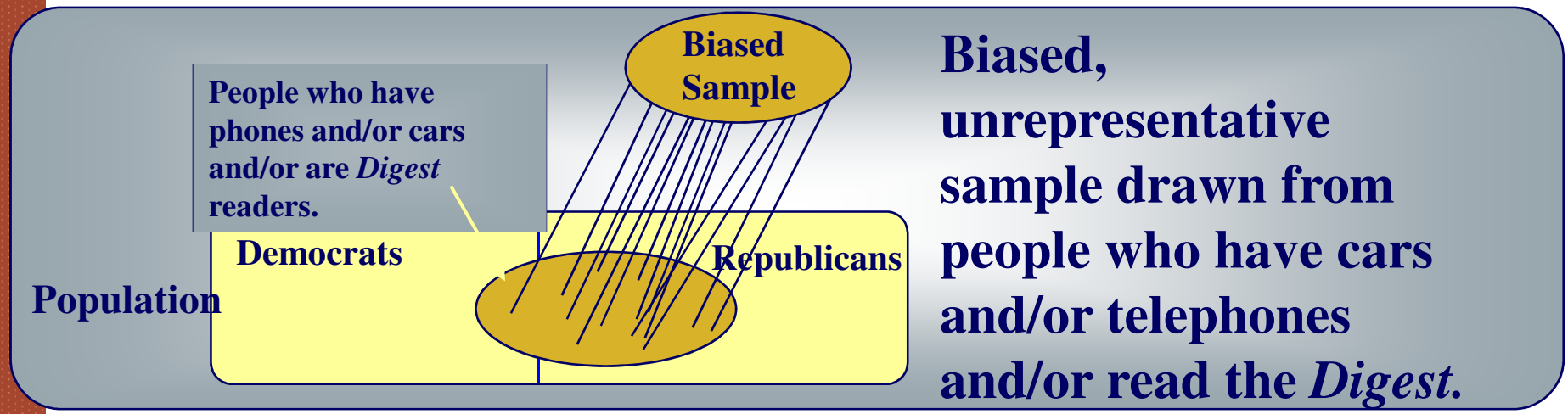
Make generalizations about the characteristics of a ***population...***

On the basis of observations of a ***sample***, a part of a population

The Literary Digest Poll (1936)



Unbiased, representative sample drawn at random from the entire population.



Biased, unrepresentative sample drawn from people who have cars and/or telephones and/or read the *Digest*.

5-2 Sample Statistics as Estimators of Population Parameters

- A **sample statistic** is a numerical measure of a summary characteristic of a sample.

A **population parameter** is a numerical measure of a summary characteristic of a population.

- An **estimator** of a population parameter is a sample statistic used to estimate or predict the population parameter.
- An **estimate** of a parameter is a *particular* numerical value of a sample statistic obtained through sampling.
- A **point estimate** is a single value used as an estimate of a population parameter.

Estimators

- The sample mean, $\bar{x}$, is the most common estimator of the population mean, μ.
- The sample variance, s^2, is the most common estimator of the population variance, σ^2.
- The sample standard deviation, s, is the most common estimator of the population standard deviation, σ.
- The sample proportion, $\hat{p}$, is the most common estimator of the population proportion, p.

Estimators

Estimator (Sample Statistic)	→	Population Parameter
$\overline{X}$	*estimates*	μ
S^2	*estimates*	σ^2
S	*estimates*	σ
$\hat{p}$	*estimates*	p

Population and Sample Proportions

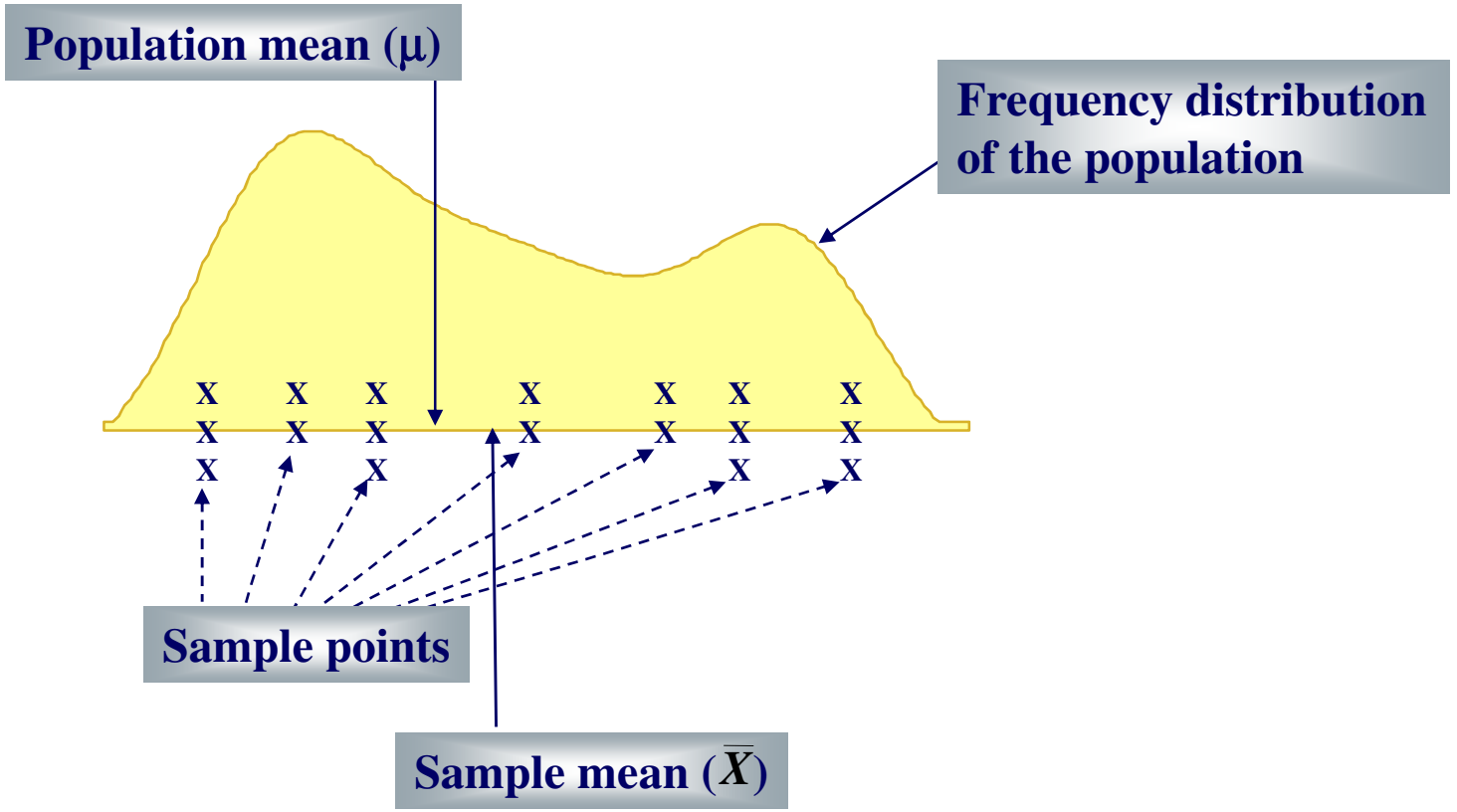
- The **population proportion** is equal to the number of elements in the population belonging to the category of interest, divided by the total number of elements in the population:

$$p = \frac{X}{N}$$

- The **sample proportion** is the number of elements in the sample belonging to the category of interest, divided by the sample size:

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

A Population Distribution, a Sample from a Population, and the Population and Sample Means



Common Sampling Methods

- **(Simple) Random sampling:** every unit in the population has the same chance of being included in the sample – sampling is unbiased – most common, most popular.
- **Stratified sampling:** in stratified sampling, the population is partitioned into two or more subpopulation called strata, and from each stratum a desired sample size is selected at random.
- **Cluster sampling:** in cluster sampling, a random sample of the strata is selected and then samples from these selected strata are obtained.
- **Single-stage Cluster sampling:** in single-stage cluster sampling, a cluster is chosen at random and ***every*** item or person in that cluster is sampled.

