05. Sampling Distribution

Adapted From :

Probability & Statistics for Engineers & Scientists, 9th Ed.

Walpole/Myers/Myers/Ye (c)2010

Introduction to Business Statistics, 5e

Kvanli/Guynes/Pavur (c)2000

South-Western College Publishing

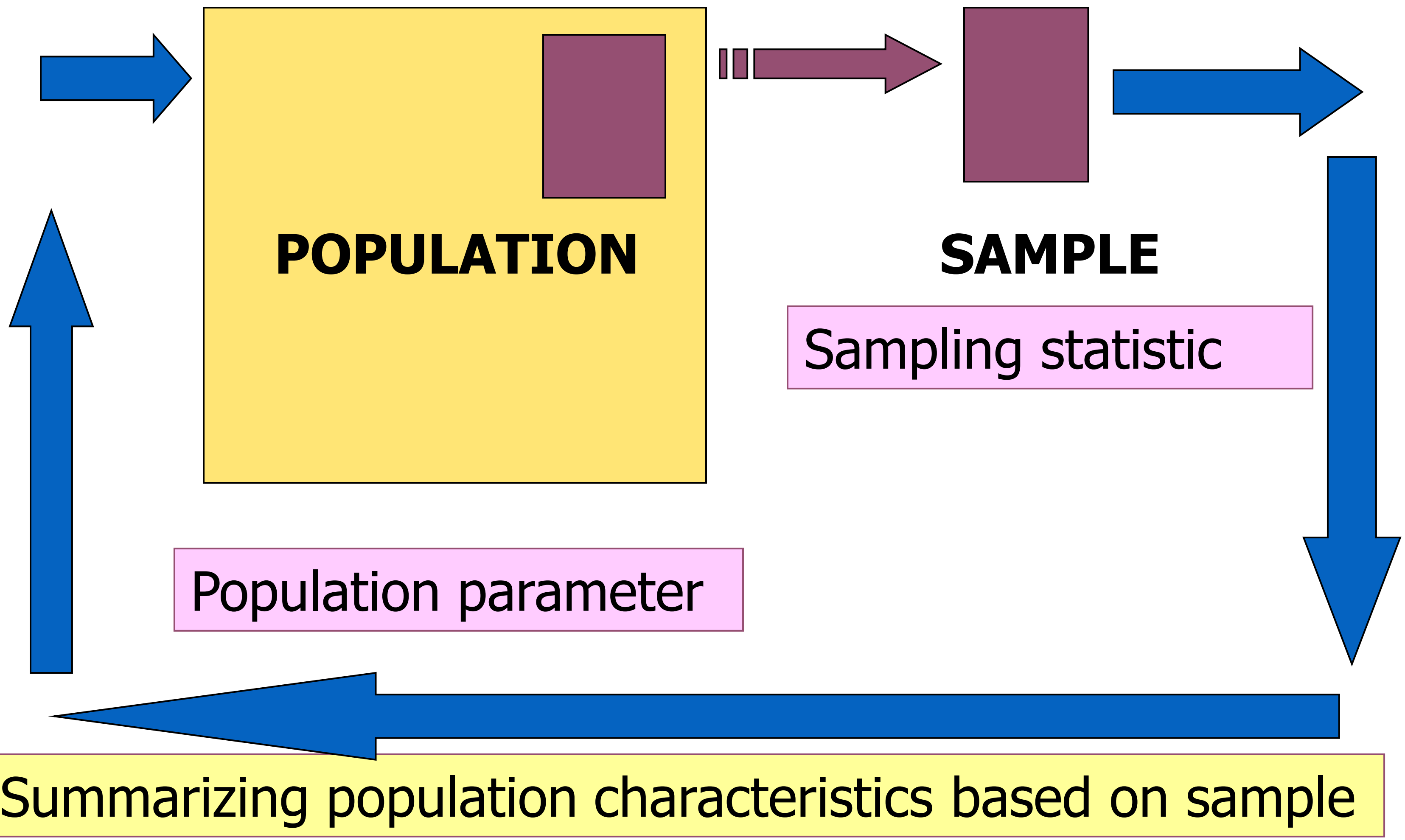
Statistics for Managers

Using Microsoft® Excel 4th Edition

Populations & Samples

- A **population** is the set (possibly infinite) of all possible observations.
  - Each observation is a random variable X having some (often unknown) probability distribution, f (x).
  - If the population distribution is known, it might be referred to, for example, as a normal population, etc.
- A **sample** is a subset of a population.
  - Our goal is to make inferences about the population based on an analysis of the sample.
  - A biased sample, usually obtained by taking convenient, rather than representative observations, will consistently over- or under-estimate some characteristic of the population.
  - Observations in a random sample are made independently and at random. Here, random variables X_1, X_2, …, X_n in the sample all have same distribution as the population, X.