Other Sampling Methods

- **Two-stage Cluster sampling:** in two-stage cluster sampling, a cluster is chosen at random and random samples of the items or persons are taken from that cluster.
- **Multistage Cluster sampling:** in multistage cluster sampling, a random sample of the strata is selected, then samples of items or persons from these strata are taken and then samples from these items and persons are selected.

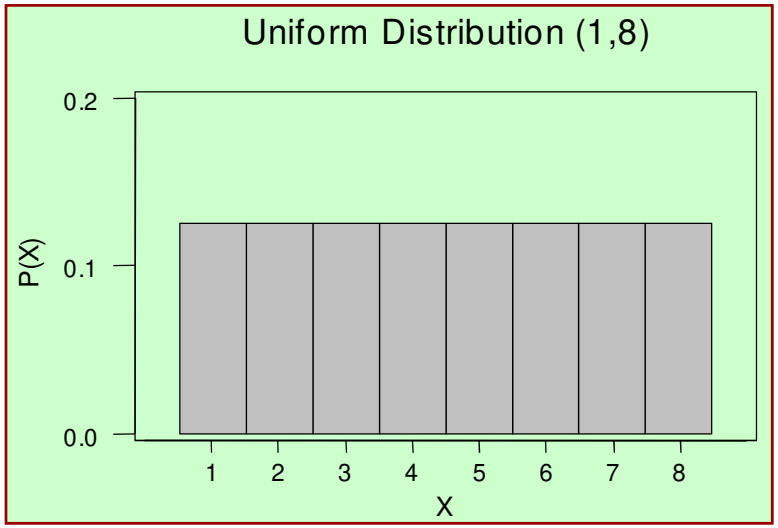
5-3 Sampling Distributions

- The **sampling distribution** of a statistic is the probability distribution of all possible values the statistic may assume, when computed from random samples of the same size, drawn from a specified population.
- The **sampling distribution of $\overline{X}$** is the probability distribution of all possible values the random variable $\overline{X}$ may assume when a sample of size n is taken from a specified population.

Sampling Distributions (Continued)

Uniform population of integers from 1 to 8:

X	P(X)	XP(X)	$(X-\mu_x)$	$(X-\mu_x)^2$	$P(X)(X-\mu_x)^2$
1	0.125	0.125	-3.5	12.25	1.53125
2	0.125	0.250	-2.5	6.25	0.78125
3	0.125	0.375	-1.5	2.25	0.28125
4	0.125	0.500	-0.5	0.25	0.03125
5	0.125	0.625	0.5	0.25	0.03125
6	0.125	0.750	1.5	2.25	0.28125
7	0.125	0.875	2.5	6.25	0.78125
8	0.125	1.000	3.5	12.25	1.53125
	1.000	4.500			5.25000



$E(X) = \mu = 4.5$
$V(X) = \sigma^2 = 5.25$
$SD(X) = \sigma = 2.2913$

Sampling Distributions (Continued)

- There are 8*8 = 64 different but equally-likely samples of size 2 that can be drawn (with replacement) from a uniform population of the integers from 1 to 8:

	Samples of Size 2 from Uniform (1,8)							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8
2	2,1	2,2	2,3	2,4	2,5	2,6	2,7	2,8
3	3,1	3,2	3,3	3,4	3,5	3,6	3,7	3,8
4	4,1	4,2	4,3	4,4	4,5	4,6	4,7	4,8
5	5,1	5,2	5,3	5,4	5,5	5,6	5,7	5,8
6	6,1	6,2	6,3	6,4	6,5	6,6	6,7	6,8
7	7,1	7,2	7,3	7,4	7,5	7,6	7,7	7,8
8	8,1	8,2	8,3	8,4	8,5	8,6	8,7	8,8

Each of these samples has a sample mean. For example, the mean of the sample (1,4) is 2.5, and the mean of the sample (8,4) is 6.

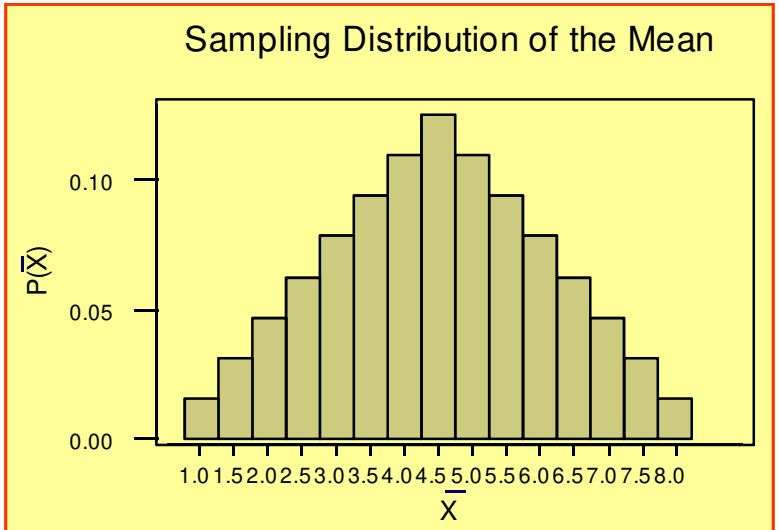
	Sample Means from Uniform (1,8), n							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5
2	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
3	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5
4	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0
5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	6.5
6	3.5	4.0	4.5	5.0	5.5	6.0	6.5	7.0
7	4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5
8	4.5	5.0	5.5	6.0	6.5	7.0	7.5	8.0

Sampling Distributions (Continued)

The probability distribution of the sample mean is called the **sampling distribution of the the sample mean**.

Sampling Distribution of the Mean

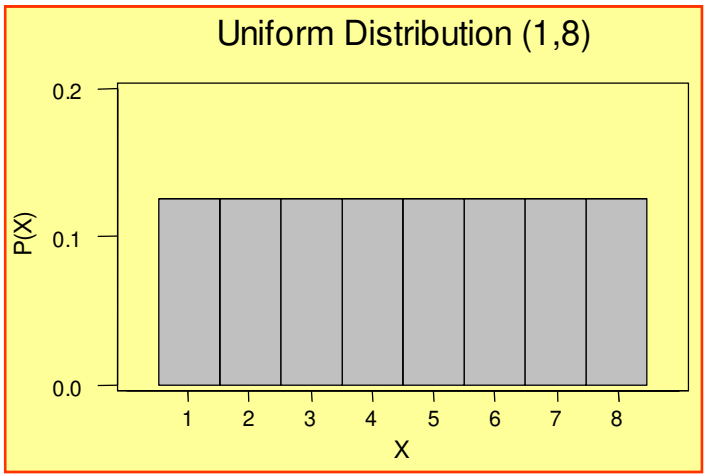
$\bar{X}$	$P(\bar{X})$	$\bar{X}P(\bar{X})$	$\bar{X}-\mu_{\bar{X}}$	$(\bar{X}-\mu_{\bar{X}})^2$	$P(\bar{X})(\bar{X}-\mu_{\bar{X}})^2$
1.0	0.015625	0.015625	-3.5	12.25	0.191406
1.5	0.031250	0.046875	-3.0	9.00	0.281250
2.0	0.046875	0.093750	-2.5	6.25	0.292969
2.5	0.062500	0.156250	-2.0	4.00	0.250000
3.0	0.078125	0.234375	-1.5	2.25	0.175781
3.5	0.093750	0.328125	-1.0	1.00	0.093750
4.0	0.109375	0.437500	-0.5	0.25	0.027344
4.5	0.125000	0.562500	0.0	0.00	0.000000
5.0	0.109375	0.546875	0.5	0.25	0.027344
5.5	0.093750	0.515625	1.0	1.00	0.093750
6.0	0.078125	0.468750	1.5	2.25	0.175781
6.5	0.062500	0.406250	2.0	4.00	0.250000
7.0	0.046875	0.328125	2.5	6.25	0.292969
7.5	0.031250	0.234375	3.0	9.00	0.281250
8.0	0.015625	0.125000	3.5	12.25	0.191406
	1.000000	**4.500000**			**2.625000**

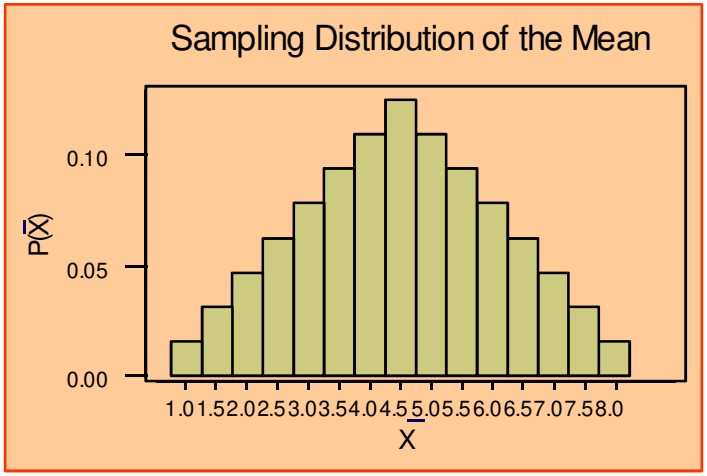


$$E(\bar{X}) = \mu_{\bar{X}} = 4.5$$
$$V(\bar{X}) = \sigma^2_{\bar{X}} = 2.625$$
$$SD(\bar{X}) = \sigma_{\bar{X}} = 1.6202$$

Properties of the Sampling Distribution of the Sample Mean

- Comparing the population distribution and the sampling distribution of the mean:
  - ✓ **The sampling distribution is more bell-shaped and symmetric.**
  - ✓ **Both have the same center.**
  - ✓ **The sampling distribution of the mean is more compact, with a smaller variance.**





Relationships between Population Parameters and the Sampling Distribution of the Sample Mean

The **expected value of the sample mean** is equal to the population mean:

$$E(\bar{X}) = \mu_{\bar{X}} = \mu_X$$

The **variance of the sample mean** is equal to the population variance divided by the sample size:

$$V(\bar{X}) = \sigma^2_{\bar{X}} = \frac{\sigma^2_X}{n}$$

The **standard deviation of the sample mean, known as the standard error of the mean**, is equal to the population standard deviation divided by the square root of the sample size:

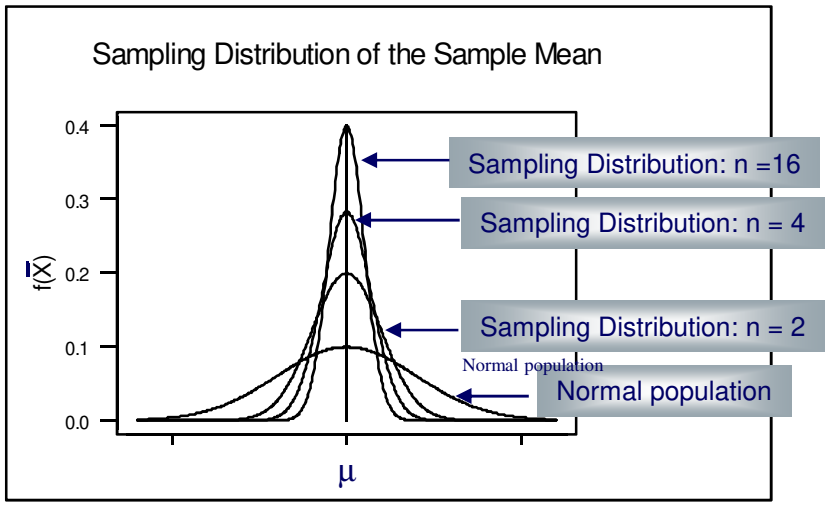
$$SD(\bar{X}) = \sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$$

Sampling from a Normal Population

When sampling from a **normal population** with mean μ and standard deviation σ, the sample mean, $\overline{X}$, has a **normal sampling distribution:**

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

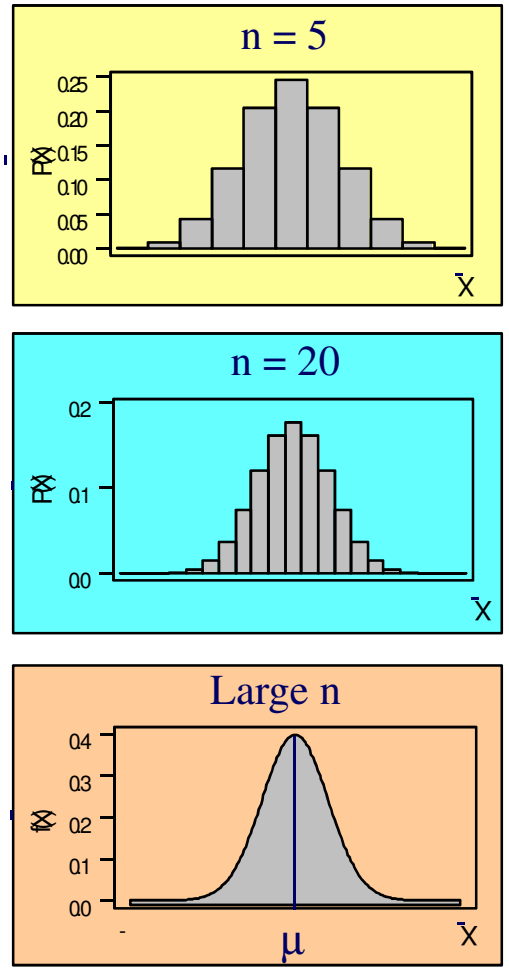
This means that, as the sample size increases, the sampling distribution of the sample mean remains centered on the population mean, but becomes more compactly distributed around that population mean.