Inferential Statistics



Parameter and Statistic

- *"The height average of Indonesian people"*

  Parameter

Population

- *"The height average of Jakarta citizen"*

  Statistic

Sampel

Sampling Distributions

- A sampling distribution is a distribution of all of the possible values of a statistic for a given size sample selected from a population

Sample Statistics

- Any function of the random variables $X_1, X_2, \ldots, X_n$ making up a random sample is called a statistic.
- The most important statistics, as we have seen are the sample mean, sample variance and sample standard deviation:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1} = \frac{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2}{n(n-1)}$$

Sampling Distributions

- Starting with an unknown population distribution, we can study the sampling distribution, or distribution of a sample statistic (like X_{bar} or S) calculated from a sample of size n from that population.
  - The sample consists of independent and identically distributed observations $X_1, X_2, \ldots, X_n$ from the population.
  - Based on the sampling distributions of X_{bar} and S for samples of size n, we will make inferences about the population mean and variance μ and σ.
  - We could approximate the sampling distribution of X_{bar} by taking a large number of random samples of size n and plotting the distribution of the X_{bar} values.

Developing a Sampling Distribution

- **Assume there is a population …**
- Population size N=4
- Random variable, X, is age of individuals
- Values of X: 18, 20, 22, 24 (years)

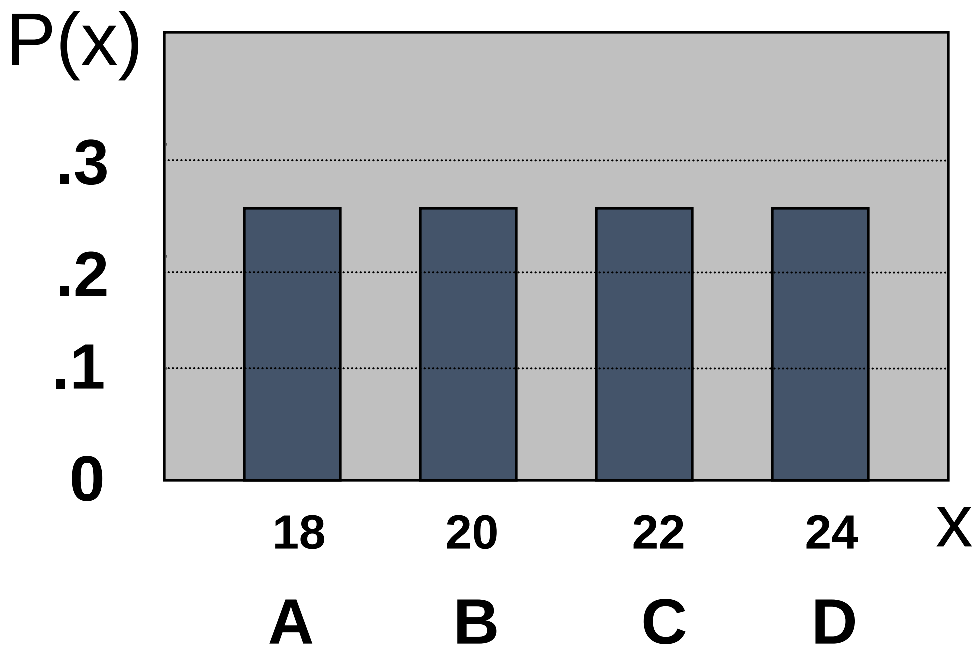
Developing a Sampling Distribution

(continued)

Summary Measures for the Population Distribution:

$$\mu = \frac{\sum X_i}{N} = \frac{18 + 20 + 22 + 24}{4} = 21$$

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}} = 2.236$$



Uniform Distribution

Developing a Sampling Distribution

(continued)

Now consider all possible samples of size n=2

1st Obs	2nd Observation			
	18	20	22	24
18	18,18	18,20	18,22	18,24
20	20,18	20,20	20,22	20,24
22	22,18	22,20	22,22	22,24
24	24,18	24,20	24,22	24,24

16 possible samples (sampling with replacement)

16 Sample Means

1st Obs	2nd Observation			
	18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