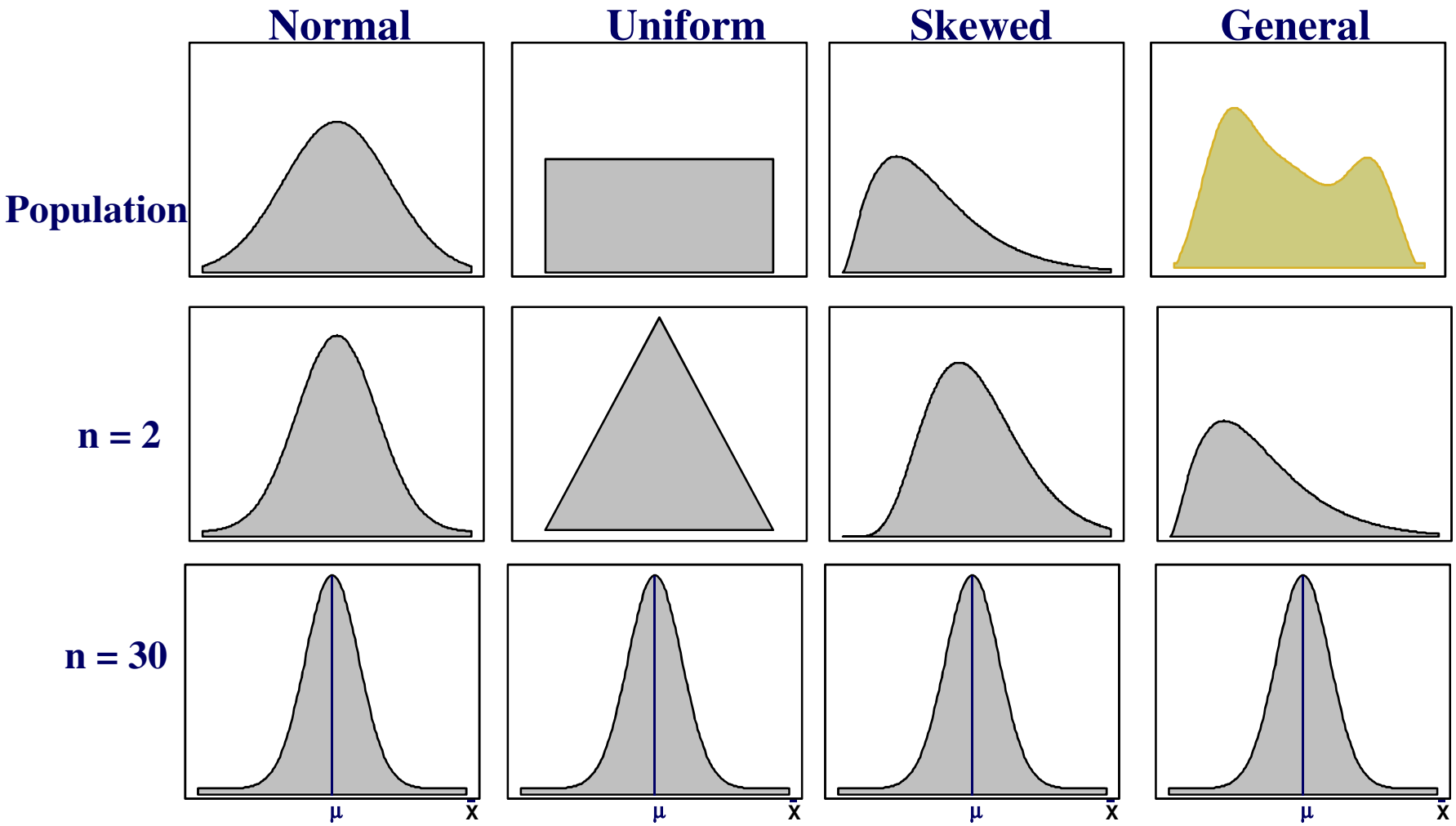
The Central Limit Theorem

When sampling from a population with mean μ and finite standard deviation σ, the sampling distribution of the sample mean will tend to a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ as the sample size becomes large (n >30).

For "large enough" n: $\bar{X} \sim N(\mu, \sigma^2 / n)$



The Central Limit Theorem Applies to Sampling Distributions from Any Population



The Central Limit Theorem (Example 5-1)

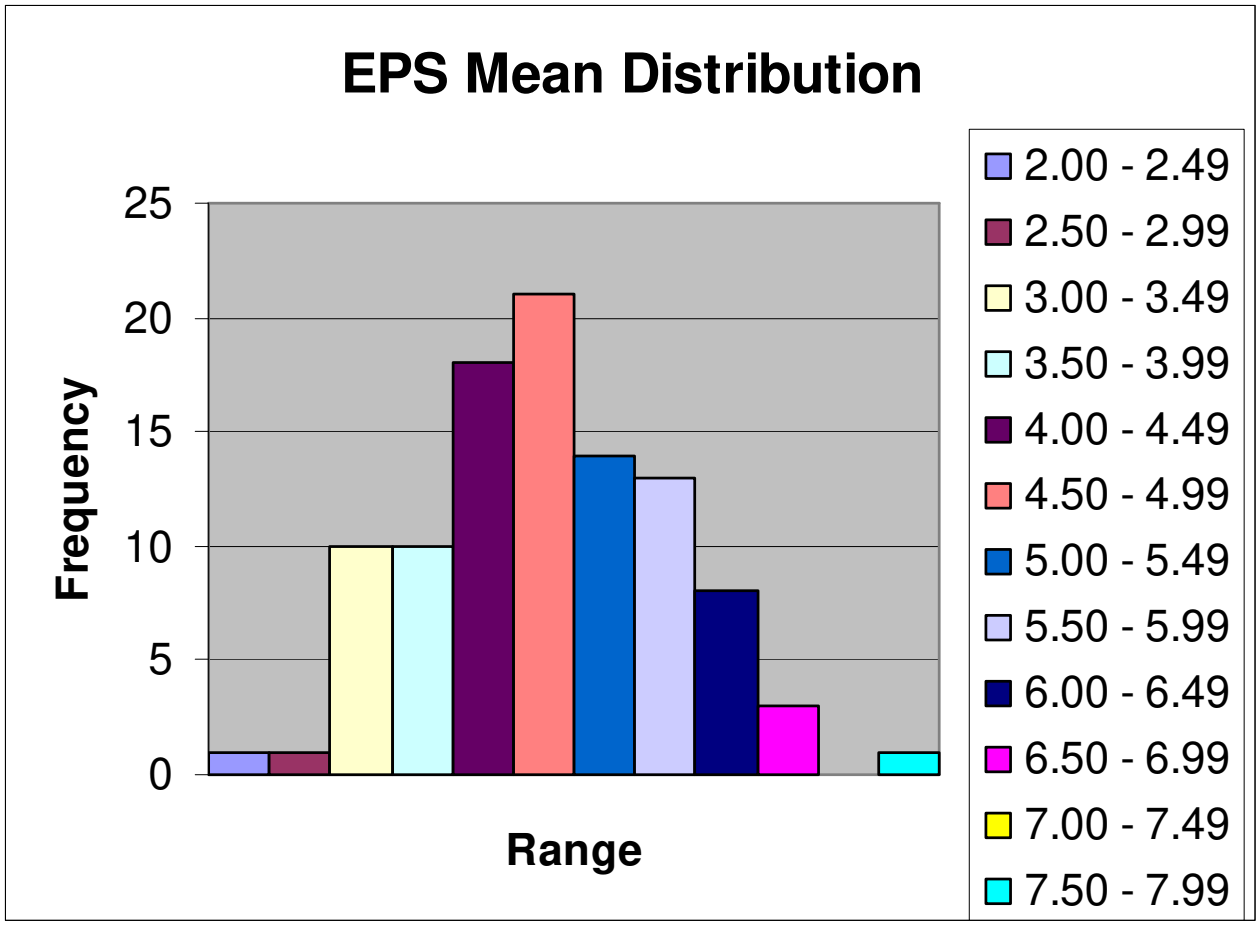
Mercury makes a 2.4 liter V-6 engine, the Laser XRi, used in speedboats. The company's engineers believe the engine delivers an average power of 220 horsepower and that the standard deviation of power delivered is 15 HP. A potential buyer intends to sample 100 engines (each engine is to be run a single time). What is the probability that the sample mean will be less than 217HP?