Developing a Sampling Distribution

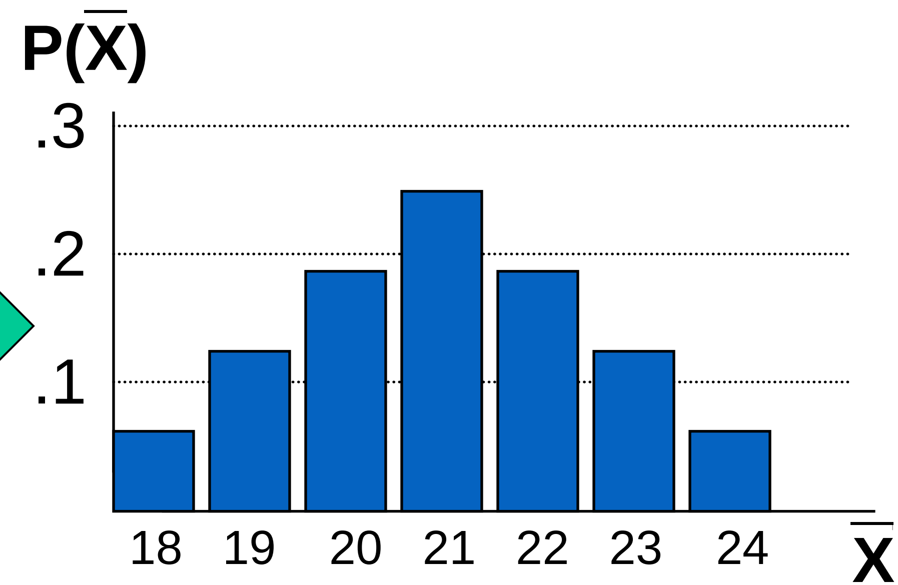
(continued)

Sampling Distribution of All Sample Means

16 Sample Means

1st Obs	2nd Observation			
	18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

Sample Means Distribution



(no longer uniform)

Developing a Sampling Distribution

(continued)

Summary Measures of this Sampling Distribution:

$$\mu_{\bar{X}} = \frac{\sum \bar{X}_i}{N} = \frac{18 + 19 + 21 + \cdots + 24}{16} = 21$$

$$\sigma_{\bar{X}} = \sqrt{\frac{\sum (\bar{X}_i - \mu_{\bar{X}})^2}{N}}$$

$$= \sqrt{\frac{(18 - 21)^2 + (19 - 21)^2 + \cdots + (24 - 21)^2}{16}} = 1.58$$

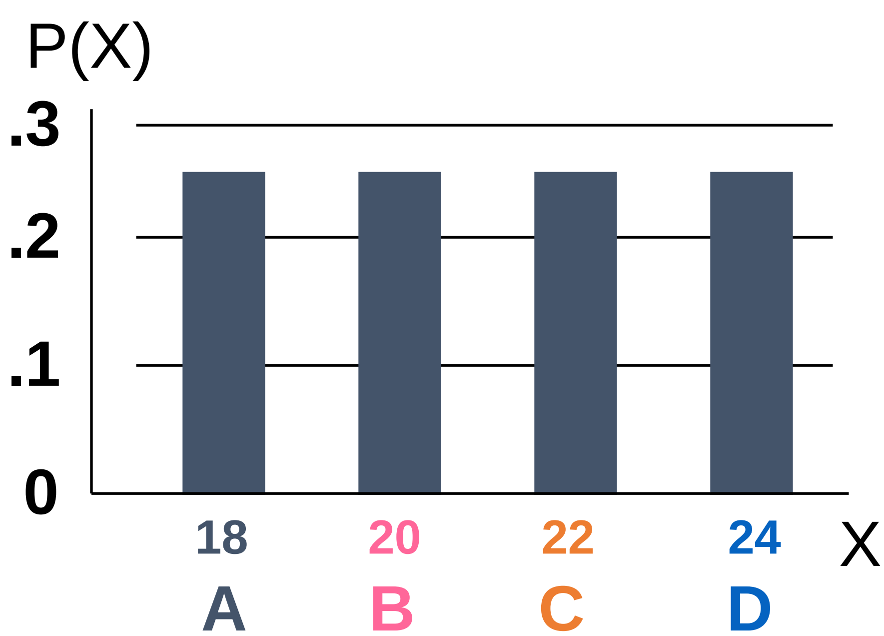
Comparing the Population with its Sampling Distribution