$$P(\bar{X} < 217) = P\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{217 - \mu}{\frac{\sigma}{\sqrt{n}}}\right)$$

$$= P\left(Z < \frac{217 - 220}{\frac{15}{\sqrt{100}}}\right) = P\left(Z < \frac{217 - 220}{\frac{15}{10}}\right)$$

$$= P(Z < -2) = 0.0228$$

Example 5-2

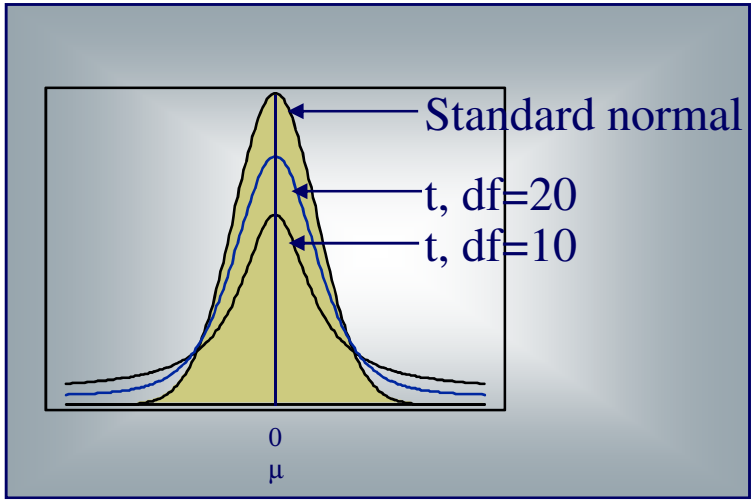

EPS Mean Distribution
Frequency
25
20
15
10
5
0
Range
2.00 - 2.49
2.50 - 2.99
3.00 - 3.49
3.50 - 3.99
4.00 - 4.49
4.50 - 4.99
5.00 - 5.49
5.50 - 5.99
6.00 - 6.49
6.50 - 6.99
7.00 - 7.49
7.50 - 7.99

Student's *t* Distribution

If the population standard deviation, σ, is ***unknown***, replace σ with the sample standard deviation, *s*. If the population is normal, the resulting statistic: $t = \dfrac{\overline{X} - \mu}{s / \sqrt{n}}$

has a ***t distribution* with *(n - 1) degrees of freedom***.

- The *t* is a family of bell-shaped and symmetric distributions, one for each number of degree of freedom.
- The expected value of *t* is 0.
- The variance of *t* is greater than 1, but approaches 1 as the number of degrees of freedom increases. The *t* is flatter and has fatter tails than does the standard normal.
- The *t* distribution approaches a standard normal as the number of degrees of freedom increases.

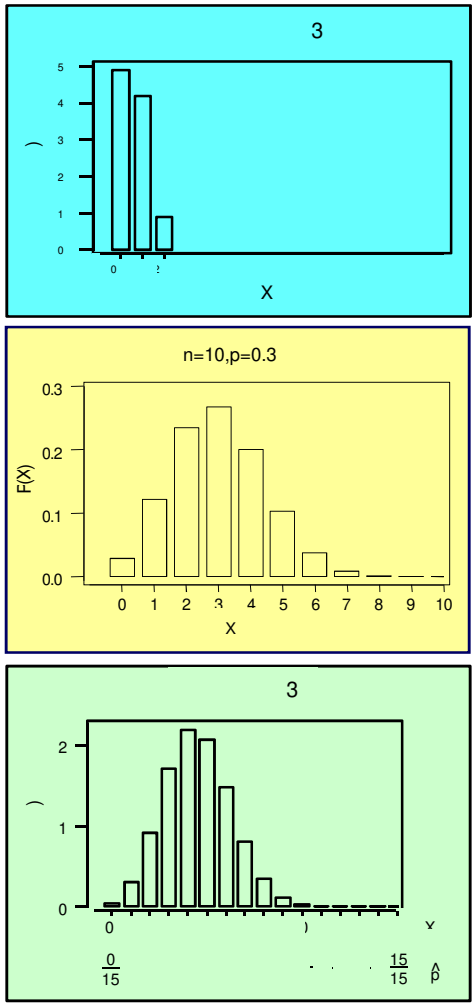


The Sampling Distribution of the Sample Proportion, $\hat{p}$

The **sample proportion** is the percentage of successes in n binomial trials. It is the number of successes, X, divided by the number of trials, n.

Sample proportion: $\hat{p} = \dfrac{X}{n}$

As the sample size, n, increases, the sampling distribution of $\hat{p}$ approaches a **normal distribution** with mean p and standard deviation $\sqrt{\dfrac{p(1-p)}{n}}$



Sample Proportion (Example 5-3)

In recent years, convertible sports coupes have become very popular in Japan. Toyota is currently shipping Celicas to Los Angeles, where a customizer does a roof lift and ships them back to Japan. Suppose that 25% of all Japanese in a given income and lifestyle category are interested in buying Celica convertibles. A random sample of 100 Japanese consumers in the category of interest is to be selected. What is the probability that at least 20% of those in the sample will express an interest in a Celica convertible?

$$n = 100$$

$$p = 0.25$$

$$np = (100)(0.25) = 25 = E(\hat{p})$$

$$\frac{p(1-p)}{n} = \frac{(.25)(.75)}{100} = 0.001875 = V(\hat{p})$$

$$\sqrt{\frac{p(1-p)}{n}} = \sqrt{0.001875} = 0.04330127 = SD(\hat{p})$$

$$P(\hat{p} > 0.20) = P\left(\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} > \frac{.20 - p}{\sqrt{\frac{p(1-p)}{n}}}\right)$$

$$= P\left(z > \frac{.20-.25}{\sqrt{\frac{(.25)(.75)}{100}}}\right) = P\left(z > \frac{-.05}{.0433}\right)$$

$$= P(z > -1.15) = 0.8749$$

5-4 Estimators and Their Properties

An **estimator** of a population parameter is a sample statistic used to estimate the parameter. The most commonly-used estimator of the:

Population Parameter		Sample Statistic
Mean (μ)	is the	Mean ($\bar{X}$)
Variance (σ^2)	is the	Variance (s^2)
Standard Deviation (σ)	is the	Standard Deviation (s)
Proportion (p)	is the	Proportion ($\hat{p}$)

- Desirable properties of estimators include:
  - ✓Unbiasedness
  - ✓Efficiency
  - ✓Consistency
  - ✓Sufficiency

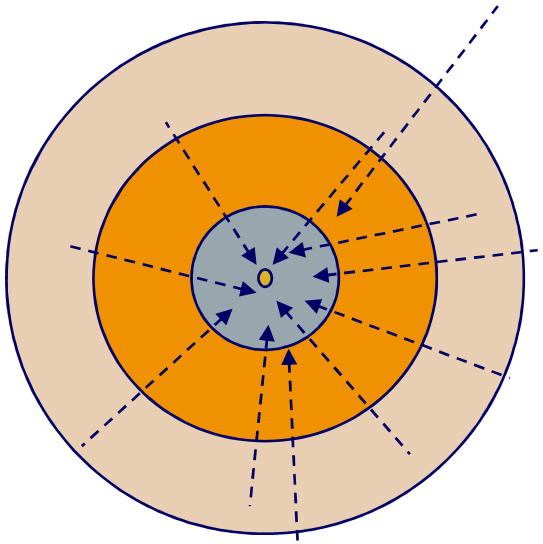
Unbiasedness

An estimator is said to be **unbiased** if its expected value is equal to the population parameter it estimates.