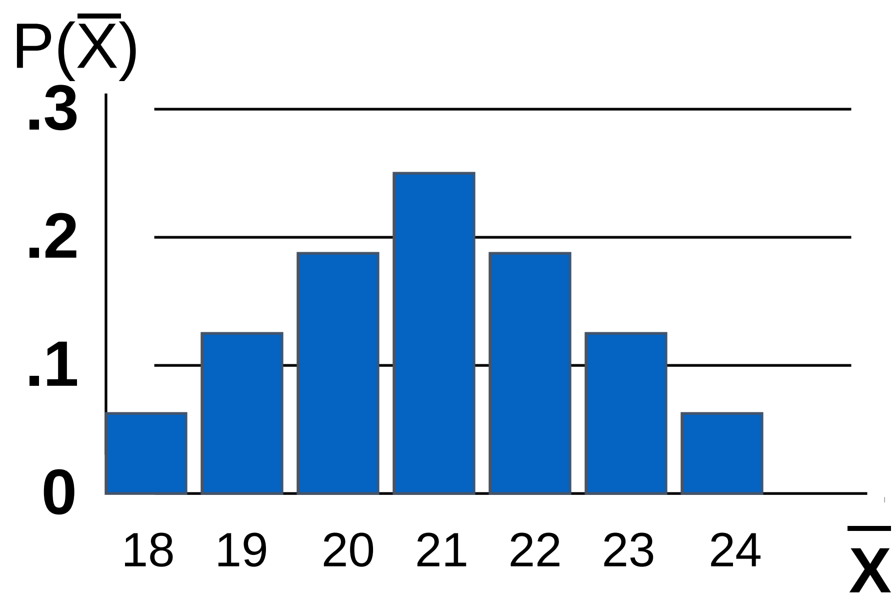
Population
N = 4

$\mu = 21 \qquad \sigma = 2.236$

Sample Means Distribution
n = 2

$\mu_{\bar{X}} = 21 \qquad \sigma_{\bar{X}} = 1.58$





Sampling Distribution Summary

- Normal distribution: Sampling distribution of X_{bar} when σ is known for any population distribution.
  - Also the sampling distribution for the difference of the means of two different samples.
- t-distribution: Sampling distribution of X_{bar} when σ is unknown and S is used. Population must be normal.
  - Also the sampling distribution for the difference of the means of two different samples when σ is unknown.
- Chi-square (χ^2) distribution: Sampling distribution of S^2. Population must be normal.
- F-distribution: The distribution of the ratio of two χ^2 random variables. Sampling distribution of the ratio of the variances of two different samples. Population must be normal.

Standard Error of the Mean

- Different samples of the same size from the same population will yield different sample means
- A measure of the variability from sample to sample is given by the Standard Error of the Mean:

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

- Note that the standard error of the mean decreases as the sample size increases

If the Population is Normal

- If a population is normal with mean μ and standard deviation σ, the sampling distribution of $\bar{X}$ is also normally distributed with

$$\boxed{\mu_{\bar{X}} = \mu}$$

and

$$\boxed{\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}}$$

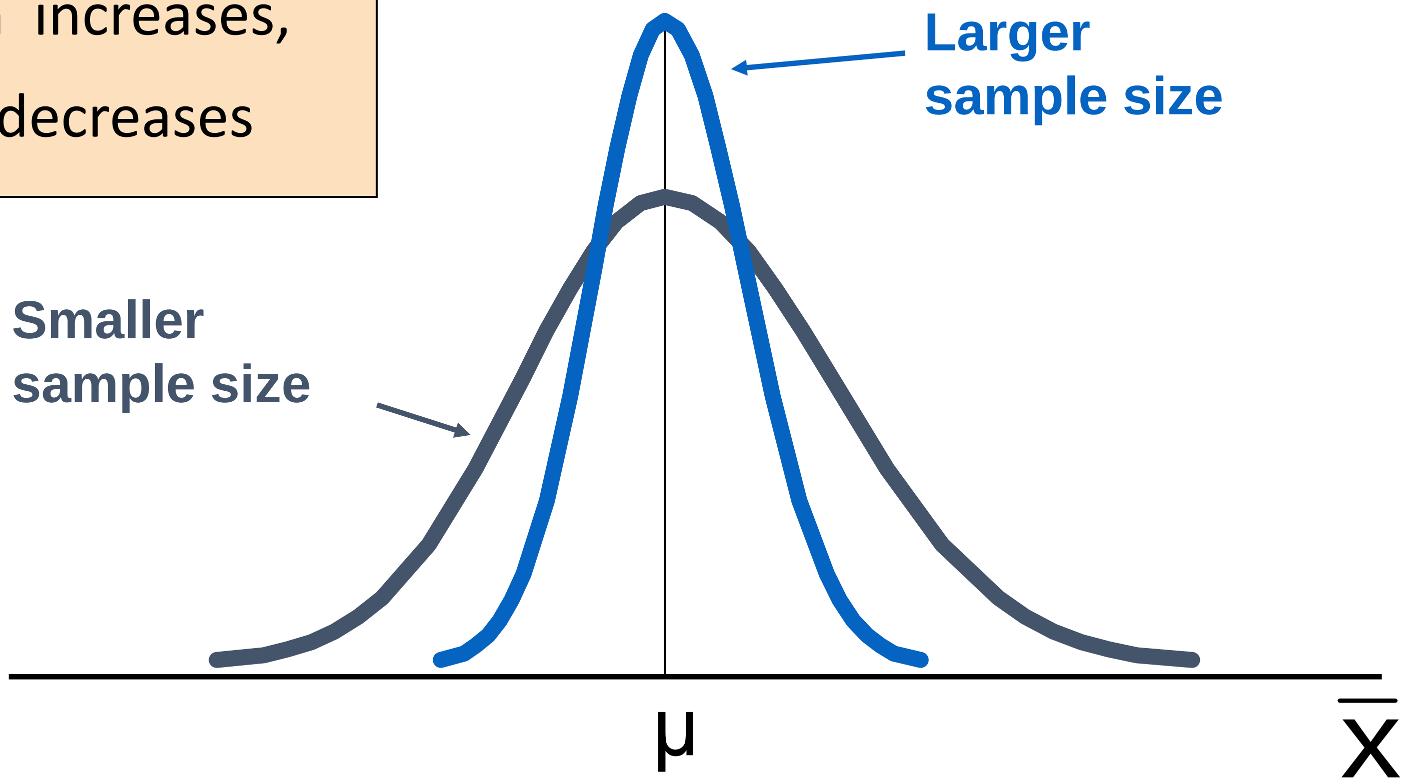
(This assumes that sampling is with replacement or sampling is without replacement from an infinite population)