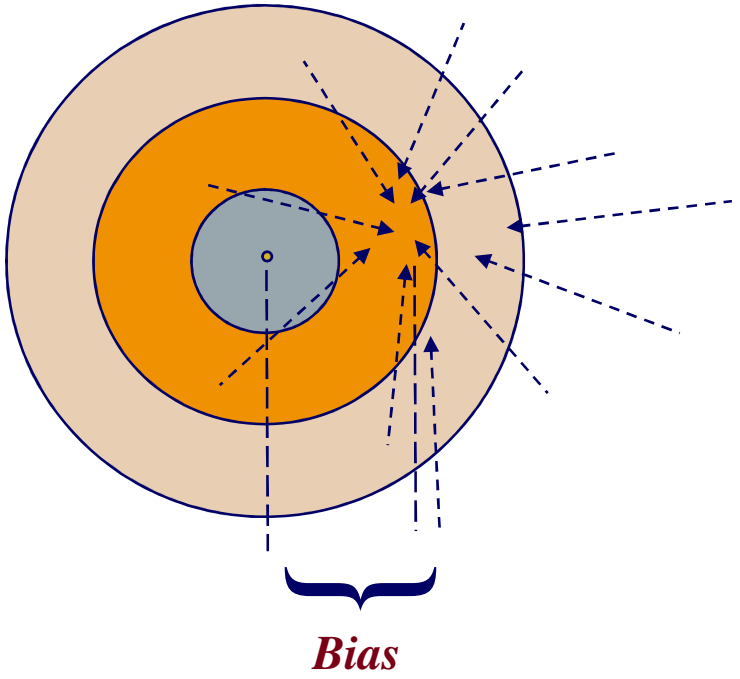
For example, $E(\bar{X})=\mu$, so the sample mean is an unbiased estimator of the population mean. Unbiasedness is an average or long-run property. The mean of any single sample will probably not equal the population mean, but the average of the means of repeated independent samples from a population will equal the population mean.

Any ***systematic deviation*** of the estimator from the population parameter of interest is called a **bias**.

Unbiased and Biased Estimators



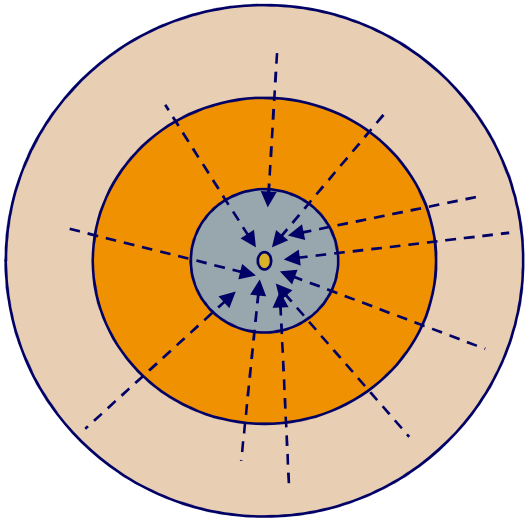
An **unbiased** estimator is on target on average.



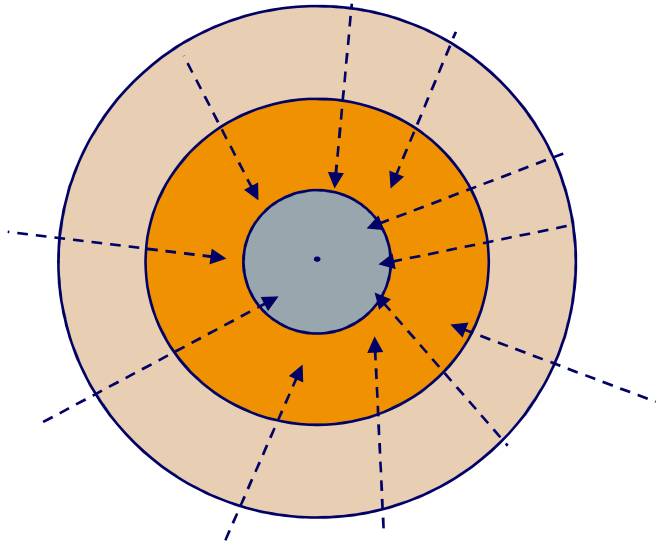
A **biased** estimator is off target on average.

Efficiency

An estimator is **efficient** if it has a relatively small variance (and standard deviation).



An **efficient** estimator is, on average, closer to the parameter being estimated..

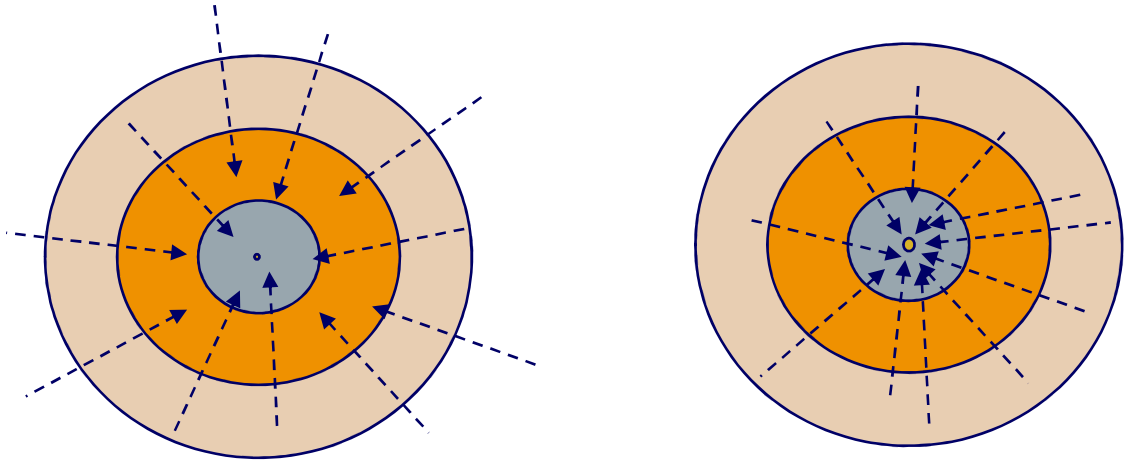


An **inefficient** estimator is, on average, farther from the parameter being estimated.

Consistency and Sufficiency

An estimator is said to be **consistent** if its probability of being close to the parameter it estimates increases as the sample size increases.

Consistency



n = 10

n = 100

An estimator is said to be **sufficient** if it contains all the information in the data about the parameter it estimates.

Properties of the Sample Mean

For a normal population, both the sample mean and sample median are ***unbiased estimators*** of the population mean, but the sample mean is both more ***efficient*** (because it has a smaller variance), and ***sufficient***. Every observation in the sample is used in the calculation of the sample mean, but only the middle value is used to find the sample median.

In general, the sample mean is the ***best*** estimator of the population mean. The sample mean is the most efficient unbiased estimator of the population mean. It is also a consistent estimator.