Sampling Distribution Properties

- For sampling with replacement:

As n increases,

$\sigma_{\bar{x}}$ decreases



Central Limit Theorem

- The central limit theorem is the most important theorem in statistics. It states that
- If X_{bar} is the mean of a random sample of size n from a population with an arbitrary distribution with mean μ and variance σ^2, then as $n \to \infty$, the sampling distribution of X_{bar} approaches a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.
- The central limit theorem holds under the following conditions:
  - For any population distribution if $n \geq 30$.
  - For $n < 30$, if the population distribution is generally shaped like a normal distribution.
  - For any value of n if the population distribution is normal.

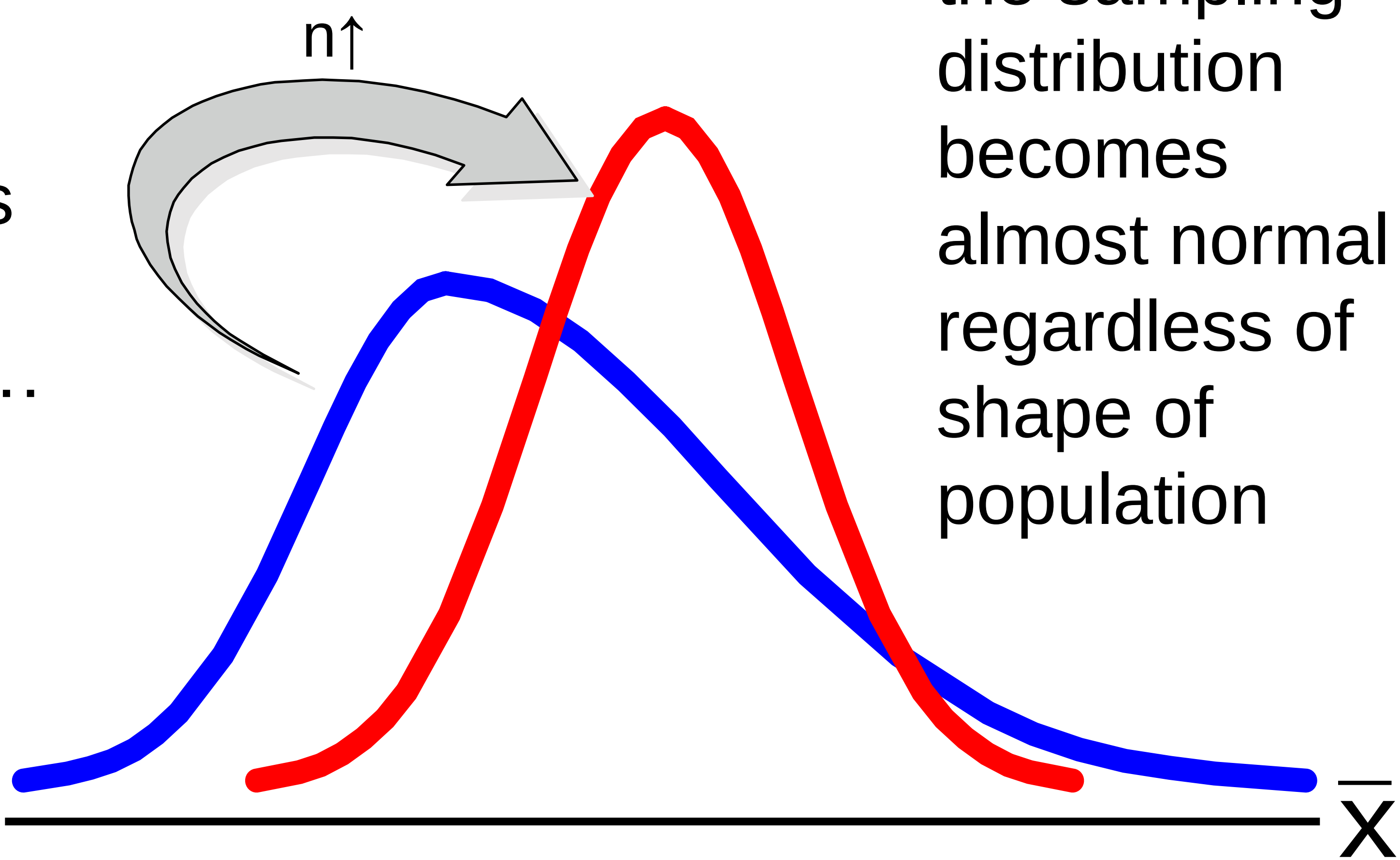
If the Population is not Normal

- We can apply the Central Limit Theorem:
  - Even if the population is not normal,
  - ...sample means from the population will be approximately normal as long as the sample size is large enough
  - ...and the sampling distribution will have

$$\boxed{\mu_{\bar{x}} = \mu}$$ and $$\boxed{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}}$$

Central Limit Theorem

As the sample size gets large enough…



the sampling distribution becomes almost normal regardless of shape of population

If the Population is not Normal *(continued)*

Sampling distribution properties:

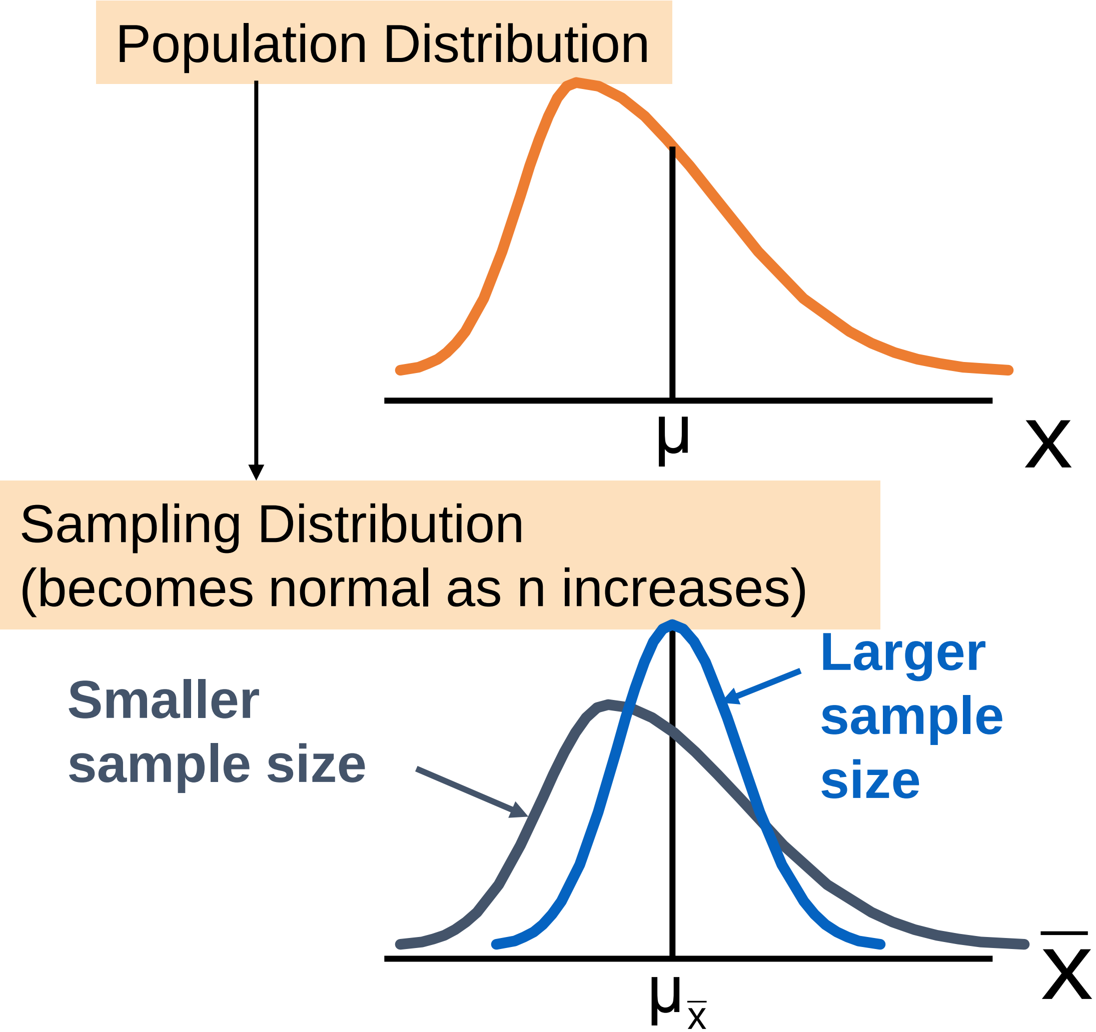
Central Tendency

$$\mu_{\bar{x}} = \mu$$

Variation

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

(Sampling with replacement)



How Large is Large Enough?

- For most distributions, $n > 30$ will give a sampling distribution that is nearly normal
- For fairly symmetric distributions, $n > 15$
- For normal population distributions, the sampling distribution of the mean is always normally distributed

Example