Properties of the Sample Variance

The ***sample variance*** (the sum of the squared deviations from the sample mean divided by *(n-1)* is an ***unbiased estimator*** of the population variance. In contrast, the ***average squared deviation*** from the sample mean is a ***biased*** (though *consistent*) estimator of the population variance.

$$E(s^2) = E\left(\frac{\sum (x-\bar{x})^2}{(n-1)}\right) = \sigma^2$$

$$E\left(\frac{\sum (x-\bar{x})^2}{n}\right) < \sigma^2$$

5-5 Degrees of Freedom

Consider a sample of size n=4 containing the following data points:

$x_1=10$ $x_2=12$ $x_3=16$ $x_4=?$

and for which the sample mean is: $\bar{x} = \frac{\sum x}{n} = 14$

Given the values of three data points and the sample mean, the value of the fourth data point can be determined:

$$\bar{x} = \frac{\sum x}{n} = \frac{12+14+16+x_4}{4} = 14$$

$$12+14+16+x_4 = 56$$

$$x_4 = 56 - 12 - 14 - 16$$

$$x_4 = 14$$

Degrees of Freedom (Continued)

If only two data points and the sample mean are known:

x_1=10 x_2=12 x_3=? x_4=? $\bar{x} = 14$

The values of the remaining two data points cannot be uniquely determined:

$$\bar{x} = \frac{\sum x}{n} = \frac{12 + 14 + x_3 + x_4}{4} = 14$$

$$12 + 14 + x_3 + x_4 = 56$$

Degrees of Freedom (Continued)

The number of ***degrees of freedom*** is equal to the total number of measurements (these are not always raw data points), less the total number of *restrictions* on the measurements. A restriction is a quantity computed from the measurements.

The sample mean is a restriction on the sample measurements, so after calculating the sample mean there are only ***(n-1)*** **degrees of freedom** remaining with which to calculate the sample variance. The sample variance is based on only *(n-1)* free data points:

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

Example 5-4

A sample of size 10 is given below. We are to choose three different numbers from which the deviations are to be taken. The first number is to be used for the first five sample points; the second number is to be used for the next three sample points; and the third number is to be used for the last two sample points.

Sample #	1	2	3	4	5	6	7	8	9	10
Sample Point	93	97	60	72	96	83	59	66	88	53

i. What three numbers should we choose in order to minimize the SSD (sum of squared deviations from the mean).?

- Note: $SSD = \sum (x - \bar{x})^2$

Example 5-4 (continued)

Solution: Choose the means of the corresponding sample points. These are: 83.6, 69.33, and 70.5.

ii. Calculate the SSD with chosen numbers.

Solution: SSD = 2030.367. See table on next slide for calculations.

iii. What is the *df* for the calculated SSD?

Solution: $df = 10 - 3 = 7$.

iv. Calculate an unbiased estimate of the population variance.

Solution: An unbiased estimate of the population variance is SSD/df = 2030.367/7 = 290.05.

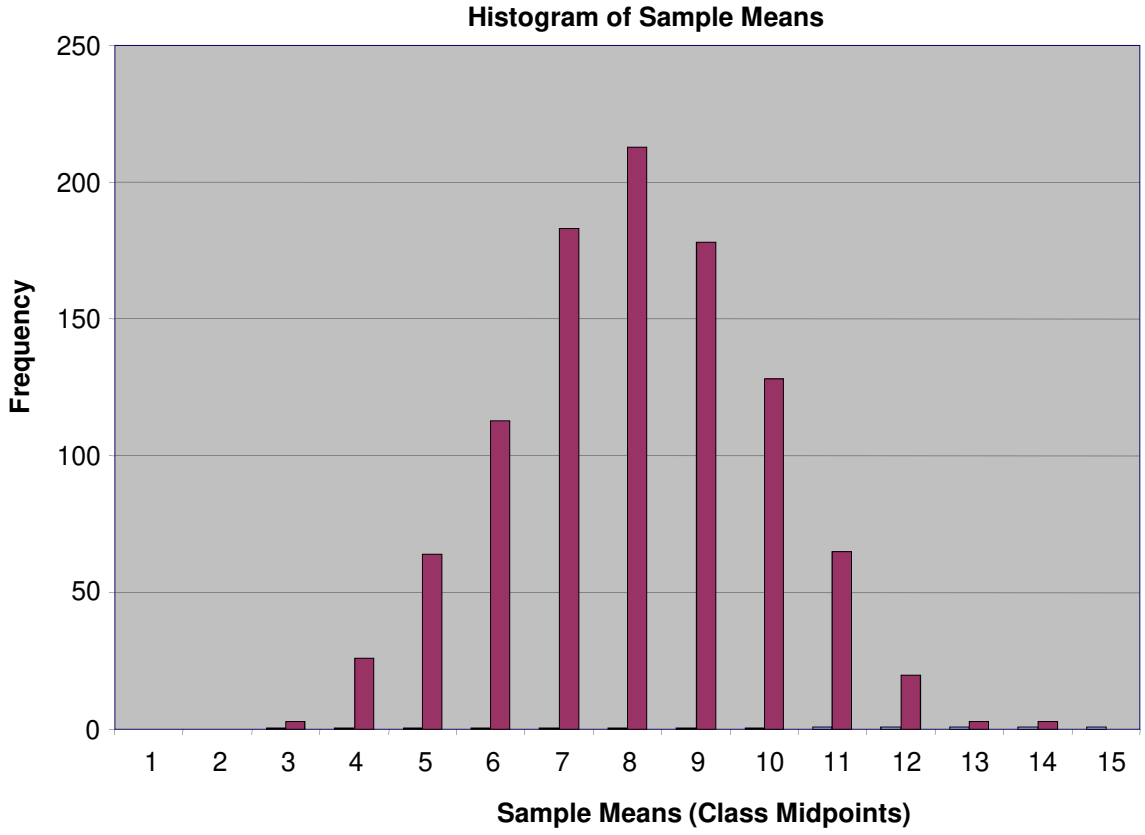
Example 5-4 (continued)

Sample #	Sample Point	Mean	Deviations	Deviation Squared
1	93	83.6	9.4	88.36
2	97	83.6	13.4	179.56
3	60	83.6	-23.6	556.96
4	72	83.6	-11.6	134.56
5	96	83.6	12.4	153.76
6	83	69.33	13.6667	186.7778
7	59	69.33	-10.3333	106.7778
8	66	69.33	-3.3333	11.1111
9	88	70.5	17.5	306.25
10	53	70.5	-17.5	306.25
			SSD	2030.367
			SSD/df	290.0524