- Suppose a population has mean $\mu = 8$ and standard deviation $\sigma = 2$. Suppose a random sample of size $n = 25$ is selected.

- What is the probability that the sample mean is between 7.8 and 8.2?

Example:

$$\mu = 8 \qquad \sigma = 2 \qquad n = 25$$

$$P\left(7.8 < \bar{X} < 8.2\right) = ?$$

$$P\left(7.8 < \bar{X} < 8.2\right) = P\left(\frac{7.8-8}{2/\sqrt{25}} < \frac{\bar{X}-\mu_{\bar{X}}}{\sigma_{\bar{X}}} < \frac{8.2-8}{2/\sqrt{25}}\right)$$

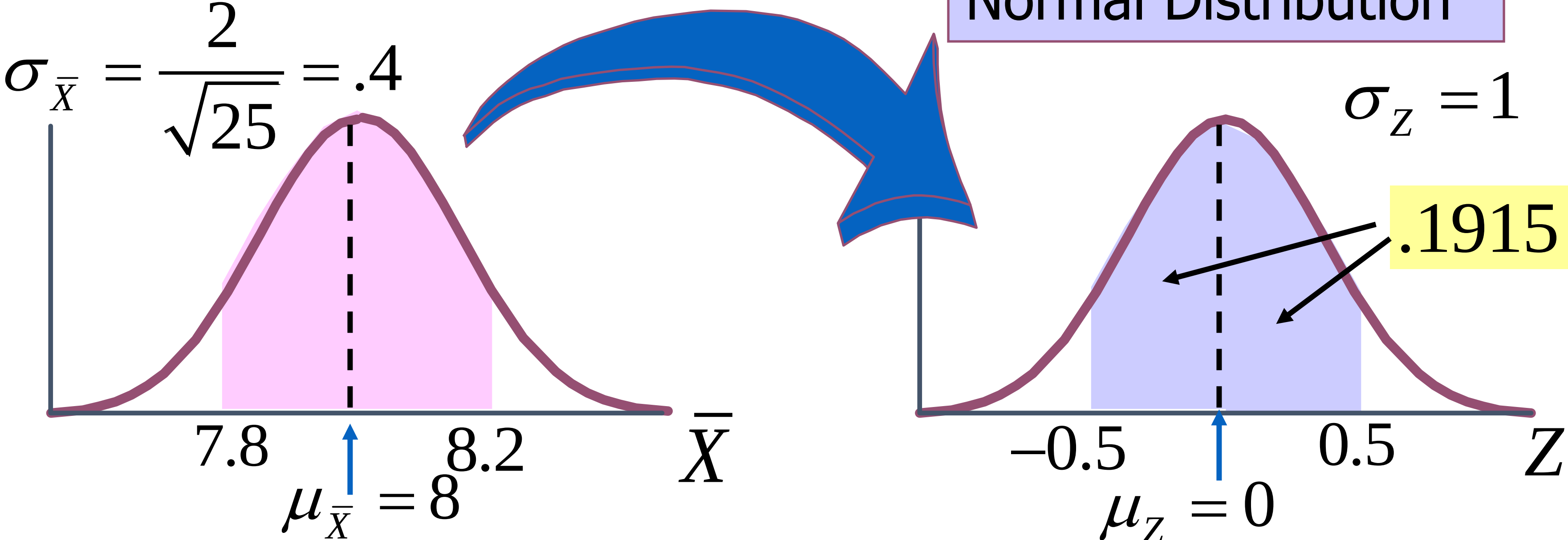
$$= P\left(-.5 < Z < .5\right) = .3830$$