Using Excel to Generate Random Data

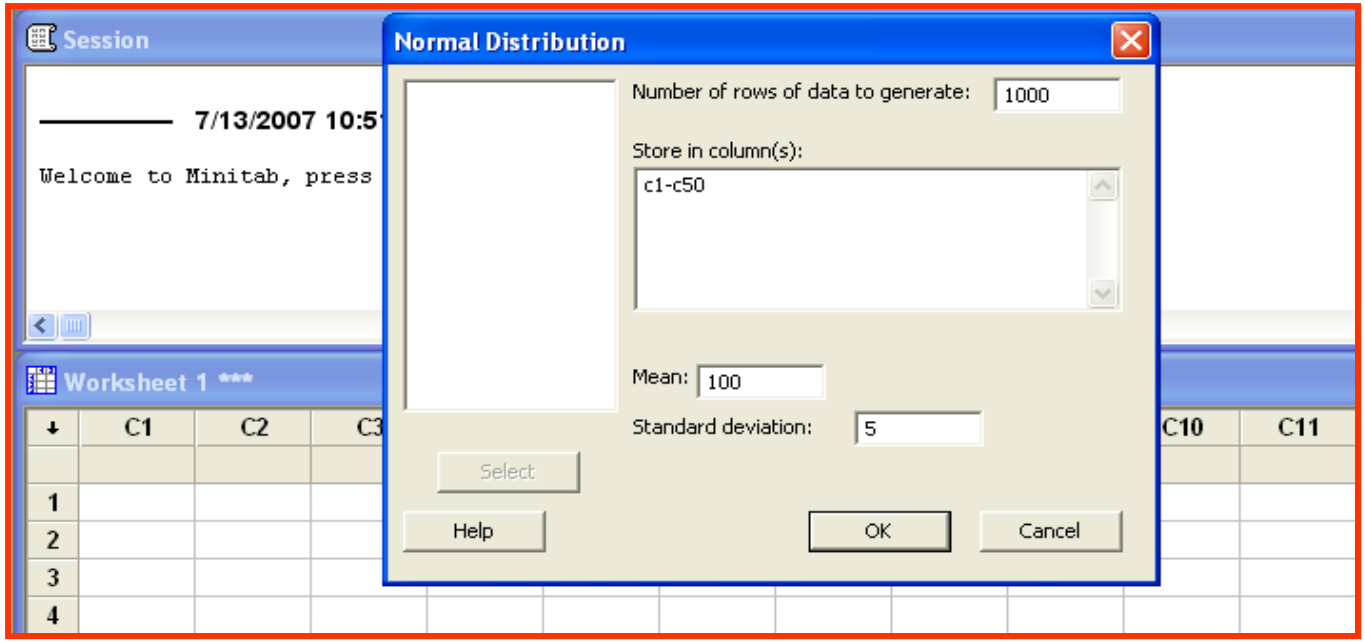
Constructing a sampling distribution of the mean from a uniform population (n = 10) using EXCEL (use RANDBETWEEN(0, 1) command to generate values to graph):

CLASS MIDPOINT	FREQUENCY
0.15	0
0.2	0
0.25	3
0.3	26
0.35	64
0.4	113
0.45	183
0.5	213
0.55	178
0.6	128
0.65	65
0.7	20
0.75	3
0.8	3
0.85	0
	999



Using Minitab to Generate Random Data

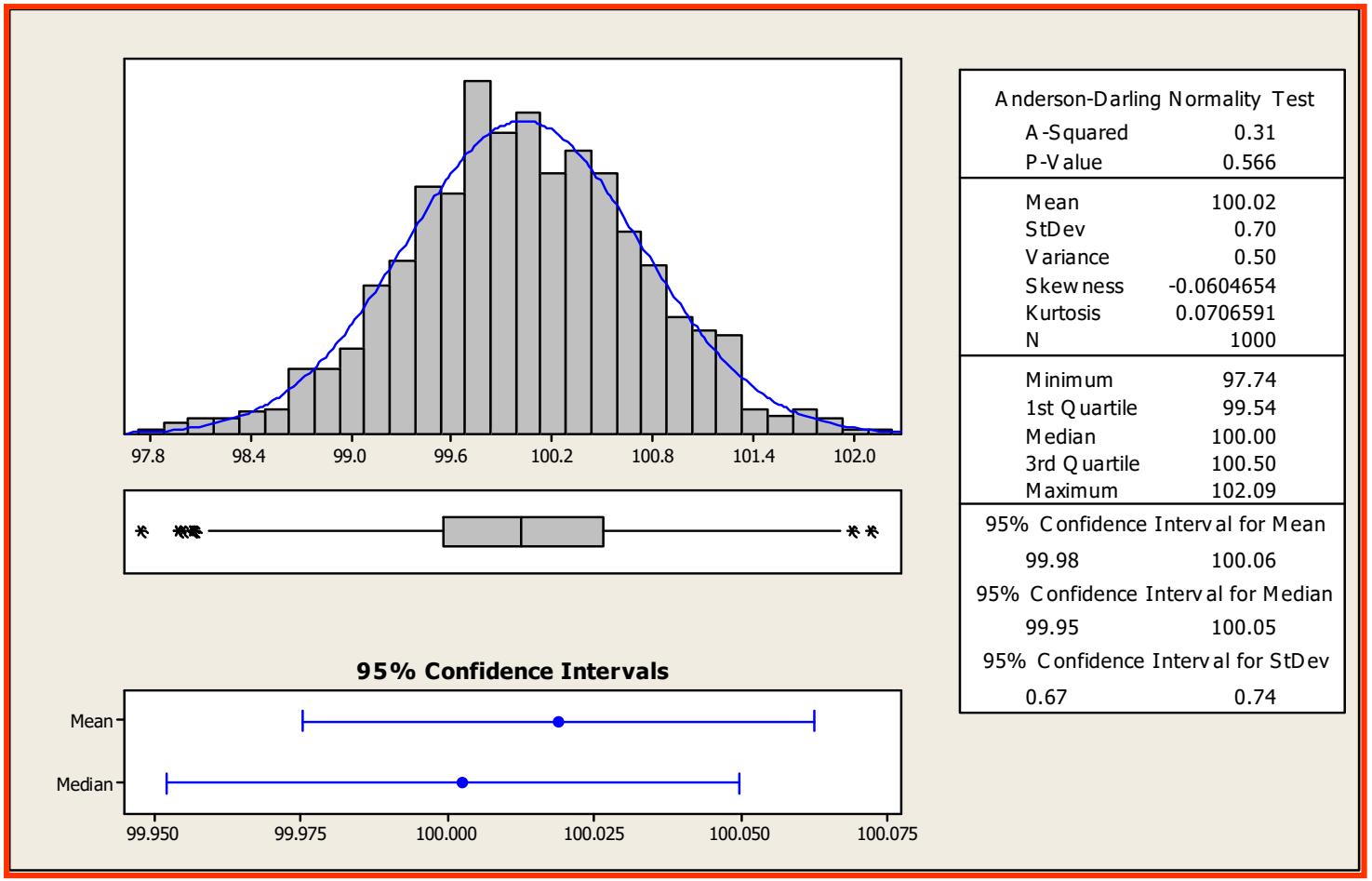
Constructing a sampling distribution of the mean from any distribution using MINITAB can be achieved by selecting CALC→RANDOM DATA and then generating the data from a selected distribution. For example, we can generate data from a normal distribution with mean = 100 and standard deviation = 5.



Using Minitab to Look at the Sampling Distribution of the sample Mean (n = 50)

- If we would like to look at the distribution of the sample means for the previous simulation, we can select different sample sizes from the columns (C1 to C50). If for instance we select samples of size n = 50, then we will have 1000 of these based on the previous simulation.
- In this simulation $\mu = 100$ and $\sigma = 5$.
- We can compute the row means for these 50 columns and save in the next available columns.
- Thus for the sampling distribution of the sample means, its mean will be 100 and standard deviation will be $5/\sqrt{50} = 0.7071$.
- The next slide shows this situation. Observe that the distribution for these simulated sample means is approximately normally distributed with a mean of 100.02 and a standard deviation of 0.70. These values are very close to the theoretical values.

Using Minitab to Look at the Sampling Distribution of the sample Mean (n = 50)



Anderson-Darling Normality Test	
A-Squared	0.31
P-Value	0.566
Mean	100.02
StDev	0.70
Variance	0.50
Skewness	-0.0604654
Kurtosis	0.0706591
N	1000
Minimum	97.74
1st Quartile	99.54
Median	100.00
3rd Quartile	100.50
Maximum	102.09
95% Confidence Interval for Mean	
99.98	100.06
95% Confidence Interval for Median	
99.95	100.05
95% Confidence Interval for StDev	
0.67	0.74