Sampling Distribution

Standardized Normal Distribution

$$\sigma_{\bar{X}} = \frac{2}{\sqrt{25}} = .4$$

$$\sigma_Z = 1$$



Inferences About the Population Mean

- We often want to test hypotheses about the population mean (hypothesis testing will be formalized later).
- Example:
  - Suppose a manufacturing process is designed to produce parts with $\mu = 6$ cm in diameter, and suppose σ is known to be .15 cm. If a random sample of 80 parts has $x_{bar} = 6.046$ cm, what is the probability (P-value) that a value this far from the mean could occur by chance if μ is truly 6 cm?

$$z = \frac{6.046 - 6.00}{.15/\sqrt{80}} = 2.74$$

$$P[|\overline{X} - 6.0| \geq .046] = P[|Z| \geq 2.74] = ?$$

$$P[|Z| \geq 2.74] = 2P[Z \geq 2.74] = 2(1 - .9969) = .0062$$

Difference Between Two Means

- In addition, we can make inferences about the difference between two population means based on the difference between two sample means.
- The central limit theorem also holds in this case.
- If independent samples of size n_1 and n_2 are drawn at random from two populations, discrete or continuous, with means μ_1 and μ_2, and variances σ_1^2 and σ_2^2, then as n gets large, the sampling distribution of X_{bar1} - X_{bar2} approaches a normal distribution with

$$\mu_{\bar{X}_1-\bar{X}_2} = \mu_1 - \mu_2 \quad \text{and} \quad \sigma^2_{\bar{X}_1-\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Difference of Two Means Example

- Example:
  - Suppose we record the drying time in hours of 20 samples each of two types of paint, type A and type B. Suppose we know that the population standard deviations are both equal to 1/2 hour. <u>Assuming that the population means are equal</u>, what is the probability that the difference in the sample means is greater than 1/2 hour?

$$\mu_{\bar{X}A-\bar{X}B} = \mu_A - \mu_B = 0$$

$$\sigma^2_{\bar{X}_A-\bar{X}_B} = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B} = \frac{.25}{20} + \frac{.25}{20} = .025$$

$$z = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{\sqrt{(\sigma_A^2 / n_A) + (\sigma_B^2 / n_B)}} = \frac{.5 - 0}{\sqrt{.025}} = 3.16$$

- P(z > 3.16) = .0008. Very unlikely to happen by chance.

t-Distribution (when σ is Unknown)

- The problem with the central limit theorem is that it assumes that σ is known.
  - Generally, if μ is being estimated from the sample, σ must be estimated from the sample as well.
  - The t-distribution can be used if σ is unknown, but it requires that the original population must be normally distributed.
- Let $X_1, X_2, \ldots, X_n$ be independent, normally distributed random variables with mean μ and standard deviation σ. Then the random variable T below has a t-distribution with $\nu = n - 1$ degrees of freedom:

$$T = \frac{\overline{X} - \mu}{S / \sqrt{n}}$$

  - The t-distribution is like the normal, but with greater spread since both μ and σ have fluctuations due to sampling.

Using the t-Distribution

- Observations on the t-Distribution:
  - The t-statistic is like the normal, but using S rather than σ.
  - Table value depends on the sample size (degrees of freedom).
  - The t-distribution is symmetric with $\mu = 0$, but $\sigma^2 > 1$. As would be expected, σ^2 is largest for small n.
  - Approaches the normal distribution ($\sigma^2 = 1$) as n gets large.
  - For a given probability α, table shows the value of t_α that has $P(\alpha)$ to the right of it.
- The t-distribution can also be used for hypotheses concerning the difference of two means where σ_1 and σ_2 are unknown, as long as the two populations are normally distributed.
- <u>Usually if $n \geq 30$, S is a good enough estimator of σ, and the normal distribution is typically used instead</u>.

Thank You

"We trust in GOD, all others must bring data"